Understanding Human Navigation using Bayesian Hypothesis Comparison

vorgelegt von
Martin Becker

Dissertation
zur Erlangung des naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg

Abstract

Understanding human navigation behavior has implications for a wide range of application scenarios. For example, insights into *geo-spatial* navigation in urban areas can impact city planning or public transport. Similarly, knowledge about navigation *on the web* can help to improve web site structures or service experience.

In this work, we focus on a hypothesis-driven approach to address the task of understanding human navigation: We aim to formulate and compare ideas — for example stemming from existing theory, literature, intuition, or previous experiments — based on a given set of navigational observations. For example, we may compare whether tourists exploring a city walk "short distances" before taking their next photo vs. they tend to "travel long distances between points of interest", or whether users browsing Wikipedia "navigate semantically" vs. "click randomly".

For this, the Bayesian method HypTrails has recently been proposed. However, while HypTrails is a straightforward and flexible approach, several major challenges remain: i) HypTrails does not account for heterogeneity (e.g., incorporating differently behaving user groups such as tourists and locals is not possible), ii) HypTrails does not support the user in conceiving novel hypotheses when confronted with a large set of possibly relevant background information or influence factors, e.g., points of interest, popularity of locations, time of the day, or user properties, and finally iii) formulating hypotheses can be technically challenging depending on the application scenario (e.g., due to continuous observations or temporal constraints). In this thesis, we address these limitations by introducing various novel methods and tools and explore a wide range of case studies.

In particular, our main contributions are the methods MixedTrails and SubTrails which specifically address the first two limitations: MixedTrails is an approach for hypothesis *comparison* that extends the previously proposed HypTrails method to allow formulating and comparing heterogeneous hypotheses (e.g., incorporating differently behaving user groups). SubTrails is a method that supports hypothesis *conception* by automatically discovering interpretable subgroups with exceptional navigation behavior. In addition, our methodological contributions also include several tools consisting of a distributed implementation of HypTrails, a web application for visualizing geo-spatial human navigation in the context of background information, as well as a system for collecting, analyzing, and visualizing mobile participatory sensing data.

Furthermore, we conduct case studies in many application domains, which encompass — among others — geo-spatial navigation based on photos from the photo-sharing platform Flickr, browsing behavior on the social tagging system BibSonomy, and task choosing behavior on a commercial crowdsourcing platform. In the process, we develop approaches to cope with application specific subtleties (like continuous observations and temporal constraints). The corresponding studies illustrate the variety of domains and facets in which navigation behavior can be studied and, thus, showcase the expressiveness, applicability, and flexibility of our methods. Using these methods, we present new aspects of navigational phenomena which ultimately help to better understand the multi-faceted characteristics of human navigation behavior.

Zusammenfassung

Menschliches Navigationsverhalten zu verstehen, kann in einer Reihe von Anwendungsgebieten große Fortschritte bringen. Zum Beispiel können Einblicke in räumliche Navigation, wie etwa in Innenstädten, dabei helfen Infrastrukturen und öffentliche Verkehrsmittel besser abzustimmen. Genauso kann Wissen über das Navigationsverhalten von Benutzern im Internet Entwickler dabei unterstützen, Webseiten besser zu strukturieren oder generell die Benutzererfahrung zu verbessern.

In dieser Arbeit konzentrieren wir uns auf einen hypothesengetriebenen Ansatz, um menschliches Navigationsverhalten zu verstehen. Das heißt, wir formulieren und vergleichen Hypothesen basierend auf beobachteten Navigationspfaden. Diese Hypothesen gründen zumeist auf existierenden Theorien, Literatur, vorherigen Experimenten oder Intuition. Beispielsweise kann es interessant sein, zu vergleichen, ob Touristen, die eine Stadt erkunden, eher zu nahegelegenen Sehenswürdigkeiten laufen als vornehmlich große Strecken zurückzulegen. Weiterhin kann man in Online-Szenarien vergleichen, ob Benutzer zum Beispiel auf Wikipedia eher semantisch navigieren als zufällig Artikel anzusurfen.

Für diese Szenarien wurde HypTrails entwickelt, ein Bayes'scher Ansatz zum Vergleich von Navigationshypothesen. Doch obwohl HypTrails eine einfach zu benutzende und sehr flexible Methode darstellt, hat es einige deutliche Schwachstellen: Zum einen kann HypTrails keine heterogenen Prozesse modellieren (z.B., um das Verhalten von verschiedenen Nutzergruppen, wie etwa von Touristen und Einheimischen, zu unterscheiden). Außerdem bietet HypTrails dem Benutzer keine Unterstützung bei der Entwicklung neuer Hypothesen. Dies stellt vor allem in Kombination mit großen Mengen an Hintergrundinformationen und anderen Einflussgrößen (z.B., Sehenswürdigkeiten, Beliebtheit von Orten, Tageszeiten, oder verschieden Benutzereigenschaften) eine große Herausforderung dar. Außerdem kann sich das Formulieren von adäquaten Hypothesen abhängig vom Anwendungsszenario als schwierig erweisen (z.B. aufgrund von kontinuierlich räumlichen Koordinaten oder zeitlichen Nebenbedingungen). In dieser Arbeit setzen wir an eben jenen Problemstellungen an.

Unsere Hauptbeiträge bestehen dabei aus den Ansätzen *MixedTrails* und *SubTrails*, die vor allem die ersten beiden genannten Schwachstellen adressieren: MixedTrails stellt einen Ansatz zum Vergleich von Hypothesen dar, der auf HypTrails basiert, es aber ermöglicht heterogene Hypothesen zu formulieren und zu vergleichen (z.B., bei Benutzergruppen mit unterschiedlichem Bewegungsverhalten). Während SubTrails eine Methode darstellt, die das Entwickeln neuer Hypothesen unterstützt, indem es die automatische Entdeckung von interpretierbaren Subgruppen mit außergewöhnlichen Bewegungscharakteristiken ermöglicht. Weiterhin stellen wir drei weitere Beiträge vor: eine verteilte und hochparallele Implementierung von HypTrails, ein Werkzeug zur Visualisierung von räumlicher Navigation zusammen mit Hintergrundinformationen, sowie ein System zur Sammlung, Analyse und Visualisierung von Daten aus dem Bereich des Participatory Sensing.

Schließlich führen wir mehrere Studien in verschiedenen Anwendungsbereichen durch. Wir untersuchen etwa räumliche Navigation basierend auf Photos der Onlineplattform Flickr, Browsing-Verhalten der Nutzer auf dem Verschlagwortungssystem BibSonomy, und das Arbeitsverhalten von Nutzern einer kommerziellen Crowdsourcing-Plattform.

Dabei entwickeln wir mehrere Ansätze, um mit den Eigenheiten der spezifischen Szenarien umgehen zu können (wie etwa kontinuierliche räumliche Koordinaten oder zeitliche Nebenbedingungen). Die Ergebnisse zeigen die Vielzahl von Anwendungsgebieten und Facetten, in denen Navigationsverhalten analysiert werden kann und illustrieren so die Ausdrucksstärke, vielseitige Anwendbarkeit und Flexibilität unserer Methoden. Gleichzeitig, geben wir neue Einblicke in verschiedene Navigationsprozesse und ermöglichen so einen wichtigen Schritt hin zum Verständnis der vielfältigen Ebenen menschlichen Navigationsverhaltens.

Acknowledgements

Navigating the woe- and wonderful waters to my dissertation has been a task that I never could have done alone. First and foremost my thanks goes out to Andreas Hotho, my supervisor, captious critic, and greatest supporter. We have come a long way, and without his guidance and friendship I would not be where I am now. I am deeply grateful for the time that I had at his research group DMIR. Secondly, I would like to thank Florian Lemmerich who has been a colleague, mentor, and inspiration to me since my diploma thesis. He is the perfect partner for discussions about research, science, and life. Florian was also part of "my GESIS group" including Philipp Singer, and Markus Strohmaier, who were an amazing team for thinking up new ideas, and publishing awesome papers. They supported me greatly throughout the last three years of my PhD for which I am profoundly thankful. This also includes Denis Helic from TU Graz who provided key ideas for our work.

I also thank all my colleagues from the Chair of Computer Science VI at the University of Würzburg and especially the gals and guys from the DMIR Group: Thomas Niebler, Alexander Dallmann, Lena Hettinger, Daniel Zoller, Daniel Schlör, Albin Zehe, and Florian Lautenschlager. I sincerely appreciate every last one of them. They are the greatest group of researchers there is to work and play with. Period. Here, I want to single out Thomas Niebler, who has been with me since the very beginning of our academic career. He has been a friend and great support in so many ways! I also want to thank Daniel Zoller for his zealous way of pushing me towards writing up my thesis.

Beyond my academic colleagues, I want to thank Annika Drumm for being my dancing partner for a major part of my time as a PhD student. She definitely helped to keep me sane! Thanks also to my trainer and training partner Luca Agnetta, as well as the Salsa crew which I spent the rest of my free time with.

Finally, my thanks goes out to Laura Kemmer, Johannes and Sabrina Wehner, as well as my family Wolfgang, Magdalena, Peter, Anna, and Robin Becker: Thanks to Laura for her love and support, and for not killing me (yet). Thanks to Johannes and Sabrina for keeping me well fed. And last but not least, thanks and thanks again to my family for supporting me all these years and grounding me like nobody else can.

Thank you very, very much, everyone!

Martin Becker
Würzburg, 21.12.2017 / 21.03.2018

Copyrighted Material

In several of our figures, we use map tiles from external sources which require and deserve attribution:

- On the thesis cover as well as in Figures 1.1 and 1.2, we use map tiles (Carto Light "Positron") by Carto (under CC BY 3.0[1]) which incorporate data by OpenStreetMap (under ODbL[2]).
- The rest of our figures with map material is based on the standard tile layer from OpenStreetMap (©OpenStreetMap contributors under CC BY-SA[3]).

[1] https://creativecommons.org/licenses/by/3.0
[2] https://opendatacommons.org/licenses/odbl
[3] https://www.openstreetmap.org/copyright

Contents

List of Figures

List of Tables

1. Introduction

Understanding human navigation behavior has been of interest to researchers and practitioners for well over a century: One of the earliest "modern studies" [377] in the area of *geo-spatial* navigation is from 1885 by Ravenstein who investigated migration patterns in several countries using census data [417]. Then, with the significant urban growth throughout the 20th century [512], understanding human navigation became more and more important in order to address the challenges arising in urban planning. For example, there is work related to human navigation with regard to commuting behavior [184], land use [95], or travel demand [351].

However, human navigation behavior is not restricted to the geo-spatial domain. Rather, navigation is defined more generally as "The process or activity of accurately ascertaining one's position and planning and following a route."[1]. This definition also encompasses navigation on information environments such as text books or library catalogs where the user browses or searches for information. Understanding human navigation in this context has become increasingly relevant with the advent of online systems and the world wide web where users have to find their way through vast amounts of information on a daily basis. For example, users navigate Wikipedia [519] to find specific information, browse videos on YouTube [33], or search for products to buy in online shops [112].

At first glance, geo-spatial navigation and navigation on the web (e.g., as shown in Figure 1.1) are fundamentally different. That is, the former can be experienced in the real-world, while the latter is virtual. Nevertheless, the problem settings in each scenario are very similar: for example managing traffic (e.g., cars [95] vs. webpage traffic [141]), optimizing infrastructures (improving transport systems [180] vs. introducing new hyper links [429]), or supporting users in their navigation tasks (routing [135] vs. product recommendation [419]). Similar analogies can be drawn to other fields, such as navigating music playlists [68] or app usage on cellphones [542].

To address these (and other) problem settings, it is — independent of the application domain — essential to *understand the underlying processes of human navigation behavior*. For example, knowing that people try to minimize the time to travel between home and work (rather than, e.g., using a scenic route) can help officials to plan new public transportation systems. Similarly, understanding that users tend to follow certain strategies when looking for information on Wikipedia introduces possibilities to improve Wikipedia's category system or, generally, the link network between articles.

[1] `https://en.oxforddictionaries.com/definition/navigation`, accessed: December 2017

(a) Photo trails in Manhattan based on Flickr **(b)** Navigation between articles on Wikipedia

Figure 1.1.: An illustration of human navigation. Human navigation can be observed in many application domains. This figure depicts geo-spatial navigation (a) and navigation on the web (b). The former shows photo trails collected from Flickr [44] where red transitions represent tourists and black transitions represent locals (the trails are restricted to pedestrians selected via speed and travel distance). The latter shows transition counts on a subset of articles on Wikipedia collected in the context of the game Wikispeedia [520] (the articles and counts are restricted to those reachable by a single link from the article "United States"). These examples illustrate the sheer complexity of human navigation behavior. In this thesis, we present novel methods to understand such behavior and provide insights into the underlying processes of several complex application domains.

1.1. Comparing hypotheses about human navigation

Understanding the underlying processes of human navigation behavior is not a trivial task. For example, navigational characteristics may be different depending on the application domain, i.e., international travel may be governed by different laws than urban navigation and browsing on Wikipedia can be very different from user behavior on Facebook. Also there is a multitude of factors that may influence the underlying processes of the observed navigation behavior. For example, in the geo-spatial context this may encompass the infrastructure of a city or the influence of points of interest, and for web pages their similarity to other pages or their general popularity may play an important role.

Thus, to study these factors and to find concise explanations for human navigation behavior in various settings, we need adequate methodology to formulate and compare our ideas about the corresponding underlying processes. To this end, HypTrails [453] has been proposed recently. It is a flexible Bayesian approach for formulating and comparing hypotheses about human navigation behavior in very different application domains. Such hypotheses usually stem from existing theory, literature, previous experiments or intuition (cf. explanatory modeling [446]) and can incorporate many different aspects of human

behavior. For example, HypTrails can be applied to compare hypotheses in online settings, such as "users navigate semantically" on Wikipedia vs. they "simply browse randomly" [144], as well as in a geo-spatial context, e.g., for analyzing if tourists prefer to walk "short distances" before taking their next photo when exploring a city vs. they tend to "travel long distances between points of interest" before their next shot [44]. This empowers a wide variety of case studies on human navigation in very different application scenarios. However, while HypTrails is a straightforward and flexible approach, it has limitations with regard to formulating hypotheses, the complexity of the underlying processes of human navigation, as well as the conception of novel hypotheses. We outline the corresponding challenges in the next section.

1.2. Challenges of hypothesis comparison

As mentioned in the previous section, HypTrails is a powerful tool for understanding human navigation behavior. It provides a flexible framework for formulating and comparing hypotheses about the underlying processes of navigational behavior. However, HypTrails has limitations which we outline in this section.

1.2.1. Complexity of human behavior

Human navigation behavior is inherently complex. One major aspect illustrating this is grounded in its heterogeneous characteristics. That is, there are often several sub-processes responsible for observed navigational phenomena. For example, in the geo-spatial context, it was shown that individual movement is very different from aggregated views on human mobility [99]. Other studies break down human mobility into sets of characteristic components [157, 435]. Also see Figure 1.1a illustrating the difference in behavior of tourists and locals taking pictures in New York City. Similarly, on the web, navigation behavior is often categorized into several classes (e.g., searching, general browsing, and serendipitous browsing [92]) and different user groups have been shown to exhibit specific behavioral traits (for example younger and older populations [353]). This illustrates how important it is to understand human navigation as a heterogeneous process instead of assuming that all users in any situation show the same navigation behavior. To this end, *background information*, such as different user properties (e.g., being a tourist or not) or the context in which the navigation process is performed (e.g., the time of the day), play an important role. However, HypTrails assumes a homogeneous process underlying the observed navigation behavior and does not allow to model heterogeneous hypotheses.

Another aspect illustrating the complexity of human navigation behavior is its inherently *large scale* nature (e.g., consider the English Wikipedia with more than 254 million page views per day on over five million articles)[2]. This requires the methodology applied to analyze and explain human behavior to be able to handle processes spanning extensive

[2] `https://tools.wmflabs.org/siteviews/?platform=all-access&source=pageviews&agent=user&start=2016-01-01&end=2016-12-31&sites=en.wikipedia.org`, accessed: December 2017

dependency structures as well as massive amounts of observations. However, HypTrails is not specifically designed to cope with such large-scale environments.

1.2.2. Hypothesis conception

The conception of novel hypotheses about human navigation behavior can be a challenging task. Especially when considering the vast amount of possible background information available to the practitioner for explaining its underlying processes. This issue is even more prominent when prior domain knowledge or specific ideas about possible hypotheses are limited. For example: Is the distance between places, the attractiveness of points of interests, the popularity of locations, or a combination of all three of them the best approach to explain urban navigation? In addition, the inherent heterogeneity of human behavior (as mentioned in the previous section) introduces further complexity into the procedure of conceiving hypotheses. For example: Is it important to distinguish between younger and older people, tourists and locals, or both? Should we consider the time of the day (e.g., rush hour vs. night times)? These aspects result in an exponentially growing search space of possible explanations of human navigation behavior. However, by itself, HypTrails only allows to compare *existing* hypotheses, e.g., from literature or intuition, and does not support the process of *conceiving novel hypotheses* leaving the selection of relevant background information to the user.

1.2.3. Formulating hypotheses

Finally, even without considering the heterogeneous nature of human navigation (Section 1.2.1)and assuming that there is no lack of ideas to formulate hypotheses Section 1.2.2, the process of formulating hypotheses can still be challenging. That is, there are many different application domains in which human navigation behavior can be observed (e.g., navigating urban areas, or browsing Wikipedia, cf. Figure 1.1). However, while HypTrails provides a flexible framework, each domain has its own characteristic properties and may not fit the HypTrails framework directly. For example, in the geo-spatial domain, navigation is a continuous process, whereas HypTrails requires a discrete state space to formulate hypotheses. Also, structural factors can play an important role: When studying web systems, it is important to consider network structures in order to formulate realistic hypotheses (e.g., there is no link from the article *crocodile*[3] to the article *teacup*[4] on Wikipedia). Similarly, some web pages may only be available for a limited amount of time (such as items in an online store). This illustrates that the process of formulating hypotheses in the HypTrails framework can be challenging and requires careful consideration depending on the application domain in order to yield interpretable results.

[3] `https://en.wikipedia.org/wiki/Crocodile`, accessed: December 2017
[4] `https://en.wikipedia.org/wiki/Teacup`, accessed: December 2017

1.3. Contribution

In this thesis, we extend existing methodology for as well as research on human navigation behavior and contribute to understanding its underlying processes. In particular, we address the issues of *comparing complex hypotheses* as well as *hypothesis conception* (as covered in the previous Sections 1.2.1 and 1.2.2, respectively) by introducing novel methodology with a focus on the inherent heterogeneity of of human navigation. Furthermore, we provide insights into human navigation behavior through an *extensive set of studies* and — in the process — develop several specialized approaches for *formulating hypotheses* (Section 1.2.3) in the context of various application domains. In the following, we give more details on our contributions which we structure as visualized by Figure 1.2.

1.3.1. Methods

On a methodological level, we categorize our work into two mutually beneficial strategies: hypothesis comparison and hypothesis conception (see Figure 1.2). *Hypothesis comparison* refers to formulating *existing* ideas as hypotheses (stemming from theory, domain knowledge, previous experiments, or intuition) and comparing them based on observed data. *hypothesis conception*, on the other hand, refers to deriving *novel* ideas and hypotheses about the underlying processes of human navigation based on a given set of observed navigational data, e.g., by visualizing, exploring or automatically discovering regularities and patterns. In the following, we summarize our approaches and contributions along these two concepts.

1.3.1.1. Comparing complex hypotheses

As mentioned in Section 1.1, we employ the recently proposed HypTrails approach for comparing hypotheses about human navigation behavior. However, HypTrails has limitations with regard to the complexity inherent to navigational processes (cf. Section 1.2.1). The next two paragraphs summarize our work to address these limitations consisting of contributions on a methodological as well as on a tooling level.

In Section 1.2.1, we have emphasized the homogeneous nature of HypTrails which does not allow to explicitly incorporate heterogeneity into hypotheses, (e.g., in the form of differently behaving user groups such as tourists and locals as depicted in Figure 1.1a). Thus, as one of our main contributions, we extend HypTrails to account for the heterogeneity inherent to human navigation behavior. That is, we propose the MixedTrails approach (see Chapter 4), which allows to model and compare hypotheses composed of a set of *different* navigation processes, instead of assuming the *same* process for the complete set of observed data (as HypTrails does). For example, in the context of our Flickr case study (Chapter 7), we formulate the hypothesis that "tourists exploring a city are more likely to take their next photo after a short distance while locals are more selective resulting in longer distances between photos". Similarly, for search-trails on Wikipedia (Section 11.2), we test whether it is plausible that "users navigate to central articles first, before using semantic relatedness to find the article they search for".

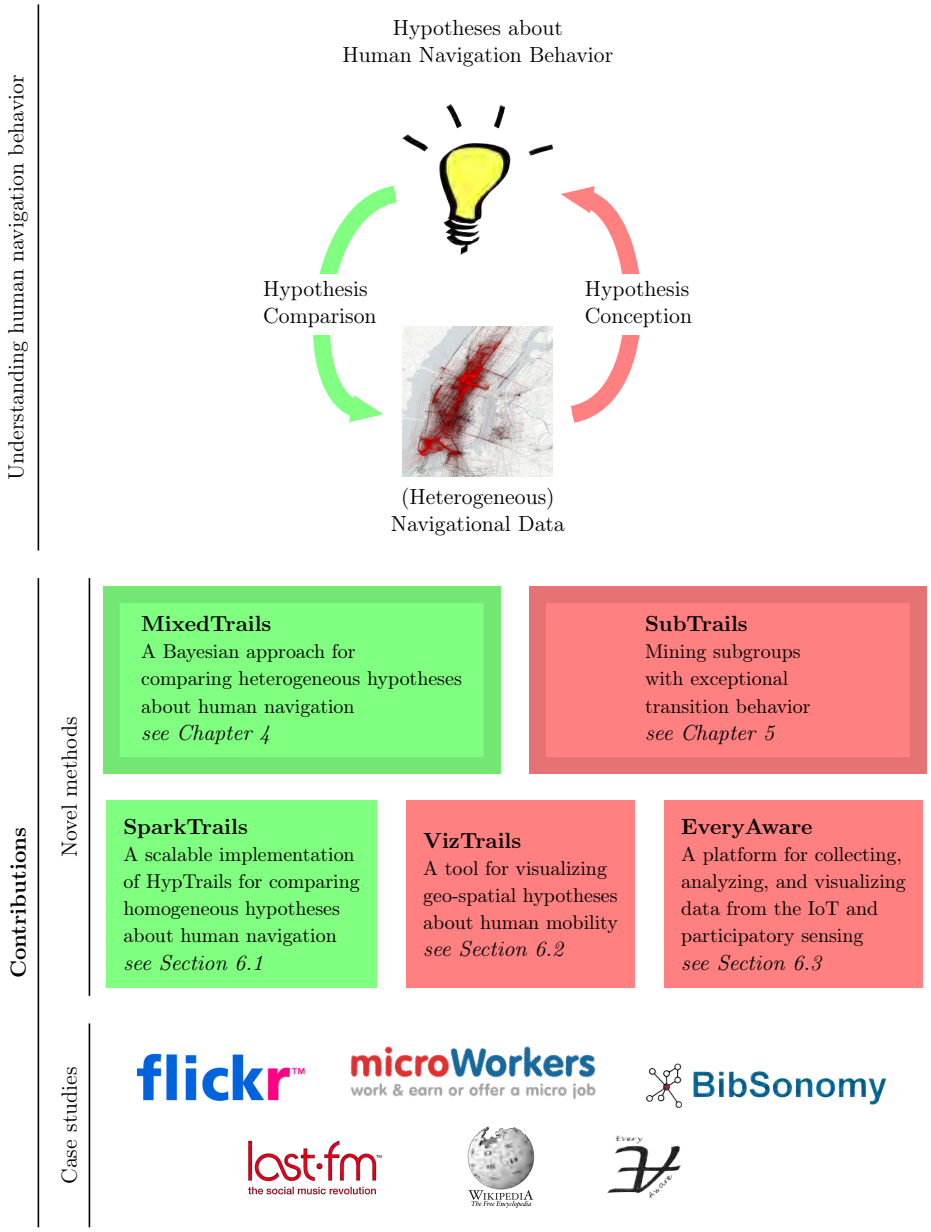


Figure 1.2.: Our contributions towards understanding human navigation behavior. The core concept of this thesis are understandable hypotheses (explanations) about human navigation behavior. To analyze a given dataset with respect to such hypotheses, we can either formulate a set of candidates (e.g., from theory, literature, previous experiments, or intuition) and *compare* them based on how well they explain the observations. Or we can directly analyze the data in order to *conceive* plausible explanations, e.g., based on the regularities and patterns we observe. In both areas we contribute a novel method (MixedTrails and SubTrails, respectively). We also provide a variety of tools including, for example, an algorithm for efficiently handling large problem settings as well as several visualization mechanisms. Finally, we analyze human navigation behavior in various complex applications domains and — in the process — introduce approaches to handle challenges like continuous navigation data or temporal constraints.

On the tooling level, we furthermore introduce SparkTrails (Section 6.1), a distributed implementation of HypTrails based on the MapReduce paradigm. It allows to apply HypTrails in large-scale application scenarios enabling hypothesis comparison for extensive dependency structures as well as massive amounts of observations. This method was used in several of our case studies (Part III).

1.3.1.2. Conception of novel hypotheses

As outlined in Section 1.2.2, methods for supporting the conception of hypotheses about human navigation behavior are required in cases when prior domain knowledge or specific ideas are missing (see red elements in Figure 1.2). In this thesis, we contribute several methods and tools based on descriptive analysis and exceptional model mining in order to support the process of conceiving novel hypotheses and to better understand the underlying processes of human navigation.

As one of our main contributions in this thesis, we exploit the descriptive nature of subgroup discovery by employing the framework of exceptional model mining — an extension of subgroup discovery — to propose SubTrails, a novel method for mining subgroups of sequence data (see Chapter 5). By design, SubTrails returns interpretable subgroups with exceptional transition behavior based on a set of attributes provided by background information. This allows us to find patterns like "tourists exhibit exceptionally different navigation behavior compared to the overall population when navigating a city". The local nature of this approach, i.e., the fact that we find *subsets* of the data with exceptional properties, allows to further explore heterogeneity in human navigation behavior (cf. Sections 1.2.1 and 1.2.2).

On the tooling level for conceiving novel hypotheses, we present VizTrails (Section 6.2) and the EveryAware platform (Section 6.3). VizTrails is an interactive visualization tool for better understanding how navigation behavior in a geo-spatial context materializes. And the EveryAware system is a holistic platform for collecting, analyzing, and visualizing mobile environmental measurements specifically featuring the environmental aspects of noise pollution and air quality measurements. Besides covering the mobile aspects of human navigation represented by collecting geo-spatial tracks, EveryAware explicitly incorporates subjective annotations by users allowing to study navigation behavior in a unique scenario.

1.3.2. Case studies

Human navigation behavior can be observed in a wide variety of settings and exhibits very specific characteristics in each scenario. In this thesis, we present work in several such domains broadly applying our methodology and tools mentioned in Section 1.3.1. In the process, we develop approaches to handle several challenges mentioned in Section 1.2.3 by formulating hypotheses in the context of applications with continuous navigation data or temporal constraints.

In particular, we explore photo trails using the full range of our methods, i.e., analyzing overall behavior by comparing hypotheses about urban navigation (e.g., based on proximity

and popularity of points of interest) using HypTrails, extracting subgroups of the data using SubTrails, and formulating heterogeneous hypotheses accounting for different user groups using MixedTrails (in this case, tourists and locals). A special challenge in this case study is handling continuous navigation data. Furthermore, with our study on explorative components of geo-spatial navigation in the context of a participatory air quality sensing campaigns, we cover a previously seldom inspected domain of navigation behavior. Here, in addition to a continuous navigation process, we have to cope with temporally dense observations requiring a very restricted type of hypotheses. Additionally, in the context of online social bookmarking systems, we formulate novel hypotheses about the browsing behavior of users in folksonomies and study the characteristics of different user groups on the BibSonomy platform[5]. And for crowdsourcing environments, we introduce one of the first studies which compares hypotheses about task-choosing behavior (e.g., based on monetary incentives or category consistency) on actual log data instead of using surveys. For this case study, we handle temporal constraints caused by the limited availability of campaigns. Finally, we cover several small scale case studies which include a preliminary analysis of navigational processes when exploring urban noise pollution as well as two examples of applying MixedTrails and SubTrails to navigation on Wikipedia[6] and the music platform last.fm[7], respectively.

Overall, our case studies illustrate the variety of domains and facets in which navigation can be studied and, thus, showcase the applicability and flexibility of our approaches. In the process, we present new aspects of navigation phenomena which ultimately help to better understand the multi-faceted characteristics of human navigation behavior.

1.4. Structure of this work

This thesis is divided into three parts: background (Part I), methodological contributions (Part II), and case studies (Part III).

In the background part (Part I), we first cover the current state of understanding human navigation behavior (Chapter 2) in the context of geo-spatial navigation (Section 2.1) as well as navigation on the web (Section 2.2). Then, we introduce the methodological foundations of this thesis (Chapter 3). This includes information on discrete navigational processes (Section 3.1) and Markov chains (Section 3.2) which define the underlying concepts of our contributions. In the same section, we continue by reviewing the HypTrails approach for comparing navigational hypotheses (Section 3.3.2) as well as exceptional model mining (Section 3.4). Each of these two concepts is used as the foundation of one of our main methodological contributions (MixedTrails and SubTrails, respectively). Furthermore, many of our case studies strongly rely on HypTrails.

After covering the required background information, we introduce several novel methods for analyzing human navigation behavior (Part II). Our main contributions are the two methods MixedTrails (Chapter 4) and SubTrails (Chapter 5). The hypothesis comparison

[5] `https://www.bibsonomy.org`, accessed: December 2017
[6] `https://www.wikipedia.org`, accessed: December 2017
[7] `https://www.last.fm`, accessed: December 2017

approach, MixedTrails, extends the previously proposed HypTrails method to allow formulating and comparing *heterogeneous* hypotheses, while the hypothesis conception approach, SubTrails, represents an algorithm for mining subgroups with *exceptional* navigation behavior. Our methodological contribution also includes several tools in Chapter 6 which consist of a distributed implementation of HypTrails (Section 6.1), a tool for visualizing geo-spatial human navigation in the context of background information (Section 6.2), as well as a system for collecting, analyzing, and visualizing mobile participatory sensing data (Section 6.3).

The final part (Part III) of our thesis consists of a broad variety of studies on real-world human navigation behavior. This includes work on geo-tagged photos from Flickr in Chapter 7, on exploration processes in the context of a participatory air quality sensing campaign (Chapter 8), on browsing behavior on the social tagging system BibSonomy in Chapter 9, on task choosing behavior on a commercial crowdsourcing platform (Chapter 10), as well as a several small scale studies on navigation in the context of urban noise pollution exploration, Wikipedia, and music play lists (Chapter 11). Finally, Chapter 12 closes this thesis with a summary as well as remarks on future work.

Part I.

Background

2. Current state of understanding human navigation behavior

Our introduction in Chapter 1 illustrates that understanding human navigation behavior has implications for a variety of application domains, such as city planning and public transport in geo-spatial contexts, or improving web site structures and service experience in online settings. Consequently, many studies have been conducted in both areas supported by a steadily increasing number of novel data sources, such as GPS tracks or location based social media check-ins for geo-spatial navigation, and large-scale log datasets from internet service providers or online platforms in the context of navigation on the web.

In this thesis, we also contribute towards understanding human navigation behavior by introducing novel methodology to analyze and explore observed data in geo-spatial as well as online settings, and conduct a wide range of different case studies. Thus, to place our work, this chapter reviews existing work on human navigation behavior in the context of geo-spatial navigation (Section 2.1), as well as navigation on the web (Section 2.2). In particular, for both fields, we first outline early work and give a general overview of data sources and domains. Then, we introduce a set of existing modeling approaches in order to illustrate how other studies have explained observed navigational data. We also list work on regularities and patterns exhibited by human behavior illustrating the different factors that can influence such observations. Finally, we reiterate over the reviewed research and highlight studies where the heterogeneous nature of human behavior is particularly prominent (e.g., in the form of differently behaving user groups or characteristic navigation patterns on different days of the week). This emphasizes the value of our main contributions, i.e., enabling and supporting the process of exploring, discovering, and explaining the heterogenous aspects of navigational data (cf. Chapter 1).

2.1. Geo-spatial behavior

As we have argued in Chapter 1, human navigation behavior can be observed in a variety of domains. Besides navigation on the web, as we will cover in Section 2.2, this specifically includes human mobility in the geo-spatial context.[1] Human mobility studies are highly relevant with regard to our daily lives, since they have implications for a wide variety of applications such as understanding human migration patterns [234, 417, 450], improving urban planning and traffic management [61, 143, 207, 267], crime prevention [74, 85],

[1] Due to their ubiquitous nature, geo-spatial mobility, movement, and navigation have been studied in a wide variety of scenarios. Corresponding studies do not only center on human society but also include work about animal mobility [53, 158, 428, 452], and sometimes even its relation to human movement characteristics [250].

predicting the spread of diseases [32, 48, 430], or recommending places or travel routes for locals and tourists [106, 189, 295, 476]. In this section, we place our methodological approaches (Part II) as well as our case studies (Part III) into the corresponding context of human mobility by giving a brief overview of this broad field of research.

Specifically, we give a short overview of early work on human mobility (Section 2.1.1), and follow up with a small summary of the development of available geo-spatial data sources (Section 2.1.2). Then, we cover several aspects of modeling geo-spatial navigation behavior (Section 2.1.3), and review results on patterns and regularities discovered by previous work in Section 2.1.4. Afterwards (Section 2.1.5), we cover the notion of heterogeneity in human mobility data which is especially relevant for this thesis since our contributions specifically aim at incorporating multiple sub-processes for explaining human navigation behavior instead of employing a single, possibly oversimplified explanation (cf. Chapter 1). We close this section with a discussion on the relation of our work to the previously covered studies (Section 2.1.6).

2.1.1. Early work and development

One of the earliest "modern studies" [377] on human movement has been conducted by Ravenstein [416, 417] who analyzed census data from several countries. He found certain laws governing the process of migration. Later studies revisited these laws and put them into more modern terms such as "Zipf's law" [484], or the gravity model [88].[2] Consecutive work also studied human migration patterns [304, 458, 473, 554] and, like Ravenstein, mainly worked with census data or employed information gathered through surveys.

Human mobility studies, however, are not limited to migration. Especially with the significant urban growth throughout the 20th century [512], human mobility models became important in order to manage the corresponding challenges in urban planning. For example, there are studies on land use [e.g., 95], commuting behavior [e.g., 184], or travel demand [for an overview see, e.g., 351]. Many of these studies and the corresponding methods and models such as the gravity model [88], the model of intervening opportunities [473], trip and activity based travel demand models [351], or origin-destination flow estimation [86] still influence research on human mobility today [32, 139, 378, 535].

Nevertheless, all of these studies were limited in that the used data was very sparse with regard to spatial as well as temporal resolution. Especially information from surveys — as has been (and still is) often employed for human mobility studies — seldom covers a large number of individuals. This changed with the advent of cellular phones, the global positioning system, GPS, and the World Wide Web, as covered in the following section.

2.1.2. Data sources for geo-spatial navigation

While there are many sources to collect data about human mobility, there are three technologies which have strongly shaped human mobility research in the past two decades, i.e., mobile phones, the global position system GPS, and their combination in the form of smartphones giving rise to social networks with location features. In this section, we

[2] A form of the gravity model was even mentioned as early as 1781, cf. [450].

briefly cover the data sources associated with these technologies. For another overview on data sources in the context of human mobility, we refer to Asgari et al. [19].

2.1.2.1. Call detail records (CDR)

With the launch of the second generation of cellular technology in Finland in the early 1990s, human communication has changed tremendously [377]. In a few years the world coverage of mobile phone subscriptions grew from 12% of the world population to 96% in 2014 resulting in 6.8 billion subscribers [65]. For these users, telecommunication companies keep call detail records (CDR) from which locations can be inferred every time the user initiates or receives a call or a text message using the location of the tower routing the communication [214]. Blondel et al. [65] give an overview of the studies emerging from this kind of data in general. With regard to mobility, even though there are studies warning about bias [329, 415, 518], and even though CDR data does not provide a high spatial (e.g., $2\,\mathrm{km}^2$ to $3\,\mathrm{km}^2$, [420]) or temporal resolution (hand picked intervals or dependent on the frequency of text messages and calls), studying human mobility using large sets of CDR data is an active area of research [e.g., 46, 214, 245].

2.1.2.2. Global positioning system (GPS)

Another source of location data often used to study human mobility are GPS tracks [e.g., 91, 276, 562]. The global positioning system (GPS) was developed in the 1970's by the U.S. Military, reaching full operational capacity in 1995 [409]. Since then it was used in a wide variety of applications including mobility and movement behavior research where it was employed to enhance individual travel surveys [445, 531]. After the year 2000, when the artificial accuracy limitation (selective availability) for civilian use was deactivated [531], many industries adopted GPS. This resulted in a boom of navigation devices and the integration into smartphones, the latter starting in 1999 and being continuously improved, e.g., by introducing assisted GPS in 2004 by Qualcomm.[3] Compared to call detail records (CDR), which are sparse in time and coarse in space [46], GPS tracks are more fine grained and temporally dense. However, studies using GPS tracks are often criticized because they tend to contain small amounts of participants [415, 445], at least in the context of travel surveys. In other domains, large scale datasets are available, for example, including GPS tracks of fitness trackers or taxis [91, 400]. Even though this data, in comparison to surveys, usually lacks background information (e.g., about the users or the purpose of the trip), it can still be employed to derive interesting insights into human mobility [e.g., 185].

2.1.2.3. Social networks and locations

The introduction of the World Wide Web (WWW) in the early 1900s [59] has contributed greatly to studying human mobility [124]. For example, distributing surveys to larger amounts of participants became easier, as taken advantage of by Brockmann et al. [75], who tracked dollar bills using a web interface to reach a broad audience.[4] However, the

[3] `http://www.pcworld.com/article/2000276/a-brief-history-of-gps.html`, accessed: 11.02.2017
[4] `http://www.wheresgeorge.com/`, accessed: December 2017

most interesting change came with the increasing popularity of social networks [163] and the general adoption of smartphones at the beginning of the 21st century [377].[5] The continuously improving localization capabilities of smartphones using a combination of network, Wifi, and GPS positioning in combination with their access to the WWW gave rise to location based social networks and triggered the integration of location features into existing social networks. This led to novel research on human mobility based on dedicated location-based platforms like Gowalla[6], Brightkite[6], Foursquare[7] [116, 433] but also on existing social networks such as Facebook[8] [509], or Twitter[9] [107, 235] which started to incorporate location features into their systems. Even though these services are usually based on active check-ins, thus, sharing some draw-backs with call detail records [556] and even exhibiting an inferior temporal resolution, they provide a unique link between location sequences and semantic data such as information about the location, corresponding activities, and friendship relations [116, 233]. Other specialized systems provide very specific localized information, such as yelp![10], which implements a review system for places [cf., 82], and Flickr[11], a social photo-sharing platform supporting geo-tagged photos [cf., 37, 205, 206].

2.1.2.4. Other data sources and discussion

In Section 2.1.1 as well as this section, we have covered the "traditional" [233, 445] survey based data collection method as well as three currently often used data sources used in human mobility research, i.e., call detail records from mobile phones, GPS tracks, and social networks with location features. While there are other data sources worth mentioning, such as specifically exploiting the Wifi [368, 418, 427, 548] or Bluetooth [155, 469] capabilities of smartphones (e.g., for indoor localization), RFID technology [e.g., 94], smart cards in transport systems [334, 392], or usage data of bike sharing stations [180, 267], the formerly mentioned three sources have gained particular interest by the research community. Nevertheless, each data source has its own characteristics and may represent reality in a biased way [99]. Also note, that the different data sources can capture mobility at different scales including inter-continent scale [32, 356], inter-country scale [32, 235, 516], regional scale [193, 450], intra-city scale [378, 535], and even campus scale or within buildings [27, 94, 249]. Each can be used for different aspects of mobility, with the potential to still yield universal patterns. For more information, we also refer to surveys as by Asgari et al. [19], Chen et al. [99], and Zhao et al. [558].

[5] The IBM Simon produced in 1995 is being considered the first smartphone: https://www.bloomberg.com/news/articles/2012-06-29/before-iphone-and-android-came-simon-the-first-smartphone, accessed: 2017-02-13

[6] discontinued

[7] `https://foursquare.com/`, accessed: December 2017

[8] `https://facebook.com`, accessed: December 2017

[9] `https://twitter.com`, accessed: December 2017

[10] `https://yelp.com`, accessed: December 2017

[11] `https://flickr.com`, accessed: December 2017

2.1.3. Modeling

There are several studies categorizing and summarizing geo-spatial navigation behavior research and human mobility from different view points [e.g., 19, 99, 558]. In this section, we list some prominent results spanning the previously introduced data sources, different scales, and various approaches. Note that there are generally two fields of research, both analyzing and modeling human movement behavior with different background and methodology. They are called "travel behavior analysis" and "human mobility analysis" by Chen et al. [99]. The former consists of a longer history of transportation researchers modeling human mobility with advanced and intricate models but mostly based on small datasets and surveys. The latter is a collection of mostly computer scientists and physicists focusing on recently available larger datasets as covered in Section 2.1.2. For a comparison of both fields we refer to Chen et al. [99].

In this section, we focus on "human mobility analysis", i.e., we mostly cover results on large datasets from the various sources introduced in Section 2.1.2.[12] In particular, we first cover the prominent field of models concerning distance and opportunities for explaining and predicting human mobility characteristics. Afterwards, we review the notion of *trip displacement distributions* which is an often used concept for studying different properties of the observed data when modeling human movement processes.

2.1.3.1. Distance and opportunities

Human mobility is strongly intertwined with distance. In particular, the notion of some form of *distance decay* plays an important role in many human mobility models. In this section, we focus on models for human movement on an aggregate level. Specifically, these models predict the number of people transitioning between discrete locations: There are two prominent models, taking different vantage points on this concept, namely the *gravity approach* and the idea of *intervening opportunities*. In the following, we first introduce each model, also including the *radiation model*, a widely adopted variant of the *intervening opportunities* approach. Then, we focus on recent work comparing and extending both models. We finish with a short summary concluding that while both methods seem to be viable models of human mobility and some universal traits can be found, a universal model for human mobility has not yet been established. We also refer to Lenormand et al. [313] who give a similar overview.

The gravity model. According to Simini et al. [450] the contemporary formulation of the gravity model goes back to Zipf [570] with roots in the 18th century. It models the number of transitions between two locations proportionally to their combined population decaying with respect to some function of their distance [cf., 313]. Two frequently used decay functions, often compared against each other, are the power-law and the exponential function [105, 322]. The gravity model has seen many applications, studies and extensions in a variety contexts within the field of human mobility: For example, the original work by Zipf studied the number of persons that move between cities based on public bus travel,

[12] Note that, in the following sections as well as the rest of this thesis, we loosely use the terms "mobility", "navigation", and "movement" synonymously.

railway travel, and airway travel (with data from the office of the Federal Coordinator of Transportation) and found considerable correlation (especially for bus travel) between the observed data and the results from the gravity model. For work on the gravity model applied to transportation analysis in general, Erlander and Stewart [165] give an overview up until 1990. In more recent work, the gravity model was applied to fit the traffic flow on highways between cities in Korea [263], to explain the spreading of infectious diseases [32], and to explain patterns in check-in data from a Chinese location based social network Liu et al. [328], where that latter found a power-law distance decay effect and suspect different decays for inter- and intra-province mobility. Other studies include, but are not limited to: Gargiulo et al. [192], Griffith [218], Lenormand et al. [314], Liang et al. [322], Masucci et al. [344], and Pappalardo et al. [389].

Intervening opportunities. The intervening opportunities model was introduced by Stouffer [473]. It only indirectly models a distance decay by incorporating distance as a notion of "opportunities" between two different locations. In particular Stouffer states that "the number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities". Stouffer also notes that the concept of distance (between two locations) as well as the notion of opportunities can and must be defined in different ways in order to explain mobility in different contexts: distance can be related to units in space, cost, or time and opportunities can be defined with regard to the social situation of the study, e.g., job opportunities when studying migration or recreational areas when studying intra-urban mobility. After its introduction, the intervening opportunities model has been studied intently (see, e.g., Akwawua and Pooler [11], Haynes et al. [237], Ruiter [425], and Wills [530] listed by Lenormand et al. [313]) and was found to perform comparably to the gravity model. Even so, the gravity model has been applied more readily. Only recently [313] new models were proposed inherently building on the concept of intervening opportunities [269]. Such models include, for example, the radiation model [450], the rank-based gravity model [378], or the population weighted opportunity model [545]. Of these models the radiation model has found considerable attention in human mobility research, cf. [269, 313, 344, 388, 544].

The radiation model. The radiation model was proposed by Simini et al. [450], who argued that the gravity model, i.e., its parameterized version[13], needs parameter adjustments varying by region and suffers from analytic inconsistencies. To solve these issues they introduced the radiation model: Analogously to the gravity model the number of predicted transitions from one location to another grows as the population at both locations increases. However, it also incorporates the idea of intervening opportunities and models an absorption potential by weighing against the size of the population between both locations. They evaluated the raditaion model on hourly travel counts, migration, communication patterns, and commodity flows derived from census data, call detail records and tax documents (mostly on a state, county, or municipality scale), and found their model to fit the data better than the gravity model.

[13] For a comparison of the parameterized vs. non-parameterized gravity model we refer to, e.g., Masucci et al. [344].

Extensions. Since the introduction of the radiation model, several comparative studies, and extensions to both, the gravity and the radiation model, have emerged with contradicting results [314]. For example, in contrast to Simini et al. [450], Masucci et al. [344] compared the gravity and the radiation model and found that, generally, the (parameterized) gravity model performed better in modeling commuting flows (based on census data) on a national level, i.e., between cities and city clusters in England and Wales. Nevertheless, they also noted that the radiation model has advantages in situation were calibration data is missing and found that "for large distances and small and moderate destination population scales, the principles of the radiation model are reliable and that mobility patterns can be approached by a diffusion model [such as the radiation model] where intervening opportunities on the commuting paths prevail on the distance of such paths". Finally, they observed that on an urban level (between wards in and around the Greater London Authority area) neither model performed well. Again, slightly contradicting results were found by Palchykov et al. [388] who studied the gravity and radiation model using the number of phone calls as a proxy for movement. While they confirmed that the radiation model works better for long distances, they concluded that, on average, both models represent the processes of inter- as well as intra-city (cell tower) movement to some degree. Nevertheless, they argued that their data does not necessarily reflect reality due their use of call counts as a proxy for mobility. Finally, one of the latest comparisons of gravity and radiation models by Lenormand et al. [313] performed a systematic comparison by employing several commuting datasets. Lenormand et al. emphasized the importance of similar testing situations with regard to i) the input (population counts, jobs opportunities, etc.) and ii) the applied constraints with regard to preserving the observed number of incoming and outgoing transitions at each location (cf. Wills [530] for a discussion on constraints). They found the gravity model with an exponential decay to perform best, but — like the other studies — noted that it fails to estimate commuting flows at large distances.

Universal laws. Overall, the discussion around these models is shaped by the aim to find a universal law [e.g., 294, 378, 450], i.e., a model which i) explains human mobility at different scales (e.g., between cities and within cities) and in different contexts (such as taxi logs, call detail records or migration), and ii) with the least amount of parameters. While several studies claim that their models (extensions or variants of the previously introduced gravity, intervening opportunities, or radiation model) are universally applicable [e.g., 378, 450], a tendency was found that the gravity model performs better at short distance movement and that the radiation model is more accurate at modeling long distance mobility [e.g., 344]. Thus, in order to derive a universal law, several researchers have attempted to generalize either model. For example, Kang et al. [269] claim the formulation of a generalized version of the radiation model. They reported that it overcomes the previously found limits in modeling short range mobility at the cost of introducing several parameters (a scaling exponent and a normalization factor) in the context of using search direction and trip origin-destination (OD) constraints. Also, Simini et al. [451] formulated a model from which the gravity model, the intervening opportunity model, and the radiation model can be derived as special cases.

However, while there are numerous models, they still fail to paint a consistent picture of human mobility, either because they do not generalize to all scales and environments of human mobility or because they need a set of parameters to be fitted to the observed data. This indicates that various factors of human navigation behavior are still not understood. This thesis aims to further explore such factors incoporating a broad set of background information in order to better understand the underlying processes.

2.1.3.2. The trip displacement distribution

Studying the trip displacement (also transition length or trip distance) distribution has become a trademark for recent human mobility studies [188]. In this context, a frequent question is if trip displacement can be described as a scaling law: Many studies found [cf., 12, 188] that human travel and mobility show a power-law distribution at a larger scale (national or inter-urban) such as, for example, Liu et al. [325] and Song et al. [462], and exhibit an exponential distribution at smaller scales (e.g., at an urban level) as for example observed by Liang et al. [323] and Liu et al. [325]. Some studies also found other distributions, such as a superimposition of Poisson [188] or the log-normal distribution [12, 557]. For more information, Alessandretti et al. [12] give an extensive overview of trip displacement studies covering work from 2006 to 2016 using different data types (call detail records, taxi and user GPS traces, location based social networks, or surveys) and different scales (from 10 m to 10 000 km transitions).

Different models and approaches were applied to explain these characteristics. A variety of models exist for this purpose [e.g., 535]. However, the most common approaches are models predicting travel counts on the one hand (similar to the already mentioned gravity or radiation model [322, 328]); and trajectory based approaches on the other hand, which explicitly model individual trips. The latter models are based on random walks [188], including, in particular, Lévy flights [75, 420, 557]. Generally, Lévy flights prefer short flights with an occasional long jump in between. However, even though they cover an important aspect of human mobility they are often noted to miss other observed properties such as spatial and temporal regularities, exploration, or preferential return [cf. 214, 259, 462]. Nevertheless, these models — including for instance the random waypoint model or the concept of Brownian motion — are also extensively used in the mobile ad hoc network community [cf., 83, 420] to simulate human mobility in order to evaluate their systems.

2.1.4. Regularities and patterns

The methods and models covered in Section 2.1.3 explain or reproduce certain aspects and regularities of human mobility to an extent that can be considered a universal law. For example, independent of the scale (e.g., city or national) and the dataset (e.g., taxi or call detail records) transition counts exhibit gravity or radiation characteristics [105, 322, 450] and trip length distributions show power-law and exponential behavior [323, 462]. On a more detailed level, however, these laws and models often require fitting to the data which indicates influence factors beyond mechanical processes [e.g., 270]. Furthermore, besides these aggregate regularities it has also been widely recognized that the proposed

approaches do not cover more intricate characteristics, i.e., they do not account for the fact that a specific individual does usually not behave randomly [462]. Thus, a variety of constraints, regularities, and patterns have been studied in the process of trying to explain different aspects of human mobility. For example, to explain trip displacement as covered in Section 2.1.3.2, a variety of aspects has been studied such as the average population density in urban areas [322], place density [102], the underlying street network [259], travel times [188], or the activities of individuals [535].

In the following, we cover several patterns and regularities of human mobility which are important to understand human navigation behavior as a whole. In particular, we cover several spatial and temporal aspects (Section 2.1.4.1), the influence of activities and context (Section 2.1.4.2), as well as social factors (Section 2.1.4.3).

2.1.4.1. Spatial and temporal aspects

It has been found that individual human mobility has a high degree of spatial and temporal regularity [e.g., 107, 136, 157, 213, 214, 384, 389, 462, 488]. For example, based on call detail records, Song et al. [462, 463] showed i) that the daily mobility patterns of users are restricted to a relatively small area with relatively few explorations, i.e., they have a 1 to 10 miles radius of gyration, ii) that users exhibit preferential return[14], i.e., they visit a small set of locations frequently[15], and iii) that human mobility has predictability rate potentially as high as 93%. Furthermore, Cheng et al. [107] confirmed similar findings across several countries and cities using check-in data from location based social networks. They also found strong regularities in daily and weekly check-in frequencies and discover differences between work days and weekends. Along the same line, Kaltenbrunner et al. [267] studied the spatio-temporal activity cycles of a city using bike-sharing data. And finally, Oliveira et al. [384] found significant similarities in people's mobility habits regardless of the city and nature of the dataset (using data from OpenStreetMap[16], GeoLife [565], and call detail records). They also list three traits present in an individual's urban mobility: preference for shortest-paths, confinement, and repetitiveness, which match the patterns mentioned above very well.

2.1.4.2. Activities and context

Another important aspect of human mobility is its contextual component, i.e., understanding the incentives and purposes of human movement is essential to explain individual trajectories and improve the modeling of emerging mobility patterns. Indeed, in aggregate mobility studies and in particular in traffic and travel demand modeling there has been a shift from trip-based models, such as the four step model [351], to activity-based models [16, 61, 421], which recognize that travel demand stems from daily activity patterns. Discrete choice models are an even more fine-grained variant, which microscopically model user choices based on alternatives, trade-offs, and conditions [cf., 99].

[14] Preferential return was, for example, also used by Pappalardo et al. [389] in combination with a gravity model to simulate individual traces.

[15] In fact 70% of the time a user can be found at her most visited location. [462, 463]

[16] `https://wiki.openstreetmap.org/wiki/API`

In the following, we cover several studies which explicitly take into account activities as well as contextual information in order to model geo-spatial human navigation behavior. In particular, we review work which aims at decomposing geo-spatial behavior into different activities, we cover studies analyzing human mobility on the activity-level instead of directly considering trajectories, and we list research on coping with sparsity of contextual information for raw trajectory data needed to derive activity information.

Composition of activities. In contrast to the field of travel demand and traffic modeling, which tries to build fine-grained models, the data mining community focuses more on finding general patterns in large datasets. In this context, recent work has found that human mobility can be separated into basic daily activity patterns. For example, Eagle and Pentland [157] observed specific "eigenbehaviors" between locations like home or work which allowed for predicting the remaining activity of the day given its first half. And, based on taxi trips from Shanghai, Peng et al. [393] discovered that people travel on workdays mainly for three basic purposes: commuting between home and workplace, traveling from workplace to workplace, and a subsuming purpose including, e.g., leisure activities. Accordingly, Wang et al. [510] found that mobility is easier to predict on workdays than on weekends by using taxi, bus, and check-in data. Finally, Schneider et al. [435] studied daily motifs represented by activity networks. Using surveys and mobile phone data for different countries, they discovered 17 unique networks (between locations) in daily mobility which captured the behavior up to 90% of the population. Similar work is based on motifs using data from Boston, Vienna, and Singapore [260, 261, 526].

Activity-level analysis. The previously mentioned results indicate that the shift towards activity based models as well as the contextual information of trips can account for strong regularities in human mobility. Consequently, other work has focused on the more abstract concept of activities instead of concentrating on raw trajectory data. For example, Wu et al. [535] analyzed location check-in data and found that transition probabilities between activities change over time. They also simulated travel demand based transitions between two different activity types, i.e., fixed location (e.g., home, work) and multi-location (e.g., dinner, recreation) "activities". Furthermore, Thomas et al. [483] studied the connection between activities and distance decay based on survey data. Also, Preoţiuc-Pietro and Cohn [406] investigated Foursquare check-ins in order to explore activity class distributions over time and their respective transitions. They clustered users based on their activity profile and identified classes like "businessmen" or "students". Finally, Ying et al. [547] improved on predicting the next movement of mobile users by exploiting "semantic" trajectories (meaningful trajectories such as $bank \rightarrow park \rightarrow home$).

Discovering activities and context. However, an issue with studying activities to explain human mobility is, that, with big data, there is the dilemma that trajectory data is readily available but activity data as well as background data is sparse [213]. Thus, several approaches have been proposed to infer context such as activity locations and trip purposes from movement trajectories. Some use clustering approaches to find activity locations [82, 546], while others try to also explain the purpose of a location, for example, by using information from points of interest [213, 251, 539], or geo-tagged messages from Twitter [185, 559]. Also see Chen et al. [99] for further work in this area.

2.1.4.3. Social factors

A strong factor in human mobility was found to be social relations and social interaction. In the following, we cover general relations between social ties and human mobility, review work on predicting social ties from spatio-temporal behavior, and list a selection of models for human navigation behavior which incorporate social factors.

Influence of social ties on human mobility. On the most general level, studies show that the frequency of contact of individuals (analog and digital) is inversely proportional to spatial distance [210, 394]. With regard to social factors, Phithakkitnukoon et al. [396] discovered strong connections between proximity and social ties based on call detail records. Along this line, Lu et al. [331] found that people tend to travel to places where they have social bonds and Berg et al. [58] reported that — according to statistics from Netherland — more than 15% of trips are due to social activities. Furthermore, Scellato et al. [433] discovered similar socio-spatial properties across several location-based social networks (LBSN), i.e., Brightkite, Gowalla und Foursquare, signaling significant correlations between the users' social properties and their spatial behavior; whereas Phithakkitnukoon and Smoreda [395] deduced that people tend to have a more similar behavior with closer social ties. Finally, Backstrom et al. [29] observed and measured the relationship between geography and friendship based on addresses reported on Facebook and were even able to predict user locations based on the spatial behavior of their social ties.

Predicting social ties. As covered above, social ties stronlgy influence human movement behavior yielding characterstic navigational patterns. Thus, it is also possible to infer the underlying social ties by using such patterns. For example, Wang et al. [505] studied call detail records showing that the similarity between two individuals' movements strongly correlates with their proximity in the corresponding call network allowing to predict new links within this network. Eagle et al. [156] also used mobile phone records and argued that proximity is generally much higher for friends and find that up to 95% of friendship dyads can be accurately inferred using observational data on human behavior. There are many other works in this direction including, for example, Cranshaw et al. [129] who improved, e.g., on Eagle et al. [156], predicting friendship between two users by analyzing their location trails. Similarly, Crandall et al. [127] proposed a co-occurrence based method on data from the social photo-sharing platform Flickr and Xiao et al. [538] used GPS tracks to calculate a similarity measure based on *semantic trajectories* [cf. 547] which they employed to derive social ties.

Human mobility models incorporating social ties. With regard to modeling human mobility incorporating social factors, there is, for example, a variety of models in the MANET (mobile ad hoc networks) community incorporating social properties [73, 242, 371]. In particular, Boldrini and Passarella [66] incorporated social attraction, location attraction, and preference for short distances, which is reported to accurately model ICT (inter contact time) and jump sizes. Also, while Cho et al. [116] found that short-ranged travel is periodic both spatially and temporally and not effected by the social network structure, they also reported that long-distance travel was influenced by social network ties. Using these properties, they showed that social relationships can explain about 10%

to 30% of all human movement while periodic behavior explains 50% to 70%. Finally, Toole et al. [487] studied call detail records and i) showed that mobility similarity can be used to classify social relationships, ii) recovered semantic information about the nature of a link in the social network, and iii) proposed a human mobility model incorporating movement-based visitation patterns of social contacts.

2.1.5. Heterogeneity

Even though many researchers work on universal models when considering aggregate mobility [e.g., 378, 450], it has been widely recognized that human movement is strongly heterogeneous. This has been shown for individual trajectories but can also be observed at different levels of aggregate statistics. In this section, we cover several factors of heterogeneity starting with the difference of mobility patterns for individuals and follow up with a short overview on demographic and user-based influence factors. Furthermore, we review work on clusters and components within human navigation behavior.

Individual mobility. For individual mobility, similarly to Yan et al. [543], Chen et al. [99] warn that properties derived from aggregate mobility analysis cannot be used to derive regularities for individual movement. Along these lines, Gonzalez et al. [214] found that the radius of gyration strongly differs for individual users (based on call detail records). Similarly, Scellato et al. [433] studied LBSNs (Brightkite, Gowalla und Foursquare) and observed strong heterogeneity across users with regard to different characteristic spatial scales of interactions across both their social ties and social triads. And finally, even across single individuals the temporal variability was found to vary from over 20% to about 80% as Chen et al. reported in [99].

Demographics and user properties. Besides these general individual differences, there are also specific demographic factors as well as individual user characteristics which strongly influence human navigation behavior. For example, Kang et al. [268] studied different user properties such as age, gender, and, call time profiles. Among other results, they discovered differences in travel distances for younger and older people based on mobile phone data. Similarly, Yan et al. [543] found that students and retirees exhibit different movement behavior on travel survey data, and Kung et al. [294] formulated travel models differentiating between long and short distance commuters to better represent the observed data. There are also studies analyzing factors like the influence of gender [522], a difference in mobility depending on social status [107] or income [143], temporal aspects [201, 266, 267], or transportation mode [483].

Clusters and components. On a more general level, it has already been mentioned that human mobility can be seen and interpreted as a (finite) set of factors, such as clusters [406], eigenbehaviors [157], mobility networks [435], or factorized representations [393, 477]. One can define two ways of working with such components, either directly studying the navigation behavior of predefined subgroups of the observed data (see the paragraph about demographics and user properties above), or by finding factors or components of human mobility based on movement properties (e.g., the range of a trip, the time of the day, certain user properties, etc.) in an automated fashion in order to then interpret the

corresponding semantics afterwards. An example for the latter is provided by Preoţiuc-Pietro and Cohn [406], who clustered mobility based on transitions between activity classes and assigned "human categories", such as "student" or "workaholics", to the resulting behavioral clusters. Similarly, Espin Noboa et al. [166] interpreted sub-processes derived from taxi traces using tensor factorization. Instead of interpreting the factors based on intuition alone [as, e.g., 393, 477], they used HypTrails [453] in order to understand the semantics underlying the respective mobility factors. In this thesis, we also heavily exploit the interpretable nature of HypTrails for understanding homogeneous as well as heterogeneous navigational processes (cf. Chapter 4 and part III).

2.1.6. Discussion and relation to this work

In this section, we covered related work on *geo-spatial* navigation behavior. This encompassed an overview of modeling approaches, a collection of regularities and patterns, as well as a dedicated section on heterogeneity.

As mentioned in Section 2.1.5, many studies exist that analyze such heterogeneity to some extent. However, there are no dedicated methods to analyze and explain the corresponding heterogeneity in general. In contrast, in this thesis, we propose two novel methods which directly embrace the notion of explainable heterogeneity. In particular, we introduce the MixedTrails (Chapter 4) and the SubTrails (Chapter 5) approach as well as several analysis tools (Chapter 6). MixedTrails allows for comparing understandable hypotheses (from theory or intuition) about heterogeneous navigation processes, whereas SubTrails enables the automated discovery of interpretable subgroups of sequence data with exceptional transition behavior, thus, supporting the conception of novel hypotheses.

Besides heterogeneous aspects, Section 2.1.2 also illustrated that there are many different data sources and application scenarios for which human mobility can be studied. In this work, we contribute to this field of research by focusing on two specific case studies, i.e., we investigate human mobility based on Flickr photos (Chapter 7) and analyze navigation patterns in a participatory sensing setting (Chapter 8). Studying Flickr photos has the advantage that the recorded data represents events and carries background information on the one hand, and, on the other hand, still provides a relatively high spatial and temporal resolution. Also, human mobility in the context of (mobile) *participatory sensing* — which we address in this work — is a seldom covered subject of research. While some work exists, e.g., deriving air quality from human mobility patterns [e.g., 563], there are no studies explicitly trying to understand the underlying processes involved in the observed navigation behavior.

Overall, we contribute strongly to better understanding human navigation behavior in the geo-spatial context. In particular, we introduce novel methodology which practitioners and researchers can use to study human navigation behavior, and we provide novel insights into human mobility by exploring the corresponding underlying processes in unprecedented detail and in underrepresented application scenarios.

2.2. Navigation on the web

The goal of navigation behavior analysis on the web is to understand the navigation characteristics of users interacting with web resources from one or more web sites. The resulting insights can then be applied to website construction, adaptation, and management, marketing or personalization [cf. 5, 285]. Furthermore, there are many facets of web navigation to consider ranging from navigation between websites, over intra webpage navigation (e.g., navigation on Wikipedia), to the interactive processes between specific concepts of a web-platform (e.g., how users listen to music on online platforms or choose tasks on crowdsourcing systems).

In this section we first give an overview of early work on web navigation analysis (Section 2.2.1), and follow up with a brief review of data sources and abstractions of web navigation studied by previous work (Section 2.2.2). Afterwards, we cover some prominent modeling approaches (Section 2.2.3), and go over a variety of patterns and regularities which have been found in human navigation on the web (Section 2.2.4). Then, we specifically review work on the heterogeneity of navigational processes which corresponds to one of the main topic of this thesis (Section 2.2.5). Finally, Section 2.2.6 discusses the relation of our work to the previously reviewed studies.

2.2.1. Early work and overview

Some of the earliest work on navigational behavior "on the web" can be traced back to the 1980's [125, 485]. For example, Tolle [485] used a Markov model consisting of abstract states like "ERROR" or "FIND" to represent transaction logs on online public access catalogs (OPAC). He calculated state and transition probabilities in order to study "the current utilization of OPACs", i.e., how users interact with the system. Cove and Walsh [125] identified different browsing behavior categories on text within a single document, namely search browsing (goal driven), general purpose browsing (checking interesting pages), and serendipitous browsing (random). While the previously mentioned settings hardly correlate to browsing on the web, Carmel et al. [87] found similar browsing categories on hypertext structures (i.e., analyzing Apple's HyperCard on Macintosh). In the same direction, Marchionini [340] studied navigational behavior on a full-text electronic encyclopedia and found differences in navigation behavior (or "information seeking") between younger and older users (from a user base of third, fourth and six graders). Such work is still very relevant and closely related to current research, e.g., on the online encyclopedia Wikipedia [e.g., 144, 519].

However, as mentioned by Catledge and Pitkow [92], most of the early work mentioned so far has not been conducted on the World Wide Web [59], which Catledge and Pitkow describe as a "collaborative and exceedingly dynamic hypermedia system". Thus, the article "Characterizing Browsing Strategies in the World-Wide Web" by Catledge and Pitkow in 1995 can be considered one of the first navigational behavior studies on the web. Their goal was to derive "design and usability suggestions for WWW pages, sites and browsers". Based on descriptive analysis of log files from the XMosaic browser with over 43 000 events, they found user navigation patters on the web equivalent to the previously

mentioned studies on closed (hypertext) systems [87, 125]: searching, general browsing, and serendipitous browsing.

In the following years, many studies worked further understanding navigational behavior on the web. The corresponding research can be roughly divided by the different approaches being taken: Some researchers focused on modeling (closely coupled with prediction) [108, 429, 454] while others investigated regularities, patterns and strategies [4, 519, 523]. Also, with the increasing number of services provided by the web, navigational studies greatly expanded from investigating navigational processes between arbitrary websites to more specialized concepts like navigation on Wikipedia articles [373, 454, 519], behavior on social networks [51, 153], exploration of collections of music (e.g., on last.fm) [172, 312], or task choosing characteristics on crowdsourcing platforms [40]. We cover a selection of these studies in the following sections.

2.2.2. Navigational data on the web

While the traditional form of data for studying human navigation on the web is very low-level, i.e., consisting for example of server, proxy, or client logs, there exists a variety of abstractions depending on the application scenario. In this section, we briefly touch on classic web log data and review some selected forms of abstracted web traffic studied in the context of human navigation behavior.

Web logs. Studies on human navigation behavior on the web are traditionally concerned with navigation processes represented by web logs collected for example by servers, proxies and client programs. This includes work from the web usage mining community [123, 285, 467], which "focuses on techniques to predict user behavior while the user is interacting with the web" [285]. For example, Mobasher et al. [361] used access logs from a newsletter website for generating user profiles, and Liu et al. [326] and Meiss et al. [354] recommended news and rank web pages based on click log files, respectively. Furthermore, Agosti et al. [5] give an overview of applied web log analysis from web pages as well as digital library systems ranging from 1983 to 2011. One of the issues in this context is to derive abstract notions like users, server sessions, episodes, click-streams, and page views from the recorded logs [355, 467]. For example, Cooley et al. [122] and Munk et al. [369] list a set of pre-processing procedures and steps to derive more abstracts concepts from raw log data. On a more abstract level, some web log datasets also only provide aggregated information such as target and referrer counts [537].

Domains and abstractions. Often, work on human navigation behavior on the web focuses on certain domains or abstractions of log data. For example, most studies only had access to data from a certain set of web servers. Thus, they inherently were limited to navigation in a specific context, e.g., Liu et al. [326] used web logs from a news platform and Gündüz and Özsu [221] investigated on web logs from the NASA Kennedy Space Center as well as logs from the Metro Baltimore-Washington DC area.

While such studies usually aim at providing results for general web usage behavior, there exist many others focusing on certain domains studying and exploiting their specific characteristics. For example, Lee et al. [305] studied click streams on online stores and Bollen et al. [67] analyzed web logs from scientific publisher web portals building "maps

of science". Furthermore, current work on web navigation analysis focused on selected applications like navigation on information networks [300, 489] (like Wikipedia [373, 454, 519]), folksonomies [145, 490], social networks [51, 153], or ontologies [501, 502, 503].

In such specialized cases, it often makes sense to employ more abstract notions than actual page visits. For example in the case of Wikipedia, navigation between actual articles can be studied (cf., Wikispeedia [520] or Wikigame[17] [453, 489]) leaving out other pages like category overviews or the start page. Similarly, corresponding work on social bookmarking systems [cf. 145, 373] focuses on those web pages representing the basic entities of a folksonomy instead of considering search or overview pages. In this work, we also study the task-choosing behavior of workers on the crowdsourcing platform Microworkers [40] (cf. Chapter 10) where the users' interactions with the platform are also logged on a more abstract level storing task subscriptions directly instead of lower-level page interactions.

2.2.3. Modeling

The previous sections covered early work on web navigation as well as different domains and data sources for observing various behavioral aspects. In order to explain the observed data and to understand the corresponding underlying processes, many studies apply modeling [e.g., 78, 253, 365], i.e., building systems for explaining the observed phenomena. In this section, we cover important models for human navigation on the web focusing on work that is closely related to this thesis, i.e., which is concerned with trajectories as well as transition behavior. Specifically, we cover work applying Markov models, which are one of the most widely used approaches for modeling web navigation, as well as studies that employ the theory of information foraging, which models user behavior as a search process in an information environment.

2.2.3.1. Markov models and memory processes

Markov models are one of the most prominent model classes employed to represent online navigation behavior in the context of browsing trajectories and page transitions. In the following, we review a selected number of studies applying Markov models to research web navigation, and address the ongoing discussion of the appropriate memory-structure of navigational processes on the web. For more technical details and methodological extensions of Markov chains we refer to Section 3.2.

Applications. Descriptive and explorative work often studies transition probabilities between states of a Markov chain: for example between user actions on online public access catalogs [485], entities in folksonomies [145], or user behavior categories on social networks [51]. Further analysis approaches include the application of Markov chains and its corresponding extensions to clustering users by their behavior [e.g., 81, 406], or to studying contextual states employing *hidden* Markov models [84].

Also, a well-known application of Markov models is the PageRank algorithm [387], which uses a specialized Markov chain model for representing users surfing the web. In

[17] `http://thewikigame.com/`

particular, it models users as random surfers on the link structure imposed by web pages and ranks these pages according to the number of times each page is visited by these surfers. Besides the ranking capabilities of this approach (applied to improving search engines), Page et al. also estimated web traffic using this method. Furthermore, work by Eirinaki et al. [161] employed the page rank of web sites to impose prior probabilities on Markov chains for predicting user behavior. Generally, work on predicting distributional properties, next-clicks, or links is numerous [402, 429, 567, 571], mostly with the goal to enhance user experience, e.g., by enabling personalization.

Memory structures. Many of the previously mentioned studies try to go beyond first-order Markov chains which assume that the next page a user visits only depends on the current page. In other words, they explore higher order memory structures. Indeed, considering different memory structures has a long history. For example, Pirolli and Pitkow [402] found that, in their case, first-order Markov models performed best for predicting user behavior in the context of predicting link choices. However, there is an active discussion about the order of Markov chains to use when describing the memory structures involved in navigational processes on the web [cf. 111, 454]: While the first-order approach was often applied and confirmed by similar studies [321, 429], other work employed more complex memory structures extending the first-order Markov chain and reporting superior results. For example, Sen and Hansen [442] modeled navigation behavior on the web using Markov models of first order, second order, and a mixed variant, and found that second-order Markov models gave the best performance in their scenario of predicting session lengths and unique pages per session. Similarly, Borges and Levene analyzed variable-length Markov chains [71] and emphasized the importance of higher-order structures [70, 72] by investigating prediction accuracy of next-link choice and "summarization ability". Others built approaches combining different orders of Markov models to gain increased prediction accuracy and less model complexity [141, 571], or use tree structures in combination with varying-order models [146]. Finally, Chierichetti et al. [111] picked up on previous work and found cases where the (first-order) Markovian property does not hold. Recently, Singer et al. [454] tried to shed light into this ongoing debate about the depth of memory in web navigation by comparing different orders of Markov chains based on a variety of statistical evaluation measures. They specifically considered model complexity to mitigate overfitting effects [cf. 370], and, by doing so, showed that first-order Markov chains are the most justifiable choice for page-to-page navigation. At the same time, however, they also discovered that on a more abstract level, i.e., navigation over topics, first-order chains may not suffice to model navigation behavior. Note, that there is also work in the human mobility domain (as covered in Section 2.1) discussing similar issues [cf., 345].

2.2.3.2. Information foraging

Coming from a psychological background rather than a technical one [182], a very specific set of models has emerged around the theory of "information foraging" [401], which considers online navigation as a search process. In particular, this theory assumes that humans searching for information on the web behave similar to animals searching for

food. Consequently, the models based on information foraging all incorporate a variant of *information scent*, which "is the subjective sense of value and cost of accessing a page based on perceptual cues" [108]. There are several applications employing the notion of information scent. For example, Chi et al. [108] described two algorithms: one for simulating the paths of web users and another for inferring the information need responsible for an observed path. The theory of information foraging and the concept of information scent were further studied, extended, or implemented by a variety of researchers resulting, for example, in the CoLiDeS model by Kitajima et al. [280], the MESA model by Miller and Remington [358], or the SNIF-ACT models by Fu and Pirolli [182]. Furthermore, the ScentTrails approach [385] highlights hyperlinks to indicate paths to *nicely smelling*[18] search results. Another concept closely related to information scent is the notion of "orienteering" [381]. Applied to the world wide web [481] this corresponds to the process of starting with a set of rather general pages and then "using both prior and contextual information to close in on the actual information target, often in a series of steps, without specifying the entire information need up front". A quite similar description of user navigation in the context of searching for information is "berry-picking" [38], which refers to the process of "bit-at-a-time retrieval". In other words, user identify useful bits of information and select references while searching the web step-by-step and constantly adjusting their queries. Indeed, subsequent work [524] found that following trails to satisfy one's information need (instead of directly arriving at the destination of a search) has value with regard to relevance, topic coverage, topic diversity, novelty, and utility.

2.2.4. Regularities and patterns

Besides modeling the overall processes of user trajectories and transition behavior on the web (as covered in Section 2.2.3), there are also studies about online navigation analysis which focus more on discovering regularities and patterns within navigational data. In the following, we cover several findings from such studies. The variety of results across several dimensions — including structural and temporal aspects, navigation types and strategies, as well as semantics — implicates the complexity of human behavior and illustrates the need for further studies in this subject which we contribute to in this thesis.

2.2.4.1. Structural and temporal aspects

Human navigation behavior exhibits strong structural as well as temporal regularities. In the following, we address both dimensions.

Structural aspects. On a structural level, Huberman et al. [253] found that the frequency of website hits follows a "Zipf-like distribution", i.e., there is a small number of highly visited pages with a power-law governed drop-off towards pages which are visited less often. Similar distributional properties were observed to hold for the length of navigation trails [253, 316]. This observation has been called the "universal law of web surfing" and was confirmed for mobile web navigation as well [225]. Furthermore,

[18] Checking ... all good!

depending on the characteristics of the corresponding trails, Catledge and Pitkow [92] discovered that web pages can be characterized according to their usage patterns.

Temporal aspects. On the temporal level, Agarwal et al. [4] and Matsubara et al. [348] observed certain temporal click and navigation patterns. Such patterns especially include cyclic activity over time as also found by Zaiane et al. [550]. Agarwal et al. also observed differing activity levels of users from the US and international users during day and night. Furthermore, they found an activity decay in visitation frequencies over time which they attributed to repeated exposure to the same content. Also, Wu and Huberman [533] observed similar attention decays attributing them to the novelty of a story which decreases over time. In the same direction, Zaiane et al. [550] detected a tendency of users of a collaborative teaching and learning environment to initially explore the features of the system while becoming more and more focused over time. A similar convergence effect on the smaller, session time-scale can also be observed when playing the game Wikispeedia [519] where individuals first navigate to general articles and get more specific as the search progresses. Also studying temporal aspects of web navigation, Matsubara et al. [348] analyzed the temporal evolution of different topics such as "media" or "business". They found topical preferences by the time of the day and depending on weekdays. For example, "food"-related URLs are more frequently visited right before dinner time and "communication"-related topics are more common in the late evening. Furthermore, on a semantic level, Yang et al. [545] studied so called "progression stages" along navigation trails on the web. In particular, they proposed an approach to identify semantic units, such as "US presidents" or "countries", which users progress through during a session. Finally, in the context of navigation prediction and personalization, models often incorporate temporal dependencies [1, 226].

2.2.4.2. Navigation characteristics

Besides the general structural and temporal aspects covered in the previous section, there are also specific strategies users apply when navigating the web. In the following, we cover several of such strategies in various application domains including general navigation characteristics, subprocesses of navigation, backtracking and revisitation patterns, as well as some other strategies observed for human navigation behavior on the web.

Navigation characteristics. Navigating the web, i.e., "browsing", is generally made up of following links and backtracking [517] and has been studied as early as 1998 by Huberman et al. [253]. They observed that users proceed to another page as long the "the value of the current page exceeds a threshold". Furthermore, Weinreich et al. [517] found that following links is the most common "navigation action" (as opposed to, e.g., backtracking) and, thus, that web navigation is a "rapidly interactive process" with regard to the frequency of clicks. Benevenuto et al. [51] confirmed the prevalence of browsing on the social network Orkut where it made up more then 90% of the users' activities. One possible reason for this is given by White and Huang [524] who found that, in an information environment such as the web, browsing in general is a practice worth pursuing since — in contrast to directly accessing a page — following links (and thus following a search trail) is a strategy often more useful with regard to topic coverage,

diversity, novelty, and utility. Similarly, Downey et al. [148] concluded that pursuing links is especially useful when the user has an "information goal" that is rare, i.e., not commonly searched for.

Sub-processes. Navigating the web it is not a homogeneous process. Instead, it subsumes a variety of characteristics. For example, some of the earliest work on web-like structures has found different classes of navigation behavior. That is, Catledge and Pitkow [92] confirmed and characterized different navigation strategies, namely "serendipitous browsing", "general purpose browsing" and "search browsing", which have been previously found in different contexts by Carmel et al. [87] and Cove and Walsh [125]. Similarly, White and Drucker [523] categorized web users into "explorers" and "navigators", whereas Choo et al. [117] distinguished between various behavior modes and moves in the context of information seeking. More recently, Phoa and Sanchez [397] also established three user groups relevant for their approach: accidental users, regular users, and power users.

Backtracking and revisitation. As previously mentioned, Weinreich et al. [517] argued that backtracking is less common with more modern browsers. However, Scaria et al. [432] found that for certain tasks, i.e., in this case for navigating Wikipedia, it plays an important role. There are also quite a few studies analyzing revisitation patterns in general [3, 383, 480]. For example Obendorf et al. [383] distinguished between three revisitation patterns, namely short-, mid-, and long-term revisits, which represent the notions of backtracking/undoing, reutilizing/observing, and rediscovering, respectively.

Other strategies. There is also a variety of other strategies and characteristics of web navigation. For example, users tend to follow the already mentioned information scent (Section 2.2.3.2), they often employ their context knowledge instead of exclusively using keyword-based search [481], they leverage semantic relations (also see Section 2.2.4.3), or show a tendency to stray from shortest paths [519]. Thus, overall, individual navigation strategies can "differ dramatically and are strongly influenced by personal habits and type of site visited" [383]. We further emphasize this inherent heterogeneity in Section 2.2.5.

2.2.4.3. Semantics

Semantics, i.e., the similarity or relatedness of words and concepts, is an important factor when analyzing human navigation behavior on the web. Indeed, as argued below, it was shown that semantics are an integral part of navigation strategies on the one hand, and, on the other hand, the semantic information inherent to the corresponding navigation trails can be recovered using appropriate methods.

Semantics are part of navigation strategies. Many of the models and strategies employed by humans navigating the web are in some form based on semantics. For example, the strategy users employ to find a certain article by following links on Wikipedia is governed by semantic similarity [41, 519]. Related notions can be found for navigation on online folksonomies [373], or task choosing behavior on crowdsourcing platforms [40]. Also, the concept of information foraging and information scent strongly relies on semantics "to account for a user's efficiency in traversing a Web structure" [265]. Similarly, progression stages of user session studied by Yang et al. [545] are characterized by semantic homogeneity.

Brumby and Howes [77] also incorporated semantic similarity in their model for web navigation and early work by Pierce et al. [398] found that for menu item selection semantic relatedness of the menu items plays an important role. On a more structured level, there is even work on incorporating ontologies to model navigation processes [300]. Thus, overall, semantics are present on many layers of the navigational processes we can observe on the web.

Navigation behavior yields semantics. The inherent semantics of navigational processes on the web also become clear when considering application oriented work, e.g., by Chalmers et al. [98] or Bilenko and White [62]. In particular, Chalmers et al. used navigation path information for calculating (semantic) similarity between URLs and Bilenko and White used information about the browsing activity after a search as a feature "in learning to rank for Web search", i.e., for improving the results a search engine. Other studies employed search queries and information about the subsequently chosen web pages to extract rich semantic relations by building folksonomic structures [31, 57, 286]. Emphasizing the inherent semantic characteristics of human navigation on the web even more, current work took the approach of calculating semantic relatedness between words and concepts directly from navigation logs, e.g., in the form of word embeddings [536]. Prominently, Dallmann et al. [132], Niebler et al. [374], and Singer et al. [455] extracted semantic relatedness from navigation trails on Wikipedia.

2.2.5. Heterogeneity

Studies on human navigation behavior are part of the general research area of web usage mining [123], which is concerned with "user interactions with Web resources on one or more Web sites" [324]. There, one of the main applications of analyzing human navigation is learning user profiles and personalized user models [285]. This already indicates that the underlying navigation processes are inherently not homogeneous and differ greatly from user to user. The regularities and patterns discussed in the previous section, like the three types of browsing (general, serendipitous, and search), already give some prominent examples of this fact. In the following, we first review already covered components of human navigation behavior from Section 2.2.3 and Section 2.2.4, emphasizing the heterogeneous nature of navigational processes on the web. We follow up with several other studies further illustrating the existence of sub-processes in browsing and their differences based on the websites and platforms being used, demographics and user properties, as well as automatically generated clusters.

Browsing types, temporal factors, and topics. First off, we revisit some studies listed in Section 2.2.4.1, emphasizing the inherent heterogeneous nature of navigation on web-like structures [87, 92, 117, 125, 397, 523]. Most prominently, we cited Catledge and Pitkow who found browsing to have three sub-components, namely "serendipitous browsing", "general purpose browsing" and "search browsing". We have also already discussed temporal aspects of user navigation exhibiting heterogeneous properties including diverging temporal activity patterns for different countries [4] as well as general topic shifts over time [348]. On a session-level, Yang et al. [545] studied evolving topic stages resulting in sequences of semantic units and West and Leskovec [519] and Zaiane et al. [550] found

a shift from an exploration phase to more focused utilization of the respective information environment. Finally, beyond singular patterns, models for navigation processes also often incorporate temporal dependencies [1, 226], e.g., for prediction or personalization.

Platforms and usage aspects. In addition to these general behavioral patterns, navigation behavior also differs depending on the service being used. Prominent examples are social networks. In particular, Benevenuto et al. [51] and Schneider et al. [436] both compared different social networks and found deviating characteristics in activities and session properties. Taking a different point of view, Dunn et al. [153] investigated the difference between online social networks (OSN) and search engines. There, the results showed that users tend to stay on OSNs longer than on search engines and navigate to less popular and different types of web pages from OSNs.

Demographics and user properties. In addition to the previously covered studies, there are also other results which emphasize the heterogeneity of human navigation on the web. In particular, the notion of demographics has been recognized to play an important role when considering web navigation, especially with regard to optimizing user experience [97, 353]. For example, one of the most straight-forward influence factors on human performance when searching and browsing the web is the age or the gender of individuals [340, 350, 353, 500]. For example Marchionini [340] found that a difference in search behavior depending on age (on a full-text electronic encyclopedia), and in the work by Mead et al. [353] older adults were found to be "less efficient and somewhat less successful than younger adults when searching a 19-page Web site for the answers to specific questions". Extending such studies, Goel et al. [208] analyzed an extensive set of user attributes including education, gender, income, or age, and even found that properties like ethnicity and income can be inferred from browsing histories. Furthermore, literacy was discovered to influence navigation behavior by Zarcadoolas et al. [553] who identified "specific navigational issues" that present barriers to low-literate adults. Similarly, Stanney and Salvendy [471] studied how individuals who have a low ability to perceive spatial patterns can be supported when navigating the web. Juvina and Oostendorp [264] incorporated and extended these studies in their work and found that domain expertise, spatial ability, working memory, motivation, and interest are important determinants of task outcomes and thus ultimately influence navigation behavior.

Clusters and components. The previously mentioned studies indicate certain categories of web navigation characteristics. In web navigation research, or web usage mining, a common understanding is that establishing user groups can help to infer "user demographics in order to perform market segmentation in e-commerce applications or to provide personalized Web content to the users" [467]. To this end, a wide array of clustering algorithms exists [183, 221, 460, 475, 506] also covering other applications such as improving page performance or detecting spammers. For example, Wang et al. [506] analyzed the difference between clickstreams of real users and fake accounts on the Chinese social network "Renren". Besides finding different characteristics in the number of sessions per user, average session length in seconds and clicks, or the average inter-arrival time, they derived behavioral clusters based on navigation traces which they used to classify fake accounts. In general, clustering user navigation traces on the web has a long

history. For instance, as early as 1996, Yan et al. [541] already clusterd user behavior based on access logs and claimed that the corresponding clusters were "not apparent from the physical linkage of the pages, and, thus would not be identified without looking at the [web] logs". Along the same lines, Cadez et al. [81] and Kumbaroska and Mitrevski [293] clustered users by navigation paths using a model based approach (instead of a distance based one). They visualized the clusters, finding different page category preferences. Another approach (similar to MixedTrails in Chapter 4) specifically employed a mixture of first-order Markov chains [81]. There are also clustering approaches which allow to interpret the clusters in order to detect, classify or explain interesting user behavior. For example, Barab et al. [34] established four classes of navigational performance ("models users", "disenchanted volunteers", "feature explorers", and "cyber cartographers"), and Wang et al. [507] found unexpected or interesting sub-clusters such as inactive users or users with hostile behavior.

2.2.6. Discussion and relation to this work

Analogously to navigation on the web (cf. Section 2.1), we covered several aspects of *web navigation* analysis with respect to early work, data, regularities and patterns, as well as heterogeneity. Especially Section 2.2.4 and Section 2.2.5 emphasized the inherent heterogeneity of human navigation behavior. This means that there is no single underlying process explaining the observed data. Rather, there is a multitude of factors which may influence navigation strategies and characteristics.

While we covered some work which studies these factors, there are no general methods dedicated to analyzing and understanding such heterogeneity. To this end, this thesis introduces the MixedTrails (Chapter 4) and the SubTrails (Chapter 5) approach, as well as several analysis tools (Chapter 6). In particular, MixedTrails allows for comparing understandable hypotheses (from theory or intuition) about heterogeneous sequence data, and SubTrails, enables the discovery of interpretable subgroups of sequence data with exceptional transition behavior.

Additionally, Section 2.2.2 illustrated that web navigation analysis can cover a large array of data types, abstractions, and application domains. These domains ever expand and new application domains or systems emerge for which novel kinds of navigation characteristics have to be investigated. In this context, Agosti et al. [5] mention a set of future trends for web log analysis, including research on social bookmarking systems (e.g., Delicious[19] and BibSonomy[20]), as well as the increasing focus on online encyclopedias like Wikipedia[21]. Matching these predictions, in this thesis, we contribute a study on the social bookmarking system BibSonomy (see Chapter 9), and an exemplary case study on Wikipedia navigation (see Section 11.2). Additionally, we add to the already large array of different application domains by analyzing task-choosing behavior on the crowdsourcing platform Microworkers (Chapter 10), as well as music listening trails (see Section 11.3).

Overall, this thesis builds on the previous work covered in the sections above and

[19] `https://delicious.com`
[20] `https://bibsonomy.org`
[21] `https://wikipedia.org`

contributes strongly to understanding of human behavior on the web by introducing novel methods for analyzing heterogeneous navigation processes, and provides corresponding insights for researchers and practitioners on several interesting application domains.

3. Methodological foundations

In this thesis, our goal is to analyze and understand human navigation behavior, be it in a geo-spatial context or on the web. To this end, we employ a specific set of underlying concepts which we introduce in this section.

In particular, we first introduce the notion of discrete navigation processes (Section 3.1) as the general setting in which we study human navigation behavior. That is, we argue to use sequences over a discrete state space as the unifying framework (Section 3.1.1), and emphasize the necessity of background information to cope with the multitude of underlying aspects and different contexts of navigational processes (Section 3.1.2). Thus, using background information to explain observed discrete navigation data is the main idea of this thesis. In the following section, we then review several methodological approaches applied throughout this work, i.e., Markov chains (Section 3.2), the HypTrails approach (Section 3.3.2), as well as exceptional model mining (Section 3.4.1). Specifically, we start with defining Markov chains which represent the core concept employed by this thesis. Then, we cover HypTrails, a Bayesian approach for comparing hypotheses about human navigation trails, which we extend to cope for heterogeneous hypotheses (cf. Chapter 4) and apply extensively in our case studies (cf. Part III). Finally, Exceptional model mining is the basis for our second methodological contribution called SubTrails (Chapter 5), a method for discovering subgroups with exceptional transition behavior. For a general overview on notations in the context of discrete navigational data and Markov chains, we refer to Table A.1.

3.1. Data for understanding discrete navigation

In this thesis, we aim to understand human navigation behavior in various application domains, i.e., geo-spatial navigation, and navigation on the web (cf. Chapter 1). To study these domains in a unified manner, we apply the concept of *discrete navigation behavior*, and employ *background information* to explain the observed data. In this section, we first introduce the notion of discrete navigation data (Section 3.1.1), and then give an overview of relevant background information (Section 3.1.2).

3.1.1. Discrete navigational data

At first glance, geo-spatial navigation and navigation on the web — which we study in this thesis — appear to be fundamentally different instantiations of human navigation behavior. That is, generally, geo-spatial navigation is part of our physical experience while navigation on the web is a virtual process. Also, on a technical level, the former is embedded into the continuous space of the real world while navigation on the web is

$D = \{p_{\text{alice}}, p_{\text{bob}}\}$
$S = \{s_{\text{central}}, s_{\text{times}}, s_{\text{statue}}, s_{\text{empire}}\}$

$p_{\text{alice}} = [s_{\text{central}}, s_{\text{times}}, s_{\text{statue}}, s_{\text{central}}, s_{\text{times}}]$
$p_{\text{bob}} = [s_{\text{statue}}, s_{\text{empire}}, s_{\text{central}}, s_{\text{times}}]$

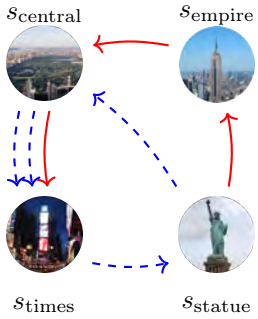


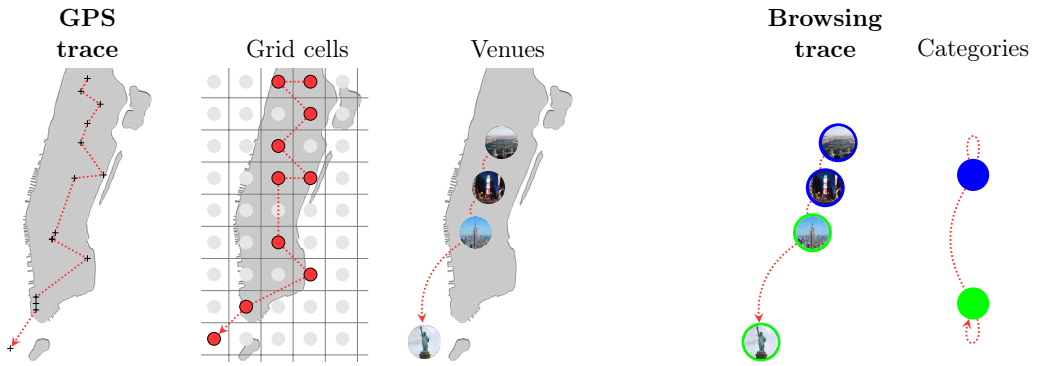
Figure 3.1.: Example for discrete navigational data. This figure shows a path dataset D based on state space S of locations in New York City (Central Park, Times Square, Statue of Liberty, Empire State Building). The dataset contains two paths, one by Alice (p_{alice}) and one by Bob (p_{bob}). Alice starts at Central Station s_{central} and Bob at the Statue of Liberty s_{statue}.

restricted to a discrete set of web pages (cf. Figure 1.1 in Chapter 1). However, many studies in the context of geo-spatial human mobility do not directly study navigation processes in such a continuous manner. Instead, they often use data from call detail records which are restricted to a *discrete* set of cell towers [e.g., 214], or they study check-in sequences from location-based social networks where navigation is restricted to a*discrete* set of venues [e.g., 378]. And even when GPS tracks are analyzed, which are not restricted to a discrete state space, these tracks are often *discretized*, either in a preprocessing step [e.g., 202, 271] or inherently by the applied methodology [202]. Overall, discretization is a natural process since human behavior in a geo-spatial context is dictated by concrete places, venues, or activities which are often bound to certain locations. Thus, while we may loose some information on local details of human mobility, it is reasonable to describe the underlying processes based on a discrete state space. For this thesis, this allows us to formulate a general framework based on discrete navigation which covers geo-spatial human behavior as well as navigation on the web using the same methodology.

Definition 1 (Discrete navigational data)**.** *In this work, we consider navigational behavior on a finite state space* $S = \{s_1, \ldots, s_n\}$*. The navigational data we observe on such a state space consists of a (possibly very large) set of* paths D *(also called* sequences *or* trails*) generated by a set of* individuals U *(also referred to as* users*). Each path* $p \in P$ *is a sequence of states* $p = [s_{\tau_1}, \ldots, s_{\tau_{n_p}}]$ *where* $n_p > 1$ *is the number of states visited by path* p*. A path can include each state several times. A path* p *can also be represented as a sequence of* transitions $p = [t_{\tau_1,\tau_2}, t_{\tau_2,\tau_3}, \ldots, t_{\tau_{n_p-1},\tau_{n_p}}]$ *where* $t_{\tau_i,\tau_{i+1}} = (s_{\tau_i}, s_{\tau_{i+1}})$ *represents a pair of states. We also write* $t_{i,j}$ *for a transition from state* s_i *to* s_j*.*

Example 1 (Discrete navigational data)**.** *An example of a path dataset* D *is shown in Figure 3.1. It depicts a state space* S *of locations in Manhattan and two paths with individual lengths by different users.*

State spaces. State spaces, on for discrete navigation data, can be defined in a very flexible manner. That is, there are many variants which represent different levels of



(a) Geo-spatial navigation **(b)** Online navigation

Figure 3.2.: An illustration of different state spaces for human navigation. The figure shows raw data as well as different state spaces for human navigation behavior in a geo-spatial context (a) as well as on the web (b). In this work, we focus on methods for discrete, finite state spaces. Thus, for the continuous geo-spatial case, navigation traces, e.g., from a GPS, need to be discretized. (a) shows a grid-based and a venue-based discretization approach. For online navigation, web pages form a natural state space. However, different levels of abstraction may still make sense. (b) shows how navigation on articles on Wikipedia can be abstracted to navigation over categories. Choosing a discretized settings allows for applying the same methodology to both settings (geo-spatial as well as online navigation), and investigating different abstraction layers results in a more intricate understanding of human navigation behavior in general.

navigation on the same underlying data. Consider Figure 3.2 where we illustrate different state spaces with corresponding paths for geo-spatial as well as online navigation.

Figure 3.2a shows a GPS trace from a user exploring Manhattan. To apply the discrete setting this continuous trace needs to be discretized, i.e., each position is mapped to one of a finite set of states. There are many different variants to do this. The figure shows two of them: discretization i) by using a grid where states correspond to grid cells the users pass through, or ii) by considering venues where each state represents a venue that users may visit. Further variants of discretizing the geo-spatial space include, for example, other semantic discretizations (like venues) obtained from background information (e.g., tracts or neighborhoods) or using clustering approaches [e.g., 202, 312]. It is apparent that the choice of the state space strongly depends on the data as well as conceptual level of human mobility at interest. The grid-based approach may be more useful for fine-grained navigation behavior while a venue based approach focuses on the more semantic level of navigation between meaningful places. Both tackle the process of human mobility on different levels thus each helping to understand different facets of human navigation behavior. See, for example, Chapter 7 where we employ grid-based as well as tract-based discretization. For an example of a venue based study, we refer to Noulas et al. [378].

Navigation on the web poses similar challenges. Figure 3.2b shows a browsing trace, e.g., from a user navigating articles on Wikipedia [cf. 519]. There, the depicted "browsing trace" can be directly used in a discretized navigation study by considering the visited websites as states. However, even though a natural state space arises in the context of online

navigation, there are still different levels of navigation to be studied. For example, it may be interesting how users navigate between *categories* instead of articles on Wikipedia. Figure 3.2b shows the corresponding abstraction. Along the same line, one of the earliest articles on online navigation [485] studied a state space of *user actions* instead of using the actual pages as underlying concepts.

It becomes apparent that the state space to study is highly flexible and can help to gain insights into many facets of human navigation behavior. Note however, that the choice of a specific set of states is not a trivial one. For example, studies in the geo-spatial context have recognized the "modifiable areal unit problem" (MAUP) [257, 534] which refers to the problem of choosing appropriate areal units in order to study aggregate statistics. MAUP consists of two effects: the scale and the zoning effect. The former refers to variations due to the level of aggregation to be studied, e.g., by altering the size of the cells in a grid based discretization approach. The latter refers to variations due to the choice of which data points are aggregated, e.g., by altering the coordinates where the boundaries of a grid are placed. The same effects can be observed when aggregating fine-grained state spaces into more coarse ones as has been done in Figure 3.2b. This makes the MAUP also relevant for the field of web navigation.

Nevertheless, actual observations (such as GPS traces or fine-grained views on navigation on the web) can and need to be abstracted to be able to efficiently formulate hypotheses about human navigation. Thus, we need to make sure to carefully choose an appropriate state space in order to produce results that match the facets of human navigation behavior to be analyzed.

Summary. In this work, we choose a discretized setting for analyzing navigation processes. This allows us to study a very diverse range of different levels and facets of human navigation behavior using the same methodology. Exploiting this fact, we develop versatile, novel methods and tools to analyze navigation based on first-order Markov models which generally require a discrete state space, cf. MixedTrails (Chapter 4) or SubTrails (Chapter 5). We also leverage the tight coupling of geo-spatial and online navigation (cf. Figure 3.2). For example, the venues a user visits while exploring Manhattan (Figure 3.2a), may also visit the corresponding articles on Wikipedia (Figure 3.2b). For example, in Chapter 7 we incorporate knowledge about visitation patterns of venues on Wikipedia into navigation hypotheses explaining geo-spatial navigation processes.

3.1.2. Background data

Independent of the method, it is evident that — in addition to data containing navigation paths of users over a set of states — further information, i.e., *background data*, is needed to explain navigational behavior. This specifically includes properties of *states*, *users*, as well as *paths* and their individual *transitions*. These properties can be drawn from a multitude of sources. They may materialize as simple binary, categorical or real-valued attributes. However, they may also incorporate complex relations given by background information such as semantics [e.g., 373] or ontologies [e.g., 300].

In this section, we give a *non-exhaustive* list of possibly relevant background information when modeling human navigation behavior, e.g., for formulating specific hypotheses about

human mobility (cf. Chapter 7) or web navigation (cf., Chapter 9 or Section 11.2). For more information on additional factors which may influence human navigation behavior, we also refer to the sections covering respective related work (e.g., Sections 2.1.4, 2.1.5, 2.2.4 and 2.2.5).

States. In discrete navigation processes, states can have very specific properties. For example, categories of states play an important role: in the geo-spatial context, we may consider venue categories such as *public transport* for the Central Station (in New York) or *touristic* for the Statue of Liberty. Then, one hypothesis may state that transitions from public transport hubs will likely have their destination at areas with many office buildings, while transitions from touristic venues will favor other touristic destinations. Similarly, state categories also play an important role in web navigation. For example, in Chapter 9, we consider the website categories: resource, user, and tag pages.

Besides these straightforward examples, state properties can also include information on more complex procedures: For states in the geo-spatial context, we may consider to embed venues into its geographical context such as transportation networks, e.g., in order to derive actual vs. effective distance. In the context of navigation on the web, similar distance measures may be derived based on semantic similarity of states or ontologies embeddings (see West and Leskovec [519], Lamprecht et al. [300], or Chapter 10).

Users. Similar to states, individual users can be associated with properties which strongly influence their navigation behavior. Again, categories play an important role. For example, in the geo-spatial context, we may consider tourists vs. locals (cf. Chapter 7) and formulate a hypothesis that says that tourists are more likely to move towards touristic locations than locals. And in the context of web navigation, in Chapter 9, we study several user groups, e.g, based on gender or usage patterns. We have covered further examples of the influence of user categories in Sections 2.1.5 and 2.2.5.

Similar to states, users can also be described using more complex processes. For example, we may derive properties based on interactions or sentiments on social networks, such as friendship relations or emotional ties [116].

Paths. Besides states and users, individual paths (as a collection of transitions) may also exhibit specific properties. One of the properties of paths which can be used to explain the observed sequence of states is its purpose or corresponding incentives. For example, in the geo-spatial context, the purpose of one path may be to get to work, while another path is the result of a leisurely stroll through a park. The sequence of states and the probabilities of transitions will most likely be vastly different [e.g., 435]. Other interesting path properties may include the length of the path, or the time required for completing it. Similarly, for online behavior there are differences depending on the purpose of the navigational process, e.g., browsing vs. searching [e.g., 92].

Transitions. On a more fine-grained level, each individual transition within a path also exhibits inherent properties, such as the mode of travel [e.g., 483] (in the geo-spatial context), the start and stop time, or the position within the path (see, e.g., West and Leskovec [519] or Section 11.2 in the context of search behavior on Wikipedia).

Time. Navigational processes are strongly intertwined with time. For example considering navigation on Wikipedia, articles and links are constantly created or deleted. This results

in differing availabilities depending on the time of a transition. Similarly, in Chapter 10, we consider tasks-choosing in a crowdsourcing environment where the corresponding campaigns are only available for limited amount of time. In a geo-spatial setting, touristic venues may only attract tourists during their opening hours. Also, the same user may visit a city for a business meeting at one time and be a tourist at the other, which may result in strongly differing navigation paths. Finally, navigational processes in a geo-spatial setting strongly differ when comparing rush hours and night time, or weekdays and weekends.

History. Navigational behavior of an individual can change within a single path. For example, in the geo-spatial setting, a tourist may more likely visit locations which are close to subway stations if she has walked the first half of the day, while a tourist who started with a bus tour may still be more inclined to walk to places not reachable by another mode of travel. In Chapter 4, we consider a similar example in a synthetic setting, where "walkers" choose from red and blue states as their next destination depending on the history of colors of the states they have already visited.

Semantics and knowledge. As indicated above, *semantic relations*, e.g., between states (web sites, places, etc.), can be helpful in explaining human navigation. In particular, in this work, we use the notion of *semantics* in many of our studies to explain observed behavior. For example, we show that crowdsourcing workers prefer to work on semantically similar tasks. Also see Chapters 9 and 10 and Section 11.2 for further application areas. In those cases, we use the rather basic notion of semantic similarity based on cosine-distance between TF-IDF representations of textual descriptions.

However, this is a very limited definition of semantics. That is, the notion of semantics is often used, especially in the context of the semantic web [60, 337], and can represent more detailed concepts. For example, information such as *article1 Cites article2* or *category1 IsSubfieldOf category2* may help to explain behavior on publication management systems like BibSonomy (cf. Chapter 9). There is a wide range of work, defining and generalizing this notion to advanced structures like ontologies and knowledge bases [468].

We believe that such knowledge representations can be used to formulate intricate behavioral hypotheses, thus, further advancing the understanding of human navigation. However, this task is trivial and is out of the scope of this work.

Summary. In this section, we have listed a small portion of the wide variety of background information which is essential for explaining human navigation behavior, ranging from state and user properties, over path and transition properties, to their inherent dynamics. We use many of the listed properties in our case studies and, thus, refer to Part III for detailed examples.

3.2. Markov chain modeling

Markov chains, named after a study by A. A. Markov in 1913 [341], were used to model stochastic processes in a wide variety of domains, such as speech recognition [410], bio-informatics [272, 470], or weather prediction [186]. In particular, this also includes human navigation on the web [e.g., 454] and geo-spatial mobility [e.g., 189]. We also build upon

Markov chains in this thesis: We introduce novel methodology for understanding human navigation and mobility (Part II), and study various application scenarios (Part III).

In this section, review the corresponding methodological basics on Markov chain modeling. That is, we first formally define Markov chain models, and then — in order to give a broader overview — cover some applications and extensions relevant in the field of human navigation and mobility.

3.2.1. Markov chains

In this thesis, we use Markov chains to model discrete navigational data as introduced in Section 3.1.1. In the following, we first define *Markov chain* models formally[1], and secondly introduce the construct of *transition count matrices* which are required for practically applying Markov chains in several use cases.

Definition 2 (Markov chain)**.** *Given a finite, discrete state space $S = \{s_1, \ldots, s_n\}$ (cf. Definition 1), a Markov chain models finite sequences of random variables of variable length $X_1, X_2, \ldots$ with values from state space S. These variables abide by the* Markovian property*, i.e., the next state is only dependent on the previous state. Formally, we write:*

$$\Pr(X_{\tau+1} = s_j \mid X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_\tau = s_{i_\tau}) \tag{3.1}$$
$$= \Pr(X_{\tau+1} = s_j \mid X_\tau = s_{i_\tau}) \tag{3.2}$$
$$= \Pr(s_j|s_i) \tag{3.3}$$
$$= \theta_{i,j} \tag{3.4}$$

Here, $\Pr(s_j|s_i)$ and $\theta_{i,j}$ are short notations for the transition probability $\Pr(X_{\tau+1} = s_j \mid X_\tau = s_i)$ from state s_i to s_j. The transition probabilities between all pairs of states (s_i, s_j) can be subsumed in a stochastic matrix $\boldsymbol{\theta} = (\theta_{i,j})$, i.e., each row sums up to 1: $\sum_j \theta_{i,j} = 1$. Thus, overall a Markov chain is defined by a state space and the corresponding transition probabilities: $\mathcal{M} = (S, \boldsymbol{\theta})$.[2]

Example 2 (Markov chain)**.** *We use the same state space as from Figure 3.1 in Section 3.1.1 to to construct an example of a Markov chain: The state space S consists of a set of venues in Manhattan, e.g., the Central Park $s_{central}$, the Times Square s_{times}, the Statue of Liberty s_{statue}, and the Empire State Building s_{empire}. People moving between these venues (either offline, in the case of people, e.g., exploring New York as tourists, or online, in the case of people browsing, e.g., Wikipedia) exhibit sequences of states instantiating a sequence of random variables $X_1, X_2, \ldots$. Figure 3.3 visualizes a corresponding state space S with (arbitrarily chosen) transition probabilities $\boldsymbol{\theta} = (\theta_{i,j})$. It also shows user*

[1] While there are other variants of Markov processes (e.g., continuous time, or continuous space models) [422], we focus on a finite, discrete state space and discrete time. We refer to this specific notion of Markov processes as a *Markov chain* (model).

[2] Note that other work explicitly models the probability of the first state (e.g., Singer et al. [454]). However, this can be implicitly included by introducing a special start state for each sequence with appropriate transition probabilities. Similarly, when sampling from a Markov chain, it can make sense to model a stop state which signifies the end of a sequence.

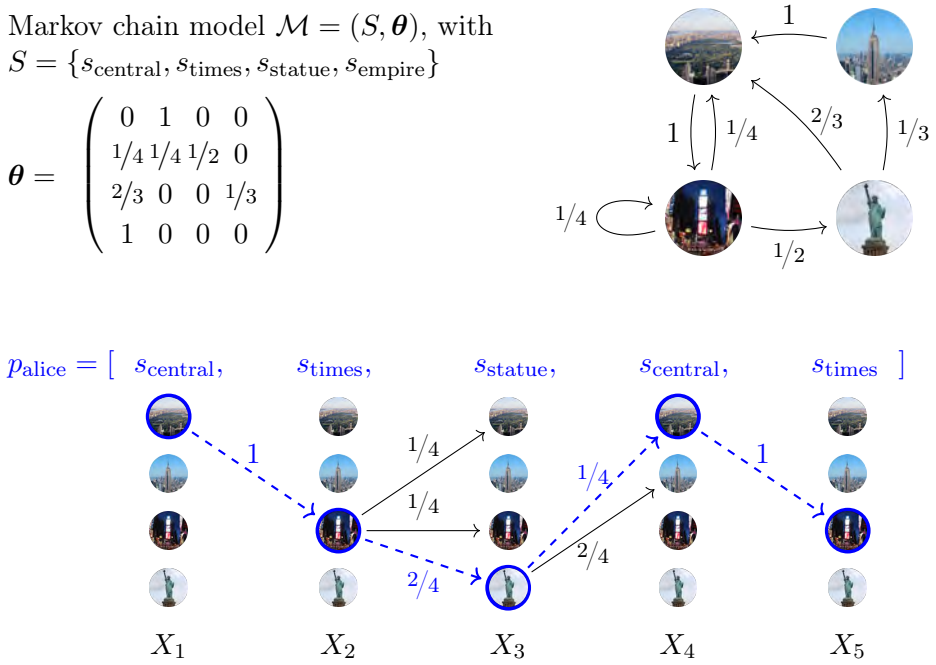


Figure 3.3.: An example of a Markov chain. We show Markov chain $\mathcal{M} = (S, \boldsymbol{\theta})$, with state space $S = \{s_{\text{central}}, s_{\text{times}}, s_{\text{statue}}, s_{\text{empire}}\}$ and transition probabilities $\boldsymbol{\theta}$. The Markov chain is visualized as a graph (right), where transition probabilities of 0 are omitted. At the bottom we illustrate how user Alice generates a path p_{alice} by starting at state s_{central} and randomly choosing subsequent states based on the given transition probabilities $\boldsymbol{\theta}$.

"Alice" generating a path by starting at state $s_{central}$ and randomly choosing subsequent states based on the transition probabilities given by the Markov chain.

In this thesis, we encounter Markov chains in two scenarios: i) for comparing hypotheses about human navigation (cf. HypTrails in Section 3.3.2 and our own approach MixedTrails introduced in Chapter 4), and ii) in the context of our pattern mining approach for discovering subgroups with exceptional transition behavior (Chapter 5). In both cases, transition counts matrices play an important role.

Definition 3 (Transition count matrix)**.** *Given a path dataset $D = [p_1, \ldots, p_m]$ as formalized in Definition 1, with paths being represented as sequences of transitions $p = [t_{\tau_1,\tau_2}, t_{\tau_2,\tau_3}, \ldots, t_{\tau_{n_{p-1}},\tau_{n_p}}]$, the corresponding transition count matrix $\boldsymbol{T}_D$ is given as:*

$$\boldsymbol{T}_D = (n_{i,j}) = \left(\sum_{p \in D} \sum_{t_{i,j} \in p} 1 \right) \tag{3.5}$$

Here, $\boldsymbol{T}_D$ is a matrix where each entry $n_{i,j}$ represents the number of transitions observed between each pair of states (s_i, s_j) over all paths $p \in D$.

Example 3 (Transition count matrix)**.** *As an example, given the path dataset* $D = \{p_{alice}, p_{bob}\}$ *from Figure 3.1 in Section 3.1.1 with*

$$p_{alice} = [s_{central}, s_{times}, s_{statue}, s_{central}, s_{times}] \tag{3.6}$$

$$p_{bob} = [s_{statue}, s_{empire}, s_{central}, s_{times}], \tag{3.7}$$

we get the following transition count matrix $\boldsymbol{T}_D$*:*

$$\boldsymbol{T}_D = \begin{pmatrix} 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{3.8}$$

As mentioned earlier, such transition count matrices play an important role with regard to the methodology applied throughout this work: i) For comparing hypotheses about human navigation, calculating the *likelihood* of a specific parameter instantiation of a Markov chain is essential. ii) For mining subgroups with exceptional transition behavior, we fit Markov models to various subsets (e.g., old vs. young people) of a path dataset in order to judge their particular navigation characteristics. In the following, given a path dataset D, we briefly cover the corresponding aspects of *calculating the likelihood* of a parameter instantiation, and of *fitting* a Markov model to the data.

Calculating likelihood. If a Markov chain $\mathcal{M} = (S, \boldsymbol{\theta})$ is given, as in Figure 3.3, we can calculate the probability $\Pr(D|\boldsymbol{\theta})$ of observing the paths in D, where $\Pr(D|\boldsymbol{\theta})$ is also called the *likelihood* of the transition probabilities $\boldsymbol{\theta}$ given the data D. After deriving the transition count matrix $\boldsymbol{T}_D = (n_{i,j})$ from D, we calculate the likelihood $\Pr(D|\boldsymbol{\theta})$ using the following formula:

$$\Pr(D|\boldsymbol{\theta}) = \prod_{i,j} \theta_{i,j}^{n_{i,j}} \tag{3.9}$$

For the Markov chain from Example 2 and the data from Example 3, we then calculate:

$$\Pr(D|\boldsymbol{\theta}) = \theta_{\text{central,times}}^{n_{\text{central,times}}} \cdot \theta_{\text{times,statue}}^{n_{\text{times,statue}}} \cdot \theta_{\text{statue,central}}^{n_{\text{statue,central}}} \cdot \theta_{\text{statue,empire}}^{n_{\text{statue,empire}}} \cdot \theta_{\text{empire,central}}^{n_{\text{empire,central}}} \tag{3.10}$$

$$= 1^3 \cdot (\frac{1}{2})^1 \cdot (\frac{2}{3})^1 \cdot (\frac{1}{3})^1 \cdot 1^1 \tag{3.11}$$

$$= \frac{1}{9} \tag{3.12}$$

Fitting to data. If no transition probabilities are given, we can fit a Markov chain to the observed data, i.e., by inferring a transition probability matrix $\boldsymbol{\theta}_D = (\theta'_{i,j})$ from the path dataset D (instead of setting it arbitrarily as in Example 2). For this, we normalize each row of the transition count matrix $\boldsymbol{T}_D$: $\boldsymbol{\theta}_D = (n_{i,j}/\sum_j n_{i,j})$. For the data from Example 3, this results in slightly different transition probabilities than the arbitrarily chosen ones in Example 2:

$$\boldsymbol{\theta}_D = (n_{i,j}/\textstyle\sum_j n_{i,j}) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{3.13}$$

If the transition probabilities $\boldsymbol{\theta}$ from Example 2 match the underlying navigational process of the observed phenomenon, the difference to the derived transition probabilities $\boldsymbol{\theta}_D$ will decrease.

3.2.2. Related work

In this section, we cover extensions and applications related to Markov chains.

3.2.2.1. Extensions

Depending on the naming scheme, Markov chains, as we use them in this work, can be considered to be a special case of the more general notion of *Markov processes*, which is an intensively studied model class. Markov models are stochastic models which essentially are all using processes based on the Markov property. This means that — in some manner — they incorporate a random sequence of states where the next state only depends on the current state. In this section we review a set of selected instances of Markov models in order to give a short overview of this model class. For this, we mainly focus on Markov models where the random sequence of states is directly observable and where the process is not influenced by external factors (i.e., autonomous).

Space and time. In general, there are two fundamental concepts associated with Markov models that are often varied, that is, the *state space* and the notion of *time*. With regard to the state space, there are countable (or finite) state spaces contrasted by the continuous (or general) state spaces. With regard to time there are discrete-time and continuous-time variants of Markov models. While we employ a discrete-time, discrete (and finite) state space model (cf. Section 3.2.1) as introduced by Markov in 1913 [341], continuous-time Markov processes are also often studied. For example, continuous-time, discrete-state models, also called *semi-Markov processes* [422], were introduced by Kolmogorov in 1931 [14] and allow to *stay* at each state for a random amount of time. The Poisson process can be considered to be an example of this model variant [164]. With regard to continuous-time, continuous-state models, processes like the Wiener process or Brownian motion [191] are well known.

Additional dependencies, higher orders, and mixture models. To address different challenges of modeling (web) navigation, such as data sparsity or overfitting [454], various variants and extensions of Markov chains were developed. All of them involve adding additional dependencies and information in some manner. Davison [134] gives an overview of such methods based on Markov models including: higher order Markov models [e.g. 69], Markov trees [e.g. 434], or PPM (prediction by partial matching) [e.g. 104]. Other models like the relational Markov model [13] exploit hierarchical relations of web pages (for personalization purposes).

Of the mentioned extensions higher order processes are an often studied field, i.e., where states depend on a longer history of observations [71, 114, 115, 141, 486]. For example, Singer et al. [454] studied which order of Markov chains best models the memory structure of human navigation on the web (also see Section 2.2.3.1). While higher order Markov chains are more expressive, their complexity increases exponentially when the

order of the model increases requiring large amounts of data in order to derive accurate transition probabilities. Thus, a variety of approaches emerged which introduce more flexibility into the Markov chain without relying on increasing the order [e.g., 292]. Such models include (but are not limited to) the Mixture Transition Model by Raftery [413], or the Variable Length Markov chain [70, 71, 79]. The former model uses a set of lag variables which adjust the transition probabilities given the current state by mixing in the transition probabilities from previous states. The latter [79] allows for variable memory structures by introducing a proxy function which — based on the complete history of states — chooses the number of past states to be considered for deriving the current transition probabilities (which results some kind of auto-correlated process).

There are also models which consider several separate transition probability matrices (cf. Figure 3.3) mixing them in some manner. In that direction, the *Mixed Markov chain* model was studied by Poulsen [405] in the context of customer behavior segmentation. Poulsen defined groups, each with its own transition probability matrix. Group membership probabilities are then assigned to each sequence of observations. Similar group memberships are used by Rendle et al. who factorized Markov chains [419] and by Gupta et al. [223] who reconstructed mixtures of Markov chains [223]. Others define more complex group assignments and transition probability mixtures (e.g., Wallach [504]). In our work [41], we also use a mixed Markov chain model, specifically to compare hypotheses on the underlying processes of heterogeneous sequence data as introduced in Chapter 4. We define a model where formulating hypotheses is flexible (e.g., it allows for group assignments on a transition level instead of a sequence level as in the work by Poulsen [405]). Also our model is by design straightforward to interpret (in contrast, e.g., to Buhlmann and Wyner [79] or Wallach [504]). See Chapter 4 for details.

Switching processes. Markov switching processes [e.g., 177, 411] model observations dependent on hidden Markovian dependency structures. Some classic instances in this class are the Hidden Markov Model [411] (HMM), the Factorial HMM [199] or the Auto-Regressive HMM [227] (also see Bengio [52] and Murphy [370] for further extensions). There are also methods based on, or related to, these methods which are used for prediction, labelling, clustering or segmentation [84, 171, 181, 212, 347]. This includes, e.g., Bayesian non-parametric methods [177, 482] which adjust their complexity based on the data. Such models are also related to our hypothesis comparison approach, MixedTrails [41], where transitions may stem from different transition probability matrices dependent on time (see Chapter 4 for details).

Further extensions. There are also a variety of other models explicitly or implicitly using Markov models. For example hierarchical models are especially studied in the context of Hidden Markov and Markov switching models [173]. Here, observations are dependent on complex hidden structures which are modeled as a hierarchy of Markov chains. Furthermore, there are Markov decision processes [49, 408] where state transitions emit rewards and do not only depend on the current state but also on a chosen action from a set of available actions per state. Usually the goal is to find a policy for choosing actions that maximize the reward. Finally, we also want to mention Markov random fields [278], also called Markov networks, where a set of random variables assumes values

from a set of vertices of an undirected graph and satisfies certain Markov properties, i.e., the pairwise, local, and global Markov property. Markov random fields are for example applied for (3D) image processing [142], or image segmentation [239]. However, they are usually not applied to analyzing or understanding sequential processes as we aim to do in this work.

3.2.2.2. Applications

Markov chains and their variants are applied in a wide range of application domains, e.g., for descriptive and explorative analysis, modeling of real world processes, or prediction. In the following, we give pointers to approaches using Markov chains for geo-spatial behavior as well as web navigation, and provide a short overview on other application domains for which Markov chains are used.

Geo-spatial behavior and web navigation. Many human mobility models as reviewed in Section 2.1.3 also are inherently based on Markovian structures, i.e., they study and model transition counts between different locations [e.g., 165, 330, 345, 378, 450]. Also, Markov chains are often used to model switching processes between a set of movement behavior classes, e.g., employing Hidden or Mixed Markov models [18, 171, 258]. And finally, prediction tasks are an important application of Markovian models in the geo-spatial domain [e.g., 189, 252].

For the application of Markov chain models in the field of web navigation, we refer to the corresponding background section (Section 2.2.3), were we explicitly discuss applications like descriptive and explorative use cases [145, 485], behavioral clustering [81, 406], or the analysis of memory structures involved in navigational processes on the web [111, 454].

Other application scenarios. Besides web navigation and human mobility, Markov himself studied the transition probabilities between vowels and consonants in Alexander Pushkin's novel *Eugene Onegin* [cf., 236, 341]. Furthermore, a classical introductory example for Markov chains is weather prediction, for which they were also used in practice [186]. Furthermore, the Markov chain model and its extensions are used in the field of genetics [277], software testing [525], or information theory [443], for analyzing and modeling genetic algorithms [209, 376], in chemistry for modeling molecule growth [296], or in finance for modeling credit risk [255], as well as for generating lyrics [36] or melodies [386]. Finally, Markov chains also are the underlying concept of Hidden Markov models which are, for example, widely employed in the fields of speech recognition [410] and bio-informatics [272, 470].

One more prominent example for the application of Markov chains worth mentioning is their integral role in the Markov Chain Monte Carlo (MCMC) framework, which is the basic concept for many other techniques, extensions, and applications [76]. Particularly, the MCMC framework is applied to evaluate posterior distributions in complex Bayesian models [203]. Here, each state in a Markov chain represents the value of a sampled variable (e.g., the parameters of a model) and the stationary distribution of the Markov chain corresponds to the probability distribution (e.g., the posterior) for that variable. In this context, two prominent instances of MCMC are the Metropolis-Hastings algorithm for arbitrary models and the more specialized Gibbs sampler which — in its basic form —

is a special case of the Metropolis-Hastings algorithm and takes advantage of conjugate priors [76]. Also note that the output of both methods can be used to estimate marginal likelihoods, which is useful for calculating Bayes factor [273] when analytical inference is infeasible [109, 110].

3.3. Comparing navigational hypotheses using HypTrails

In this thesis, we aim to explain the underlying processes of human behavior in the form of human navigation on the web as well as geo-spatial human mobility based on different theories, existing literature, domain experts, previous experiments, or intuition (cf. Part III). In other words, we formulate hypotheses which compete to describe the same set of data [287]. This problem setting is called *model comparison*. In the context of this thesis, the most important approach to model comparison, is based on Bayes factors [474], which have the advantage of an automatic, built-in Occam's razor balancing the goodness of fit with complexity [273]. In particular, a major part of our work is based on HypTrails (cf. Section 3.3.2) by Singer et al. which employs Bayes factors. We use it to "compare hypotheses about human trails on the web" [453] as well as in a geo-spatial context (cf. Part III), and we extend HypTrails to heterogeneous sequence data with our approach MixedTrails (cf. Chapter 4).

In this section, we first introduce the notion of Bayesian model comparison (Section 3.3.1). Section 3.3.2 then builds upon this to review the HypTrails approach, which is one of the methodological foundations of this work. Finally, we also briefly summarize other methods for model comparison (Section 3.3.3).

3.3.1. Model comparison using Bayes factors

Given an arbitrary set of observed data, we may have different ideas on how this data was generated. Such ideas are often formulated as statistical processes, i.e., mathematical models generating random samples of the data [289]. For example, when observing a sequence of numbers, e.g., $5, 2, 3, 1, 1, 2$ we can assume it was generated by rolling a dice (M_i), thus expecting that each number is independently drawn from a categorical distribution. Or we can model each number as a state and introduce dependencies to the previous number (M_j) as is modeled by Markov chains (as introduced in Section 3.2). Note that both models have parameters which need to be set: The dice may be fair or loaded, and for the Markov chain we need to set transition probabilities.[3]

Now, let us consider a finite set of such models $\mathcal{M} = \{M_1, M_2, \ldots\}$ which compete for describing some dataset D. The goal of *model comparison* is to establish a partial order $\sqsubseteq$ on this set of models [453], i.e., $M_i \sqsubseteq M_j$ denotes that M_j describes the data similarly well or better than M_i. To find such a partial order, the concept of Bayes factors can be used. It follows the intuition that each model M_i has a *prior probability* $\Pr(M_i)$, which represents the probability of model M_i before seeing data. Such prior probability can stem, for example, from theory, previous experiments, or intuition. After seeing the data,

[3] Markov chains are equivalent to a dice when the transition probabilities are the same for each state.

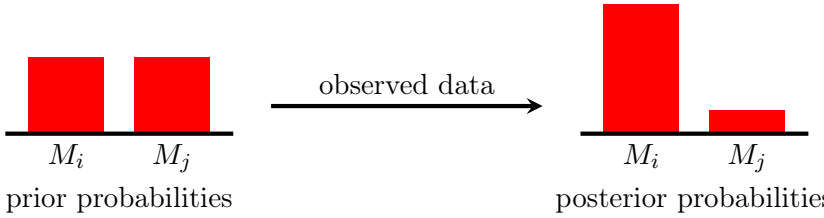


Figure 3.4.: An illustration of Bayes factor. The Bayes factor is a measure of how much the probability for each model shifts after seeing data. This figure shows two models (M_i, M_j) which are equally likely *before* seeing the data (prior). However the probability (posterior) strongly shifts in favor of M_i *after* seeing the data. If this shift is strong enough [273], then M_i is considered to describe the data better.

this (prior) probability gets redistributed between the models resulting in a *posterior probability* $\Pr(M_i|D)$, which can be calculated using Bayes rule:

$$\underbrace{\Pr(M_i|D)}_{\text{posterior of } M_i} = \frac{\overbrace{\Pr(D|M_i)}^{\text{likelihood of } M_i} \quad \overbrace{\Pr(M_i)}^{\text{prior of } M_i}}{\underbrace{\Pr(D)}_{\text{marginal likelihood over } \mathcal{M}}} \tag{3.14}$$

This redistribution is illustrated in Figure 3.4 with two models M_i and M_j. Now, Bayes factor is a pairwise measure of *how much* the probability for two models M_i, M_j shifts after seeing the data and can be expressed using prior and posterior odds [273]:

$$\underbrace{\frac{\Pr(M_i|D)}{\Pr(M_j|D)}}_{\text{posterior odds}} = B_{i,j} \cdot \underbrace{\frac{\Pr(M_i)}{\Pr(M_j)}}_{\text{prior odds}}, \text{ with } \underbrace{B_{i,j} = \frac{\Pr(D|M_i)}{\Pr(D|M_j)}}_{\text{Bayes factor}} \tag{3.15}$$

Here, the Bayes factor is the ratio of the model likelihoods:[4] $\Pr(D|M_i)$ and $\Pr(D|M_j)$. If the shift towards one of the models, e.g., M_i, is great enough then it is said that M_i describes the data better than M_j. Note that if we assume all models to be equally likely a-priori $\Pr(M_i) = \Pr(M_j)$ (as often done in Bayesian model comparison), then the Bayes factor directly implies the posterior probabilities of the models, cf. the derivation of Bayes factor in [273].

The strength of evidence and its interpretation. The likelihoods $\Pr(D|M_i)$ and $\Pr(D|M_j)$ are also called *evidence* because they provide relative evidence for one or the other model to describe the data better. To interpret if there is *enough* evidence to voice a meaningful preference, Kass and Raftery [273] give a guideline based on a threshold t. In particular, they consider the natural logarithm of Bayes factor:

$$\log_e(B_{i,j}) = \log_e\left(\frac{\Pr(D|M_i)}{\Pr(D|M_j)}\right) = \log_e(\Pr(D|M_i)) - \log_e(\Pr(D|M_j)) \tag{3.16}$$

[4] With regard to notation: The *likelihood* $\mathcal{L}$ of a model M_i given the data D is defined as the probability of the data given the model: $\mathcal{L}(M_i|D) \coloneqq \Pr(D|M_i)$. In this work, we use the probability notation.

and say, that there is "very strong" evidence for M_i to be a better description of the data if $\log_e(B_{i,j}) > t$ and that there is "very strong" evidence for M_j to be a better description if $\log_e(B_{i,j}) < -t$. For this, Kass and Raftery generally use a threshold of $t = 5$. This is more conservative than previously suggested thresholds [256]. Even so, Kass and Raftery suggest higher thresholds for specific use cases (e.g., for forensic evidence it should be set higher than in other cases). Thus, in this work, we opt for a more conservative threshold of $t = 10$ to consider one model to be superior to another. Finally, we want to emphasize that using Bayes factor inherently incorporates Occam's razor [273, 336], that is, it prevents overfitting by preferring simple models over complex ones where the complexity is not needed to explain the data. This is an important property when designing hypotheses about human navigation behavior especially in the context of heterogeneity where, for example, hypotheses can get overly complicated when introducing a large set of groups with different navigational characteristics (cf. Chapter 4).

Priors and hypotheses. As shown above, to use Bayes factors we need to calculate the likelihood $\Pr(D|M_i)$ of each model M_i given the data D. In this context, each model M_i has its own set of parameters where each parameter configuration has a specific probability. The corresponding probability distribution $\Pr(\boldsymbol{\mu}_i|M_i)$ is called the *prior* over the different parameter settings $\boldsymbol{\mu}_i$ of model M_i. Then, $\Pr(D|M_i)$ is calculated as the marginal likelihood over the parameter space of M_i:

$$\Pr(D|M_i) = \int_{\boldsymbol{\mu}_i} \underbrace{\Pr(D|\boldsymbol{\mu}_i, M_i)}_{\text{likelihood of } \boldsymbol{\mu}_i} \underbrace{\Pr(\boldsymbol{\mu}_i|M_i)}_{\text{prior of } \boldsymbol{\mu}_i} \, d\boldsymbol{\mu}_i \tag{3.17}$$

In other words the likelihood of a model is defined by marginalizing over all its possible parameter configurations weighted by their prior probability $\Pr(\boldsymbol{\mu}_i|M_i)$; hence the name "marginal probability". While the model structure defines the likelihood $\Pr(D|\boldsymbol{\mu}_i, M_i)$ of the parameters, choosing the prior is not an easy task [194, 273] since Bayes factor can be very sensitive to this choice and employing an uninformed prior (i.e., all parameter configurations are equally likely) is not always the best choice [273]. While this can be inconvenient, it can also be considered an advantage. In particular, Kruschke [287], Rouder et al. [423], and Vanpaemel [496] advocate to leverage this property to compare hypotheses. That is, they propose to encode theory-induced hypotheses into priors. Accordingly, there are many studies which discuss how to elicit informative priors appropriately [194, 382, 495, 497]. In this work, we use and extend HypTrails (cf. Section 3.3.2), which is based on exactly this notion to compare hypotheses about human navigational behavior.

Approximation. An important issue that arises when using Bayes factors is the fact that estimating the likelihood $\Pr(D|M_i)$ of a model can be analytically challenging or computationally intractable [273]. While there is an analytical solution in the case of HypTrails (cf. Section 3.3.2), which we use throughout this work, there exist a variety of other models where such a solution is not available. One example is our extension of HypTrails called MixedTrails (cf. Chapter 4) where we had to derive a sampling scheme. For the general case, there is a variety of other methods for calculating the marginal likelihood based on sampling and approximation including Markov Chain Monte Carlo methods or a Laplace approximation. For a more detailed overview on these methods we

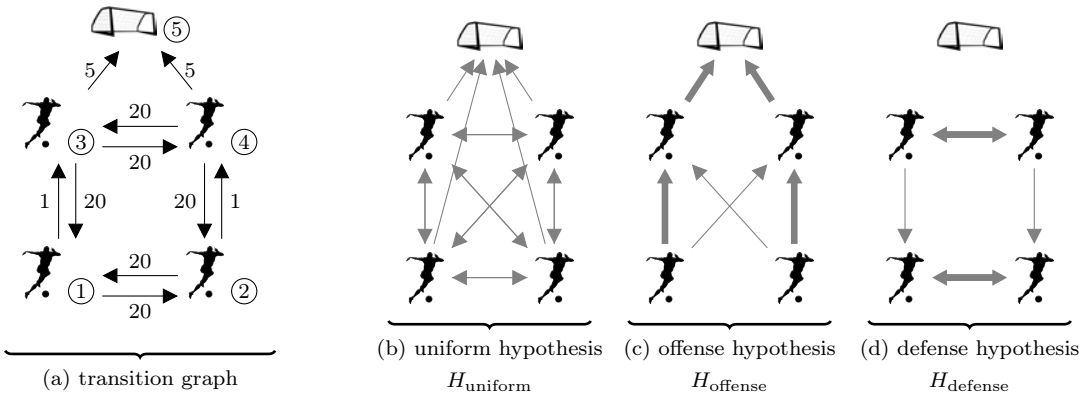


Figure 3.5.: Hypotheses about strategies in a soccer game. In this figure, we show an illustrating soccer example to which HypTrails can be readily applied: We are interested in a team's strategy in a specific game. We have recorded and counted the passes between players as well as shots on the goal and represent them as transition counts between the states of a Markov chain (a). Based on this data HypTrails allows researchers to compare hypotheses about sequential data that express beliefs in transition probabilities (b-d, strength of belief indicated by line width). Utilizing Bayesian inference, it then determines the evidence of the data (a) under these hypotheses (b-d) and ranks the hypotheses based on their plausibility; in this case, even if it is not a perfect match, the defense hypothesis (d) can be considered the relatively most plausible one (cf. Section 3.3.2.1).

refer to Kass and Raftery [273] and more recent overviews by Friel and Wyse [179] and Han and Carlin [228].

3.3.2. The HypTrails approach

Hypotheses about human navigation, as we study in this theses, are usually abstract concepts and can stem from a variety of sources including existing theories, domain experts, previous experiments, or intuition. HypTrails provides an approach to formulate and compare such hypotheses with regard to the relative plausibility for each hypothesis to have generated the data. Figure 3.5 shows an illustrating example in which we depict competing hypotheses about the strategy of a soccer team during a specific game. Figure 3.5a shows the number of passes between players and shots recorded during a game, i.e., the data. Figure 3.5 (b-d) list various hypotheses including a uniform hypothesis, where players pass around randomly, an offensive hypothesis, and a defensive hypothesis. In the course of this section, we will see that the defensive hypothesis explains the strategy of the team well but does not quite cover all aspects of the players' behavior (cf. Figure 3.6). To achieve this, HypTrails employs Bayes factors as described in the previous section (Section 3.3.1). In particular, HypTrails is a special case of Bayesian model comparison, where the model M_i is fixed to a Markov chain and hypotheses are encoded into the prior probability distributions over the model parameters $\boldsymbol{\mu}_i$.

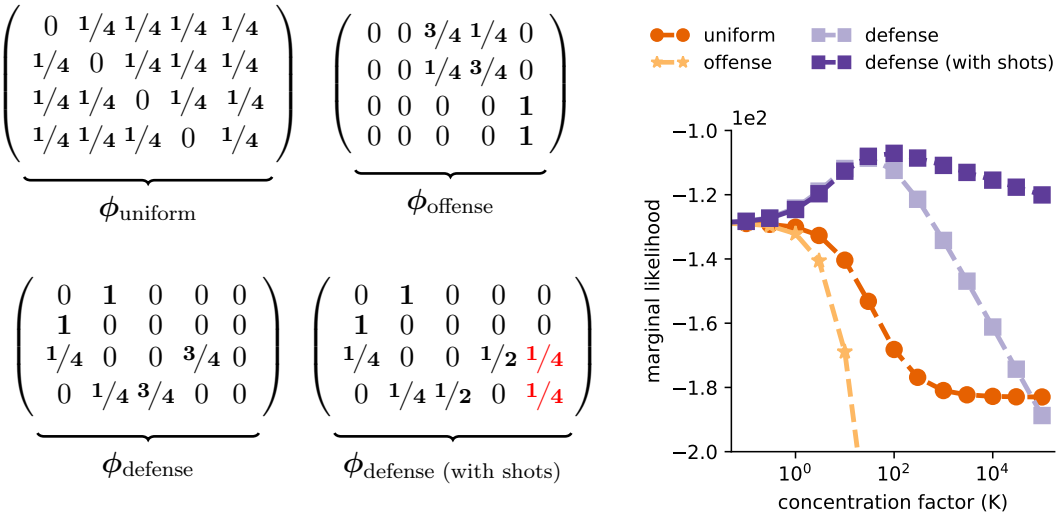
In the following, we first explain how hypotheses are formulated in the HypTrails framework and how they are compared on a practical level. In this context, we specifically focus on the notion of different belief levels (i.e., how much error a hypothesis is allowed with regard to the observed data). This enables a more detailed and robust comparison of hypotheses than when using a single parameter instantiation. Then, we summarize how the previously introduced Bayes factors are utilized to allow for comparing hypotheses based on marginal likelihoods and argue that hypotheses can be encoded as priors over Markov chain parameters within this framework. And finally, we elaborate on the process of eliciting hypotheses as such priors based on different levels of belief.

3.3.2.1. Formulating and comparing hypotheses

For comparing a set of hypotheses $\mathcal{H} = \{H_1, H_2, \ldots\}$ about the underlying processes of sequential data, HypTrails [453] builds on (first-order) Markov chain models (cf. Section 3.2). That is, it formulates each hypothesis H as a matrix of transition probabilities $\boldsymbol{\phi} = (\phi_{i,j})$ between a fixed set of states $S = \{s_1, \ldots, s_m\}$. Some examples for the hypotheses in Figure 3.5 are given in Figure 3.6a. For example, the uniform hypotheses H_{uniform} assumes that the soccer players are passing the ball randomly. Thus, we set the transition probability of the ball between all players to $1/m-1$ where m is the number of states in S. We exclude self-transitions, since players usually do not pass to themselves, hence $m-1$ instead of m. In contrast the offense H_{offense} hypothesis assumes that strikers will always shoot at the goal ($\phi_{3,5} = \phi_{4,5} = 1$), and defense players will always pass towards strikers with a preference to flank ($\phi_{1,3} = \phi_{2,4} = 3/4$, $\phi_{1,4} = \phi_{2,3} = 1/4$). Note that in Figure 3.6a, we have left out the transition probabilities from the goal to the players since we are only interested how the players from the analyzed team behave and not the goal keeper of the opposing team.

Given a sequence dataset D in the form of a transition count matrix $\boldsymbol{T}$ (derived from Figure 3.5a, cf. Definition 3), HypTrails then establishes the relative plausibility for each hypothesis to have generated the data with regard to different strengths of belief κ (also called *concentration factor*). As a measure for plausibility HypTrails calculates the marginal likelihood $P(D|H)$ (also called *evidence*) of each hypothesis H by encoding the hypothesized transition probabilities (cf. Figure 3.5) as priors of a Markov chain with regard to the given concentration factor. The results shown as evidence plots as depicted in Figure 3.6b. For technical details on encoding hypotheses as priors as well as eliciting priors as hypotheses please refer to the subsequent Sections 3.3.2.2 and 3.3.2.3.

Interpretation of evidence. The marginal likelihoods (or evidences) in Figure 3.6 are reported on a log-scale. Using this scale and considering a single concentration factor κ, a hypothesis (H_i) is considered to be more plausible than another (H_j) if its marginal likelihood $P(D|H_i)$ is *sufficiently* larger than the marginal likelihood $P(D|H_j)$ of H_j, where the sufficiency is determined by a threshold t. As already mentioned in Section 3.3.1, in this work, we opt for a threshold of $t = 10$ as inspired by Kass and Raftery [273]. Also, note that HypTrails only compares hypotheses on a *relative scale*. That is, it establishes a relative order based on which hypothesis can be *ranked*. However their *absolute* ability to model the data can not be judged.



(a) Hypothesis matrices **(b)** Marginal likelihood curves

Figure 3.6.: Several formulated hypotheses and an evidence plot created by HypTrails. (a) shows the transition probability matrices derived from the hypotheses visualized in Figure 3.5. Note that we have left out the transition probabilities from the goal to the players since we are only interested in how the players from our team behave and not the goal keeper of the opposing team. Also, we have added another hypothesis $\phi_{\text{defense (with shots)}}$ allowing for some shots at the goal while still playing defensively. (b) shows the results of HypTrails for increasing concentration factors. As it turns out our new hypothesis works best. The defense hypothesis on the other hand covers some important behavioral factors but neglects the occasional shot at the goal. This is why it first achieves high marginal likelihood values which then strongly decrease. This illustrates how HypTrails can provide a more detailed analysis of behavioral processes than scalar comparisons can (e.g., based on AIC or BIC, cf. Section 3.3.3).

Concentration factors. The concentration factor κ is a measure to weight simplicity against accuracy. Its technical formulation follows in Section 3.3.2.3. On an intuitive level, simplistic hypotheses (concentrating their probability mass on a limited set of highly frequented destinations) are favored by small concentration factors, and hypotheses fitting the overall data extremely well (possibly spreading out their probability mass to many different destinations) are favored by large concentration factors. For example, if we favor a simple hypothesis, i.e., we very strongly believe (high concentration factor κ) that the transition from state $s_i \in S$ to state $s_j \in S$ is the *only* option for transitions starting at state s_i ($\phi_{i,j} = 1$), but we observe transitions from s_i to other states as well, then the plausibility of the corresponding hypothesis will be very low, even if the hypothesis is mostly correct. If however, we set the concentration factor to a less extreme level, thus allowing for some inaccuracies, such a "simple" hypothesis that does not quite match the data can still achieve respectable plausibility values. More concretely, in Figure 3.6b, H_{defense} covers a very strong component of the observed data while being relatively simple, i.e., it believes that the players exclusively play defensively by passing the ball from side to side or backwards. H_{uniform} on the other hand assumes that all

passes and shots are equally likely. It spreads out its probability mass and looses a lot of accuracy. However, it allows for transitions that the defensive hypothesis rules out. Thus, for larger concentration factors it benefits from its vagueness and does not suffer strongly decreasing marginal likelihoods as the defense hypothesis does. Nevertheless, the fact that it does not reach the high marginal likelihood (compared to the defense hypothesis), allows to conclude that it does not cover important processes present in the observed data. The new hypotheses $H_{\text{defense (with shots)}}$ somewhat alleviates this issue by extending the defense hypothesis by including shots at the goal. This allows for a more stable marginal likelihood.

Thus, overall, using a range of different concentrations factors κ (different levels of belief), can help to compare a set of hypotheses in more detail than fixing a single belief for each hypothesis. For further examples, also see our studies in Part III where we extensively use HypTrails.

3.3.2.2. From model comparison to hypothesis comparison

To allow comparing hypotheses about sequential data via marginal likelihoods as outlined in the previous section, HypTrails operationalizes Bayesian model comparison (cf., Section 3.3.1), i.e., it uses the notion of Bayes factor. However, instead of comparing different *models*, HypTrails encodes *hypotheses* into the prior distribution over the parameters of a single class of models. As mentioned in Section 3.3.1, this approach has been advocated by a variety of researchers [287, 423, 496]. In this context, HypTrails uses (first-order) Markov chains M_{MC} (see Section 3.2) as its underlying model and encodes hypotheses into the prior distribution $\Pr(\boldsymbol{\theta}|H, M_{MC})$ over the corresponding transition probabilities $\boldsymbol{\theta} = (\theta_{i,j})$. It then calculates the marginal likelihood (also called *evidence*) of a hypothesis H given the data D (cf., Equation (3.17) where the hypothesis is included in the model):

$$\underbrace{\Pr(D|H, M_{MC})}_{\text{marginal likelihood of } H} = \int \underbrace{\Pr(D|\boldsymbol{\theta}, M_{MC})}_{\text{likelihood of } \boldsymbol{\theta}} \underbrace{\Pr(\boldsymbol{\theta}|H, M_{MC})}_{\text{prior of } \boldsymbol{\theta}} \, d\boldsymbol{\theta} \tag{3.18}$$

Since HypTrails only uses M_{MC} as the underlying model, we can simplify the notation to:

$$\underbrace{\Pr(D|H)}_{\text{marginal likelihood of } H} = \int \underbrace{\Pr(D|\boldsymbol{\theta})}_{\text{likelihood of } \boldsymbol{\theta}} \underbrace{\Pr(\boldsymbol{\theta}|H)}_{\text{prior of } \boldsymbol{\theta}} \, d\boldsymbol{\theta} \tag{3.19}$$

Now, having calculated the marginal likelihood of two hypotheses H_i and H_j they are compared using Bayes factor:

$$B_{i,j} = \frac{\Pr(D|H_i)}{\Pr(D|H_j)} \tag{3.20}$$

As mentioned before (Sections 3.3.1 and 3.3.2.1), if the logarithm of Bayes factor exceeds a certain threshold ($t = 10$) one or the other hypothesis is considered to be more plausible in the context of the given data D. Also, note that if we assume all hypotheses to be equally likely a-priori $\Pr(H_i) = \Pr(H_j)$, as often done in Bayesian model comparison, then the

Bayes factor directly implies the posterior probabilities of the models, cf. Section 3.3.1. In the following we clarify how a hypothesis H in combination with a concentration factor κ can be encoded into the prior distribution $\Pr(\boldsymbol{\theta}|H)$ in order to compare hypotheses as showcased in Section 3.3.2.1.

3.3.2.3. Eliciting priors from hypotheses

In Section 3.3.2.2, we have reviewed that model comparison can be employed for comparing hypotheses about sequential data, i.e, by encoding hypotheses as priors over the parameters of a Markov chain. In this section, we first show how Dirichlet priors can be operationalized to define the corresponding prior probability distributions. Then, we clarify how hypotheses (represented as transition probability matrices $\boldsymbol{\phi} = (\phi_{i,j})$, cf. Figure 3.6a) in combination with a concentration factor κ are converted into the parameters of a Dirichlet prior. This process is called *elicitation*.

Marginal likelihood and Dirichlet priors. HypTrails uses Dirichlet priors to encode a hypothesis as probability distributions over the parameters of a Markov chain. In particular, for each state s_i an individual Dirichlet prior $Dir(\boldsymbol{\alpha}_{s_i})$ is used. Each "state prior" $Dir(\boldsymbol{\alpha}_{s_i})$ defines a probability distribution over the transition probabilities $\boldsymbol{\theta}_{s_i}$ from each state s_i to all other states: $\boldsymbol{\theta}_{s_i} \sim Dir(\boldsymbol{\alpha}_{s_i})$. The parameters $\boldsymbol{\alpha}_{s_i}$ are vectors of positive numbers, i.e., $\boldsymbol{\alpha}_{s_i} = (\alpha_{i,1}, \dots \alpha_{i,n})$ where $\alpha_{i,j} \in \mathbb{R}^+$. A Dirichlet distribution $Dir(\boldsymbol{\alpha}_{s_i})$ with $\alpha_{i,j} = \kappa \cdot \phi_{i,j} + 1$, can be pictured by assuming $\boldsymbol{\phi}_{s_i} = (\phi_{i,1}, \dots \phi_{i,n})$ as the *core* transition probability distribution from state s_i to all other states, and κ as the concentration around $\boldsymbol{\phi}_i$ for transition probabilities $\boldsymbol{\theta}_{s_i}$ sampled from $Dir(\boldsymbol{\alpha}_{s_i})$. See Figure 3.7 for an illustration.[5]

Given a sequence dataset D in the form of a transition count matrix $\boldsymbol{T} = (n_{i,j})$, the formula to calculate the marginal likelihood $\Pr(D|H)$ of a hypotheses H represented by a Dirichlet parameter matrix $\boldsymbol{\alpha} = (\alpha_{i,j})$ is:[6]

$$P(D|H) = \prod_{s_i \in S} \frac{\Gamma(\sum_{s_j \in S} \alpha_{i,j})}{\prod_{s_j \in S} \Gamma(\alpha_{i,j})} \frac{\prod_{s_j \in S} \Gamma(n_{i,j} + \alpha_{i,j})}{\Gamma(\sum_{s_j \in S} n_{i,j} + \alpha_{i,j})} \tag{3.21}$$

Elicitation. The next step is to elicit a parameter matrix $\boldsymbol{\alpha} = (\alpha_{i,j})$ from a hypothesis H represented by a transition probability matrix $\boldsymbol{\phi} = (\phi_{i,j})$ and given a specific concentration factor κ (or strength of belief). In other words, we aim to encode the information about transition behavior ($\boldsymbol{\phi}$) together with a certainty or accuracy (κ) into the Dirichlet priors used by HypTrails. In this work, we use a slightly modified version of the approach proposed by the original HypTrails paper [453]. In particular, analogously to the example in Figure 3.7, we use a *stochastic* hypotheses matrix $\boldsymbol{\phi} = (\phi_{i,j})$, and, given a specific concentration factor κ, we calculate the entries of parameter matrix $\boldsymbol{\alpha} = (\alpha_{i,j})$ as follows:

$$\alpha_{i,j} = \kappa \cdot \phi_{i,j} + 1 \tag{3.22}$$

[5] Many thanks to Thomas Boggs for providing the code for the visualization in Figure 3.7: `https://gist.github.com/tboggs/8778945`, accessed: December 2017

[6] See Singer et al. [454] for a derivation. Note however, that we do not explicitly incorporate the probability of the start state since it can be modeled as a dedicated regular state.

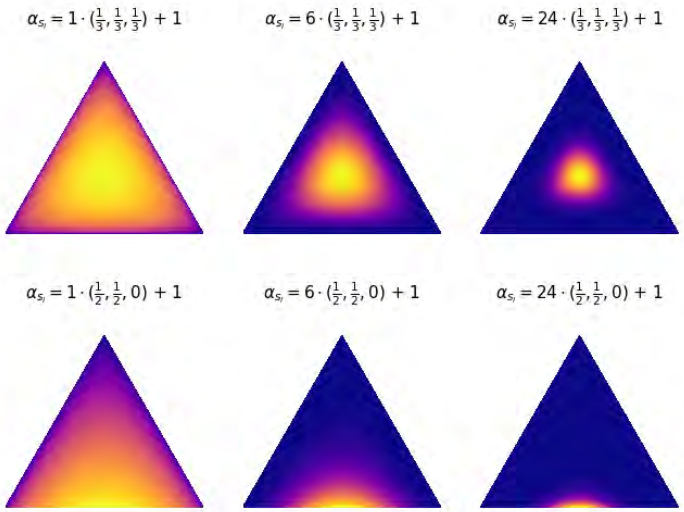


Figure 3.7.: Dirichlet distributions with various parameter settings. This figure shows the distribution of transition probabilities $\boldsymbol{\theta}_{s_i}$ from state s_i of a Markov chain with three states according to various Dirichlet distributions: $\boldsymbol{\theta}_{s_i} \sim Dir(\boldsymbol{\alpha}_{s_i})$. For the parameters $\boldsymbol{\alpha}_{s_i} = (\kappa \cdot \boldsymbol{\phi}_{s_i} + 1)$, we use two *core* probability distributions, $\boldsymbol{\phi}'_{s_i} = (1/3, 1/3, 1/3)$ and $\boldsymbol{\phi}''_{s_i} = (1/2, 1/2, 0)$, with varying concentration factors $\kappa \in \{1, 6, 24\}$. Generally, the transition probability distributions $\boldsymbol{\theta}_{s_i}$ are sampled around the respective core transition probability distribution. For increasing concentration factors κ the transition probability distributions $\boldsymbol{\theta}_{s_i}$ sampled form the corresponding Dirichlet prior will be more concentrated.

The +1 adds the proto-prior that is necessary to ensure proper priors. Also, if $\kappa = 0$, every transition probability configuration is equally likely (referred to as a flat prior, cf. Singer et al. [453]). Note that in some experiments it can make sense to scale the concentration factor κ for better interpretation. That is, we multiply the concentration factor by the number of states n, resulting in the following formula for the entries of the parameter matrix: $\alpha_{i,j} = \kappa \cdot n \cdot \phi_{i,j} + 1$. We use this scaling in most of our experiments (cf. Part III). For each case study we explicitly point out if this is the case.

Alternative elicitation processes. The original HypTrails paper [453] proposes to use a slightly different approach based on the *trial roulette* method [133, 217, 382]. It distributes a fixed number of integer-valued chips across all entries in the parameter matrix $\boldsymbol{\alpha} = (\alpha_{i,j})$. The overall number of chips is calculated as $m^2 + k \cdot m^2$ where k — similar to κ — represents the strength of belief. The chips were distributed as integer values where each entry of $\boldsymbol{\alpha}$ receives at least one chip ($\alpha_{i,j} \geq 1$). The remaining chips are distributed according to a (non-stochastic) hypothesis matrix with entries ranging from 0 to 1. For details please see the original paper [453].

However, using non-stochastic hypothesis matrices results in a different number of chips per state, that is $\sum_j \alpha_{i,j}$ is not necessarily equal to $\sum_j \alpha_{i',j}$ for two states $s_i, s_{i'} \in S$. This can lead to weighting the importance of states differently and thus making the results difficult to interpret. Also, the global distribution of integer-valued chips makes the elicitation process harder to distribute across several computation nodes (cf. Section 6.1).

Nevertheless, the original elicitation process may be considered to be more natural in a sense that integer-valued chips are used together with the notion of trial roulette. Also, it is more flexible in its ability to weight source states. With regard to our methodological contribution in Chapter 4, both processes can be used without restrictions. For our experiments in Part III, we found our own variant to be more practical due to the mentioned consistency across states as well as computational efficiency.

Concentration factors to avoid over-specification. In the general framework of Bayesian model comparison, choosing priors for the corresponding model parameters is not an easy task since usually a variety of information has to be taken into account including relevant data, literature, and in particular the corresponding certainty, i.e., *the strength of belief*. HypTrails somewhat alleviates this issue by formalizing the suggestion by Kass and Raftery [273] to compare several prior instantiations by using a range of concentration factors $\boldsymbol{\kappa} = \{\kappa_1, \kappa_2, ...\}$. This allows for a structured and detailed comparison of hypotheses as described in Section 3.3.2.1.

3.3.3. Related work

Besides using Bayes factor there are several other methods for model comparison. In the context of establishing the order of Markov chains used for modeling human navigation on the web, Singer et al. [454] summarized several methods for model comparison. We follow their example and, beyond covering Bayes factor (Section 3.3.1), outline Frequentist and information theoretic approaches. For another overview on methods for model comparison also see for example Piironen and Vehtari [399] and Vanpaemel [496].

Frequentist approach. There are two major schools in statistics, that is the *Frequentist approach* and the *Bayesian approach*. We already covered a Bayesian approach for model comparison by introducing Bayes factor in Section 3.3.1. In the Frequentist context, one way to establish if a model describes the data better/best is when all other models are rejected with regard to their goodness of fit, i.e., by using p-values [120]. In particular, after fitting a specific model to the data, a proxy measure is used to calculate the difference of the data to the model. This difference is compared to the difference of simulated data from the fitted model. If the difference of the simulated data to their respective fits is generally smaller (small p-value), the model is rejected. For a more detailed discussion see Clauset et al. [120] who describe this methodology in the context of proving a power-law fit using the Kolomogorov-Smirnov statistic [441] as a difference measure. Of course, this method can only establish a single model to describe the data best, i.e., if all other models can be rejected. Among the models which were not rejected, none can be considered better than the other.[7]

In cases where rejecting all models but one is not possible, i.e., where we have to choose from a set of non-rejected models, approaches like the likelihood ratio test [90, 372] can be used (as exemplified by Clauset et al. [120]).[8] For the likelihood ratio test, the parameters

[7] Even so, Clauset et al. suggest that, generally, a model with a very large p-value can be considered to describe the data better than models with a very small p-value.

[8] In addition to the previously introduced Bayes factor, Clauset et al. [120] also mention alternatives like cross-validation [472] and minimum description length [220].

of each model M_i are optimized using the maximum likelihood estimate [303, 424], i.e., the parameters are adjusted in order to maximize the likelihood $\Pr(D|M_i)$ of the model given the data. Then, to compare two models M_i, M_j the logarithm of their likelihood ratio $\log(\Pr(D|M_i)) - \log(\Pr(D|M_j))$ is examined. Depending on the sign, one or the other is considered to describe the data better. To ensure statistical significance, p-values are used to check if the established order has to be rejected. However, in the general case it is not an easy task to formulate the distribution of the likelihood ratio required to calculate these p-values [317]. Thus, while unified approaches exist [317], usually[9] the likelihood ratio test is only used for nested models because then the likelihood ratio is χ^2 distributed [528] and, thus, the p-value can be easily computed. Note that this way, it is only possible to *reject a simpler (nested) model* [120].

Generally, the Frequentist methods mentioned so far can only be used to reject certain models. Thus, they can not establish (or directly confirm[10]) a partial order on a set of hypotheses as is desirable in the model comparison setting. Also note that for more general tests in the Frequentist settings, the use of p-values has to be treated with care and has often been criticized [89, 121, 196, 215, 367, 380].

Information theory. Among others [e.g., 119, 231, 464, 511], there are two prominent information theoretic measures for model comparison, namely the AIC (Aikaike Information Criterion) and the BIC (Bayesian Information Criterion). While AIC [9, 10] can also be interpreted in a predictive setting, it is originally based on approximating the loss of information with regard to the Kullback-Leibler divergence [291] when using a particular model to describe the data. BIC — sometimes called the Schwarz criterion because it has been proposed by Schwarz et al. [439] — approximates the Bayes factor [273, 290, 412] assuming a "unit information prior" and can be calculated independently of a specific prior on the model parameters. This can be useful in cases were calculating the Bayes factor is analytically intractable or specifying an informed prior is not possible.

Just like the Bayes factor (introduced in Section 3.3.1), and in contrast to the already covered Frequentist measures, AIC and BIC both allow to establish a partial order on models (possibly non-nested) competing to describe some collected data. Similarly, they also do not provide a quality in an absolute sense but only establish a relative order on the tested models. Technically, both measures derive a scalar measure which weighs the power to model underlying data (i.e., the maximum likelihood of a model after optimizing its parameters), against the complexity of the model (i.e., the number of parameters to be optimized). The difference of AIC and BIC lies in the underlying theoretical approaches resulting in differing methods to account for the inherent complexity of a model. The question of which method to use is often debated, as AIC and BIC both have their advantages and disadvantages [80, 130, 290, 352, 513]. For example, Weakliem [513] argued that since the Bayes factor is sensitive to the choice of the prior, the "unit information prior", as assumed by BIC, is a too restrictive choice especially because it may be even weaker than those chosen by practitioners [274, 290]. On the other hand,

[9] Lewis et al. [317] also refer to Cox [126], Shapiro and Wilk [444], Vuong [498], and Williams [529] for approaches to apply likelihood ratio tests to non-nested models.

[10] In the case of likelihood ratio tests a suggested preference for one or the other model can be rejected but not confirmed.

AIC — in its original form — is considered to be *not consistent* in the sense that even with an increasing sample size a probability remains to select the larger (more complex) model even though a smaller model is true [290].

In this work, we heavily rely on HypTrails, which uses Bayes factors for model comparison. Thus, we generally do not need to employ the BIC measure which is an approximation of Bayes factor. Also, on top of allowing to establish a relative order on the investigated hypotheses, HypTrails utilizes the sensitivity of Bayes factor with regard to priors in order to incorporate different levels of belief in the respective hypotheses. This enables a more detailed investigation of hypotheses than when employing measures like AIC or BIC.

3.4. Exceptional model mining

For hypothesis comparison (Section 3.3.2), we *already* have to have certain theories, hypotheses, or intuitions about the real world which we want to compare based on observed data. However, we can also take the opposite approach using methods from the field of pattern mining where the data is given and we aim at finding patterns which describe sub-processes of the data in order to ultimately build *new* theories and hypotheses about the real world.

In this section, we briefly introduce *subgroup discovery* and its generalized version *exceptional model mining*. Afterwards we give a short overview on related work and applications in the context of navigation behavior mining, covering, e.g., the closely related field of sequential pattern mining.

3.4.1. From subgroup discovery to exceptional model mining

In literature, pattern discovery an integral part of the KDD (knowledge discovery in databases) process which is described as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [168, chap. 1]. Specifically, the application of data-mining methods for pattern discovery and extraction is the core of the KDD process [169]. Furthermore, "a particularly important subclass of knowledge discovery tasks is the discovery of interesting subgroups in populations, where interestingness is defined as distributional unusualness with respect to a certain property of interest." [532]. For a general definition of subgroup discovery, see Novak et al. : "*Given a dataset of individuals and a property of those individuals that we are interested in, find dataset subgroups that are statistically 'most interesting', for example, are as large as possible and have the most unusual (distributional) characteristics with respect to the property of interest.*" [379]. In the traditional subgroup discovery task the interestingness referred to in this statement is usually given by a Boolean expression over a single attribute (e.g., "class = good") and a subgroup is considered as interesting if the expressions holds more (or less) often than expected [312]. In contrast, *exceptional model mining (EMM)* [149, 307] is a framework that "allows for more complicated target concepts" [149]. That is, a subgroup is considered interesting if the model fitted to the covered data is somehow exceptional (e.g., model parameters are significantly different in the subgroup than in the overall population). This allows to apply EMM to a variety

of settings, including target concepts describing navigational processes. In particular, in Chapter 5, we introduce an approach to employ EMM for finding subgroups with exceptional transition behavior.

The exceptional model mining task. Closely following the definition by Lemmerich et al. [312], an exceptional model mining task can formally be introduced as a tuple (D, Π, C). Where D is a dataset which is represented by a set of instances $i \in I$, Π is the search space of subgroups, i.e., the set of candidates to choose interesting subgroups from, and C is a set of constraints defining the "interestingness" of a subgroup. The goal is to find a set of subgroups $R \in \Pi$ which satisfy the given constraints C, i.e., which are interesting.

Subgroups. A *subgroup* is given by a *subgroup description*, which is a Boolean function $p : D \to \{true, false\}$, and a *subgroup cover* $c(p)$, which is the set of instances covered by p, i.e., $c(p) = \{i \in I \mid p(i) = true\}$. The search space Π of candidate subgroups is usually defined by a *subgroup description language*. Assuming that each data instance $i \in I$ is associated with a set of attributes A, the description language we focus on in this work is the canonical choice of conjunctions of selection expressions over individual describing attributes $A_D \subseteq A$. For nominal attributes these selection expressions are attribute-value pairs and for numeric attributes they can be represented as intervals. Hence, an example for a subgroup description p could be: $gender = male \wedge age < 18$. Due to combinatorial explosion, a large number of subgroups can be formed even from comparatively few selection conditions. Consequently, a large amount of algorithms has emerged to solve the task of subgroup discovery and exceptional model mining [308, 310].

Interestingness. With regard to the *interestingness* of a subgroup, the constraints C usually formulate an interestingness measure $q : \Pi \to \mathbb{R}$ and either require the resulting subgroups to pass a threshold or constrain the result set $R \subset \Pi$ to contain the top k subgroups with regard to q. The interestingness measure q is often based on a set of model attributes $A_M \subseteq A$ (also called target attributes) associated with each data instance $i \in I$. For traditional subgroup discovery, in most cases, the target concept is a Boolean expression over a single attribute (e.g., $class = good$) and a subgroup is considered interesting if the expression holds more (or less) often than expected [see e.g., 283]. In exceptional model mining on the other hand, interestingness is based on more complex target concepts: Given a model class (such as correlation, a classification model, or regression), interestingness can be expressed with regard to the fit of the model parameters on the data instances defined by the corresponding *subgroup cover*. For example, a subgroup could be considered interesting if the model parameters of the model on the subgroup deviate significantly from the model parameters of the model fitted to all data instances. Lemmerich et al. [312] consider the example of correlation (model class) between two model attributes, i.e., the exam preparation time of each student and their final score for the taken course. A finding of exceptional model mining could be: "While overall there is a positive correlation between the exam preparation time and the score ($\rho = 0.3$), the subgroup of males that are younger than 18 years show a negative correlation ($\rho = -0.1$)". Exceptional model mining has been implemented for a variety of model classes including classification [307], regression [150], Bayesian networks [152], and rank correlation [147].

Summary. Given its explicit utilization of a description language to define subgroups, subgroup discovery and exceptional model mining are *descriptive* methods [308]. That is, they allow for finding interesting subgroups of data instances which are interpretable by construction. Thus, by exploring and interpreting the resulting subgroups, exceptional model mining allows to study and understand the observed data by inspecting its underlying components. This may enable the practitioner to conceive new theories for explaining the corresponding observations. In Chapter 5, we utilize this approach and formulate an exceptional model mining class which employs the same Markov chain scenario as HypTrails (cf. Section 3.3.2) in order to find interesting subgroups with regard to their aggregate transition behavior.

3.4.2. Related work and applications

In the following, we cover work related to subgroup discovery and exceptional model mining. In this context, we first give a brief overview of traditional application scenarios and list several corresponding algorithms. Then, as this thesis focuses on paths and traces resulting from human navigation behavior, we list some related work in the area of trajectory and sequential pattern mining.

3.4.2.1. Algorithms and applications

Probably the most intensively studied variant of subgroup discovery and exceptional model mining is frequent itemset mining introduced in the context of association rule mining by Agrawal et al. [6]. In this context, Han et al. [229] give a broad overview on frequent itemset mining, corresponding algorithms, extensions and applications. Some prominent algorithms for efficiently discovering frequent itemsets are the Apriori algorithm [7], Eclat [552], and the FP-growth algorithm [230]. These algorithms were applied and extended many times. For example, the FP-growth algorithm was extended to subgroup mining with categorical as well as numeric attributes [23, 26, 219], as well as the more general task of exceptional model mining [310]. For more information on different algorithms for subgroup discovery and exceptional model mining, we refer to Duivesteijn et al. [149], Herrera et al. [241], and Lemmerich [308]. With regard to general applications, Herrera et al. [241] list a vast set of scenarios ranging from the medical domain, over bio-informatics, marketing, e-learning, and social data, to the field of spatial subgroup discovery. With regard to geo-spatial data, which is highly relevant for our work, researchers explored, for example, subgroups described by tags based on geo-tagged images from the social photo-sharing platform Flickr [22, 309].

The applications mentioned so far can be mainly attributed to the field of subgroup discovery. More complex approaches, which can be considered to be part of the exceptional model mining area, include for example Atzmueller et al. [21] and Atzmueller and Mitzlaff [25] who proposed and applied an extension of SD-Map to mining interesting community structures in social networks [21, 25]. Further applications of the exceptional model mining framework are listed by Duivesteijn et al. [149] including for example the analysis of emotion on music data, or a study on exceptional subgroups with regard to the fauna

of Europe [cf., 152]. Our own work, presented in Chapter 5, also falls into the category of exceptional model mining and extends previous work by introducing a model class which allows to discover subgroups with exceptional movement or navigation characteristics.

3.4.2.2. Sequences and trajectories

While subgroup discovery and exceptional model mining are concerned with subgroups of instance based data, there is a related branch of pattern mining which focuses on sequence and trajectory data. This branch can be divided into three categories: sequence mining, web access pattern mining, and trajectory mining.

Sequential pattern mining. *Sequential pattern mining* was introduced by Agrawal and Srikant [8] and is defined as follows: *"Given a database of customer transactions [each representing a set of items], the problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support. Each such maximal sequence represents a sequential pattern."*. In Srikant and Agrawal [465], the authors further generalized this notion to incorporate time constraints, a sliding time window, and a user-defined taxonomy and developed a corresponding sequential pattern mining algorithm [391]. As for subgroup discovery, sequential pattern mining was applied to many different domains and often extended. See Fournier-Viger et al. [175] and Mooney and Roddick [366] for recent overviews. Both list some prominent algorithms[11], such as Apriori based variants [8], GSP [466], SPADE [551], SPAM [28], or PrefixSpan [390], and name applications from a variety of fields such as bio-informatics [248, 508], e-learning [176, 568], text analysis [403], or even energy reduction in smart homes [440].

Web access pattern mining. Sequential pattern mining approaches are also often applied to web logs, i.e., traces of users left when browsing the web. For example Mooney and Roddick and Fournier-Viger et al. mention [138, 391, 431, 467] which propose or list different algorithms in this context. The main difference to general sequential pattern mining is that each element in a sequence represents a web page visited by a user instead of an itemset. This scenario was named *web access pattern mining* by Pei et al. [391] and falls into the category of *web usage mining*[12], a term which El-Sayed et al. [431] traced back to an article by Cooley et al. [123] from 1997. There exists a wide variety of algorithms for web access pattern mining including a large array of WAP-tree based algorithms [333, 391, 414, 478], as well as some approaches also used for general sequential pattern mining such as GSP or PrefixSpan [167]. There is also a wide variety of related approaches (e.g., as summarized by Facca and Lanzi [167] and Gery and Haddad [198]) consisting of work from the area of association rule mining [178], frequent sequence mining (also called traversal pattern mining, [cf. 100, 101, 362]), or generalized frequent sequence pattern mining [195, 339]. Application of web access pattern mining are personalization of web content, pre-fetching and caching, usability, and e-commerce [167]. Also see Facca and Lanzi [167] for a general overview on web usage mining.

[11] A wide variety of algorithms related to sequential pattern mining is implemented by the SPMF library [174]. Also see: `http://www.philippe-fournier-viger.com/spmf`

[12] Sometimes web usage mining is also referred to as web log mining.

Trajectory (pattern) mining. Besides web access pattern mining, sequential pattern mining can also be applied to the geo-spatial domain. This scenario is referred to as *trajectory pattern mining* [202]. The main difference to sequential and web access pattern mining is that elements in a sequence are not represented as discrete items and events. Instead these elements are defined in a continuous spatio-temporal context. Several techniques were proposed to cope with this challenge. For example, Giannotti et al. [202] and Kang and Yong [271] employed a two-step approach which discretizes the continuous location space before applying sequential pattern mining approaches [e.g., 200]. However, Giannotti et al. [202] also proposed a variant which dynamically computes regions of interest during the mining process. In these approaches the temporal component also plays an important role. For example the "T-patterns" in Giannotti et al. [202] are "a set of individual trajectories that share the property of visiting the same sequence of places with similar travel times." Further work puts their focus on more fine-grained patterns [555], semantic trajectory patterns [547], or apply trajectory pattern discovery for mining travel or life patterns [546, 560], predicting next places [364], or travel recommendation [564, 566]. Also, with regard to trajectory mining in general, Feng and Zhu [170] and Mazimpaka and Timpf [349] give a broader overview and divide the field of trajectory pattern mining into: sequential/frequent pattern mining as covered so far, where several objects are moving independently of each other, periodic/repetitive pattern mining, where only a single object is moving and the goal is to find recurring sequences, and gathering/group pattern mining, where several objects move in unison. Tanuja and Govindarajulu [479] and Zheng [561] also give further information on the more general topic of trajectory mining.

Summary. Overall, all three variants of pattern mining on sequence data and trajectories are concerned with finding interesting sequences or trajectories. In contrast, our work is based on analyzing and finding user groups with interesting transition behavior (see Chapters 4 and 5). Thus, while sequence mining, web access pattern mining, and trajectory mining, can certainly help to formulate hypotheses about the underlying processes of human movement behavior, e.g., by uncovering a set of frequent trajectories or interesting sequences, they do not directly explain human movement on an aggregate level.

Part II.

Methods

4. MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data

In this thesis, we aim to understand human navigation behavior in a geo-spatial context as well as on the web. For this, the previously proposed HypTrails approach [453] (see Section 3.3.2) provides a powerful tool which enables researchers and practitioners to formulate and compare hypotheses about the underlying processes of navigation behavior. However, HypTrails only allows to formulate homogeneous hypotheses while human navigation behavior is inherently heterogeneous, as we have discussed in previous sections (cf. Sections 2.1.5 and 2.2.5). That is, there may exist subsets of the observed phenomena that exhibit strongly differing navigational characteristics (like tourists and locals who have different preferences when navigating urban areas, cf. Section 7.4.3). To address this issue, in this chapter, we propose MixedTrails, an extension of HypTrails, that allows to formulate and compare intricate heterogeneous hypotheses (cf. Section 1.2.1). The following content is based on our article on MixedTrails [41].

4.1. Introduction

Building upon Markov chains, the recently proposed HypTrails approach [453] (also see Section 3.3.2) allows to compare hypotheses about sequential data, where hypotheses represent beliefs in state transition probabilities that are derived from existing literature, theory, previous experiments, or intuition with regard to the respective application domain. This approach is used extensively to study human navigation behavior throughout this thesis (cf. Part III). Figure 4.1 shows a concrete example on soccer data. It features passes between players and shots at the goal (a). In this scenario, we are interested in the strategy a team has used in a game, e.g., an offensive strategy, a defensive strategy, or just random passing. For this purpose, we construct a Markov transition model using the players and the goal as states, and the passes and shots as transitions between these states (b). With HypTrails, we can then express and compare hypotheses (d-g) about pass sequences by specifying different beliefs in transitions. For instance, a simple hypothesis states that all transitions are equally likely (d). Other hypotheses may express predominance of offensive passing (e), a left-flank strategy (f), or defensive play (g). Given such hypotheses, HypTrails calculates the Bayesian evidence of the data under each hypothesis based on which we can rank their relative plausibility (cf. Section 3.3.2). Given the transition data (a), the approach ranks the uniform hypothesis (d) as the most plausible one, as it resembles the overall data (b) the most.

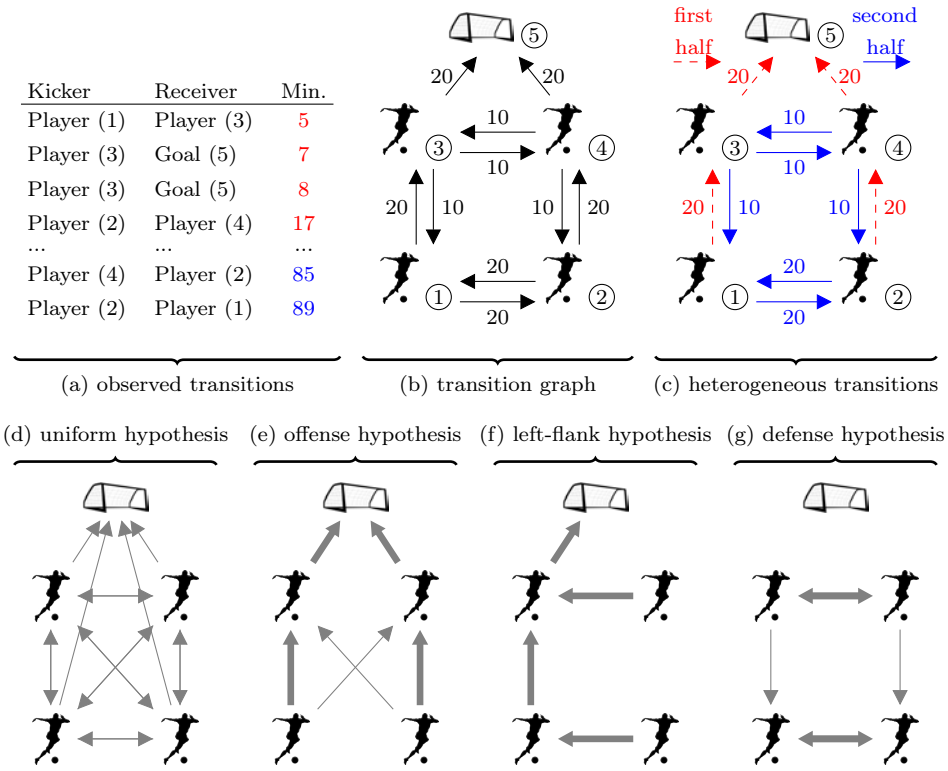


Figure 4.1.: An illustrating example for MixedTrails. In this figure, we show an illustrating soccer example: We are interested in a team's strategy in a specific game. We start with the observed data on passes and shots (a). Using a simple Markov chain, we can model these as transitions between states (b). The previously proposed HypTrails approach allows researchers to compare homogeneous hypotheses about sequential data that express beliefs in transition probabilities (d-g, strength of belief indicated by line width). Utilizing Bayesian inference, it then determines the evidence of the data (b) under these hypotheses (d-g) and ranks the hypotheses based on their plausibility; in this case, the uniform hypothesis (d) is the relatively most plausible one. However, HypTrails is limited to homogeneous data, and does not allow for more fine-grained hypotheses. Indeed, (c) reveals that splitting the data into halftimes allows for a significantly better explanation of the data: A hypothesis that assumes offense (e) in the first halftime and defense (g) in the second halftime appears to be a lot more plausible. MixedTrails enables the comparison of such hypotheses on heterogeneous data.

Problem. Simple Markov chain models, and consequently also the HypTrails approach, assume homogeneous sequence data. As such, they cannot take into account heterogeneity, i.e., behavior stemming from several underlying processes. For instance, research on mobility has found starkly differing user groups such as tourists and locals [312], and there exist different phases of Web navigation with distinct patterns [519]. Further examples are discussed in Sections 2.1.5 and 2.2.5. Also, reconsidering our soccer scenario from Figure 4.1, we can observe that the play style substantially differs for the 1st and 2nd half of the game (dashed and solid arrows). As a consequence, a hypothesis that assumes offensive play for the first halftime and defensive play for the second halftime (cf. Figure 4.1e and Figure 4.1g) could provide a better explanation for our data. However such hypotheses cannot be formulated and compared with existing approaches.

Objective. Thus, our goal in this section is to propose a method that lets researchers intuitively formalize and compare hypotheses about heterogeneous sequence data, such as "The team played according to the offense hypothesis in the first halftime, and according to the defense hypothesis in the second halftime." In this context, we aim at a general and flexible approach: allowing to group transitions by a variety of features, like user groups, state properties, or the set of antecedent transitions on the one hand, and enabling users to formulate probabilistic group assignments as required in the context of smooth behavioral shifts or uncertain classifiers on the other hand.

Approach. To this end, we introduce the *MixedTrails* approach, which covers all necessary aspects to enable the comparison of hypotheses on heterogeneous sequence data: (i) We suggest a method to formalize hypotheses as a combination of several belief matrices in combination with probabilistic group memberships; (ii) We propose the Mixed Transition Markov Chain (MTMC) model that allows to capture such hypotheses; (iii) We show how to elicit priors for this model according to the given hypotheses; (iv) We discuss exact and approximate inference for our model; (v) We provide guidance in the interpretation of the result plots. Finally, we demonstrate the benefits of our approach with synthetic and real world datasets.

Structure and references. MixedTrails is based on the concepts of Markov chains and builds upon HypTrails. For the corresponding background please see Sections 3.2 and 3.3, respectively. In Section 4.2, we first introduce MixedTrails including a formal problem statement, the definition of the underlying MTMC model, the elicitation of hypotheses as priors for MTMC, model inference, and an example illustrating how to interpret the results. Afterwards, we demonstrate MixedTrails on synthetic data (Section 4.3). For examples in the context of real-world applications, please see Sections 7.4.3 and 11.2. Finally, we discuss alternative choices and limitations of our approach in Section 4.4 and review related work in Section 4.5, before we conclude in Section 4.6.

4.2. The MixedTrails approach

In this section, we introduce our approach MixedTrails for comparing hypotheses about heterogeneous sequence data using Bayesian model comparison. To this end, we first elaborate on the specific problem setting (Section 4.2.1) and explain how hypotheses for

heterogeneous sequence data are structured. Then, we introduce the Mixed Transition Markov Chain (MTMC) model (Section 4.2.2) — an extension of the basic Markov chain model — that allows to incorporate such heterogeneity. By formulating hypotheses as elicited priors over the model parameters of this model (Section 4.2.3), we can utilize Bayesian model comparison to make relative judgments about the plausibility of the given hypotheses. Finally, we derive an approach for model inference (Section 4.2.4) and give guidelines for interpreting the results (Section 4.2.5). For illustrative purposes, we will refer to the soccer example visualized in Figure 4.1. For an overview of the methodological background of MixedTrails, we refer to Section 3.2 for a general review on Markov chain models and to Section 3.3.2 for an introduction on the HypTrails approach [453]. Furthermore, we point to Tables A.1 and B.1 for a list of the most important notations.

4.2.1. Problem statement

One of the goals of this thesis is to compare hypotheses about *heterogeneous* sequence data. That is, we consider datasets of human navigation in the form of transition sets $D = \{t_1, \ldots, t_m\}$ between a set of states $S = \{s_1, \ldots, s_n\}$. Such transition sets can be derived from path datasets as introduced in Definition 1 by merging the transitions of all paths.[1] Based on such data, we aim to compare and rate the plausibility of a set of given hypotheses $\mathcal{H} = \{H_1, H_2, \ldots\}$ that express *how* the observed transitions may have been generated. Extending HypTrails [453], we focus on transitions generated by several independent processes.

Hypotheses. We describe a heterogeneous hypothesis $H = (\boldsymbol{\gamma}, \boldsymbol{\phi})$ by two components: *group assignment probabilities* $\boldsymbol{\gamma}$ and *group transition probabilities* $\boldsymbol{\phi}$. The *group assignment probabilities* $\boldsymbol{\gamma}$ associate each transition $t \in D$ in the dataset D with a probability distribution $\boldsymbol{\gamma}_t$ which represents the probability for t to belong to one of the *groups* $G = \{g_1, \ldots, g_o\}$ defined by the hypothesis. We write all group assignment probabilities for a hypothesis as $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_t | t \in D\}$, with $\boldsymbol{\gamma}_t = \{\gamma_{g|t} | g \in G\}$. Here, $\gamma_{g|t}$ is the probability that transition t belongs to group g. Second, the *group transition probabilities* $\boldsymbol{\phi}$ describe the behavior of each group $g \in G$ by specifying respective transition probabilities between states. Formally, all group transition probabilities according to a given hypotheses are written as $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_o)$, with $\boldsymbol{\phi}_g = (\phi_{i,j|g} | s_i, s_j \in S)$, where $\phi_{i,j|g}$ is the probability of observing a transition to state s_j given state s_i within group g. Note that a *homogeneous* hypothesis can be regarded as a special case of a heterogeneous one where all transitions are assigned deterministically to one group.

Comparison. Given several hypotheses, MixedTrails — just like HypTrails — establishes a partial order $\sqsubseteq$ by employing Bayes factors to compare their relative plausibility with respect to a dataset D. This is done by converting each hypothesis H_i into Bayesian priors (see Section 4.2.3) of the generative model MTMC (see Section 4.2.2) and calculating the marginal likelihood (i.e., Bayesian evidence).

Example. For illustration, again consider the soccer game example from Figure 4.1. In the following, we specify two hypotheses for this scenario: a homogeneous one H_{hom}

[1] Note that transitions $t_i, t_j \in D$ can have the same source and target states.

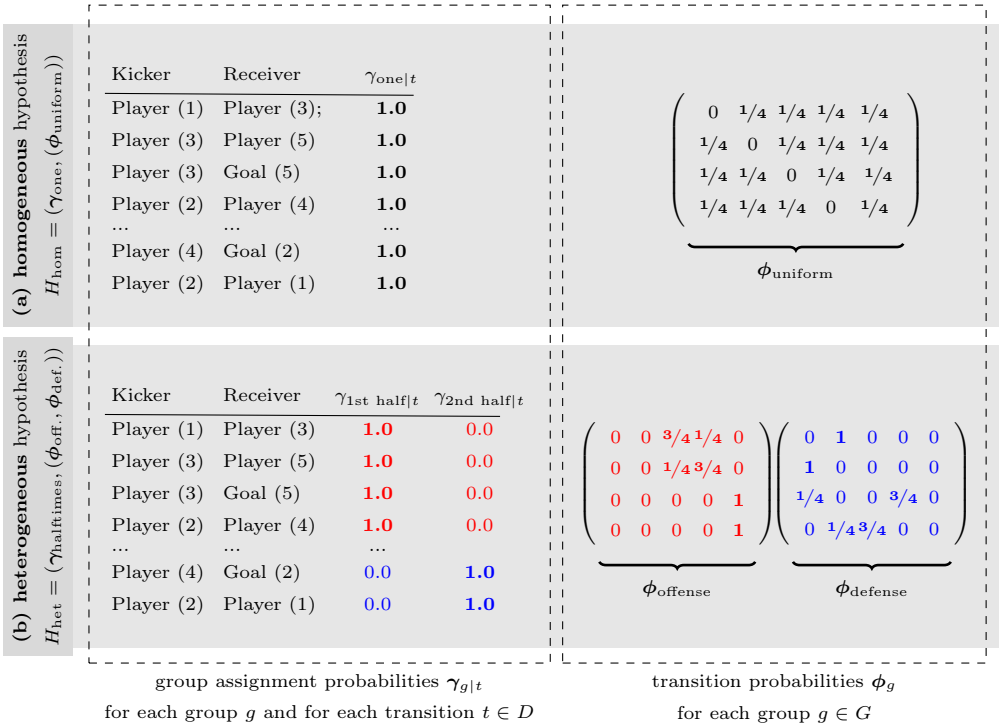


Figure 4.2.: Hypotheses for heterogeneous sequence data. In MixedTrails, we formulate hypotheses about heterogeneous sequence data. For example, in the soccer example, we define two hypotheses: The homogeneous hypothesis H_{hom} (a) assumes that players just randomly pass the ball around; the heterogeneous hypothesis H_{het} (b) assumes an offensive strategy in the first half of the game and a defensive strategy in the second half, cf. Figure 4.1. This is formalized based on two components: *group assignment probabilities* $\boldsymbol{\gamma}$, i.e., probability distributions over the set of respective groups for each transition, and a belief matrix of *group transition probabilities* $\boldsymbol{\phi}_g$ for each group g. The soccer example features a special case, where group assignments are deterministic, i.e., the probabilities are either 0 or 1.

and a heterogeneous one H_{het}. The homogeneous hypothesis H_{hom} expresses the belief that the players just kick around randomly. This can be formalized as a single matrix of transition probabilities $\boldsymbol{\phi}_{\text{uniform}}$ as shown in Figure 4.2a. Consequently, the corresponding group assignment probabilities $\boldsymbol{\gamma}_{\text{one}}$ only assign transitions to a single group. As a more fine-granular hypothesis using a heterogeneous structure, H_{het} assumes that the soccer team played by an offensive strategy in the first half of the game and by a defensive strategy in the second half. For this, we need two separate transition probability matrices ($\boldsymbol{\phi}_{\text{offense}}$ and $\boldsymbol{\phi}_{\text{defense}}$), one for each halftime. Then, we assign each transition to the group (halftime) it belongs to via $\boldsymbol{\gamma}_{\text{halftimes}}$. In this special case, transitions are assigned to halftimes without uncertainty, thus, the probabilities used are either 0 or 1. The resulting hypothesis is defined as $H_{\text{het}} = (\boldsymbol{\gamma}_{\text{halftimes}}, (\boldsymbol{\phi}_{\text{offense}}, \boldsymbol{\phi}_{\text{defense}}))$ as visualized in Figure 4.2b. Now, our approach MixedTrails determines the marginal likelihoods $\Pr(D|H_{\text{hom}})$ and $\Pr(D|H_{\text{het}})$ as a measure for the plausibility of the data under each hypothesis. Since

$\Pr(D|H_{\text{het}}) > \Pr(D|H_{\text{hom}})$, as demonstrated in Section 4.2.5, we assert that explaining the data as a result of our heterogeneous hypothesis (H_{het}) is more plausible than assuming the defined homogeneous process (H_{hom}).

Flexibility. The soccer example from above features an important special case of our approach, i.e., for the heterogeneous hypothesis, the assignment of transitions to groups is *deterministic* $\gamma_{g|t} \in \{0, 1\}$. However, our method also supports arbitrary group assignment probabilities. This is be useful when hypotheses assume gradual change between generating processes (e.g., the team continuously switches from offense to defense during a game), when they suggest that the generating entity switches between different processes (e.g., when the team unpredictably switches between offensive and defensive play), or if there is uncertain or insufficient information available (e.g., the time of some passes was not accurately recorded).

Overall, the ability to specify group assignment probabilities allows to formulate very intricate dependency structures and may serve as an interface to more complex, possibly latent processes. In particular, group assignment probabilities and consequently the transition probabilities associated with each transition can depend on any information associated with a transition, specifically including background information (e.g., user properties, length and duration of the sequence, state properties, or the time of the day), information derived from previously as well as subsequently visited states, or even information about other traces. For instance, this allows for hypotheses modeling higher order Markovian processes, i.e., by defining m^x groups (where m is the number of states and x is the order of the model) and setting the group assignment probabilities depending on the state history of each transition. Some concrete examples on defining hypotheses that take into account the overall sequence are featured in the experimental evaluation in Section 4.3. Thus, even though there are some limitations and possible extensions (cf. Section 4.4), all in all, MixedTrails provides a very flexible and easy to use framework to model a very large and possibly complex set of hypotheses.

4.2.2. The Mixed Transition Markov Chain (MTMC) model

A standard Markov chain model is unable to capture heterogeneity in sequential data. Therefore, we propose the *Mixed Transitions Markov Chain* (MTMC) model as an extension for which we can formulate heterogeneous hypotheses as beliefs over its parameters.

MTMC assigns each transition $t \in D$ in the transition dataset D to a group $g \in G = \{g_1, ..., g_o\}$, which is drawn from an individual categorical distribution with parameters $\boldsymbol{\gamma}_t = (\gamma_{g_1|t}, \ldots, \gamma_{g_o|t})$, where $\gamma_{g|t}$ denotes the probability of transition t belonging to group g. Then, given a common state space, each group $g \in G$ is associated with its own first-order Markov chain. Thus, for each source state s_i, there is a categorical distribution $\boldsymbol{\theta}_{s_i|g} = (\theta_{i,1|g}, \ldots, \theta_{i,m|g})$ over all potential target states. The parameters $\theta_{i,j|g}$ are distributed according to a (prior) Dirichlet distribution $Dir(\boldsymbol{\alpha}_{s_i|g})$ with hyperparameters $\boldsymbol{\alpha}_{s_i|g} = (\alpha_{i,1|g}, \ldots, \alpha_{i,m|g})$. For shorter notation, we write the set of transition probabilities over all states in a group as $\boldsymbol{\theta}_g = (\boldsymbol{\theta}_{s_1|g}, \ldots, \boldsymbol{\theta}_{s_m|g})$ and the set of transition probabilities over all groups as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_o)$. Similarly, we denote the set of all hyperparameters for a single group as $\boldsymbol{\alpha}_g = (\boldsymbol{\alpha}_{s_1|g}, \ldots, \boldsymbol{\alpha}_{s_m|g})$, and the set of all hyperparameters over

all groups, i.e., all Dirichlet parameters, as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_o)$. Finally, we write the set of all group assignment probabilities for all transitions in the dataset as $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_t)$ with $t \in D$. Given these definitions, considering only a single group ($|G| = 1$), MTMC is a direct generalization of the a first-order Markov chain model.

Overall, the MTMC model is described by the following generative process that, given a set of transitions $D = \{t_1, \ldots, t_m\}$, generates for each transition $t_k \in D$, a destination state dst_k for a known source state src_k and known group assignment probabilities $\boldsymbol{\gamma}_{t_k}$:

1. For each group $g \in G$ and each state $s_i \in S$,
   choose transition probabilities $\boldsymbol{\theta}_{s_i|g} \sim Dir(\boldsymbol{\alpha}_{s_i|g})$.
2. For each transition t_k:
   a) Choose the group assignment $z_k \sim Cat(\boldsymbol{\gamma}_{t_k})$.
   b) Choose the destination state $\text{dst}_k \sim Cat(\boldsymbol{\theta}_{\text{src}_k|z_k})$.

4.2.3. Eliciting priors from hypotheses

As mentioned in Section 4.2.1, MixedTrails converts hypotheses into Bayesian priors for the MTMC model (see Section 4.2.2). This process is called *elicitation* as already mentioned in the context of HypTrails (Section 3.3.2.3). However, compared to HypTrails, MTMC requires a different set of parameters: the group assignment probabilities $\boldsymbol{\gamma}$ and the prior parameters $\boldsymbol{\alpha}$. While the group assignment probabilities are directly specified by a hypotheses $H = (\boldsymbol{\gamma}, \boldsymbol{\phi})$, see Section 4.2.1, the parameters $\boldsymbol{\alpha}$ of the Dirichlet prior need to be *elicited* from the transition probabilities $\boldsymbol{\phi}$ consisting of transition probability matrices of *several* groups.

Deterministic Assignments. For *deterministic* group assignments, i.e., $\gamma_{g|t} \in \{0, 1\}$, we determine the parameters $\boldsymbol{\alpha}_g$ of the Dirichlet distributions for each group $g \in G$ separately similar to the approach described for HypTrails in Section 3.3.2.3. That is, for each group $g \in G$ and each state s_i, we set the Dirichlet parameters based on the core distributions $\boldsymbol{\phi}_g$ defined by the hypothesis and a given concentration factor κ. Formally, this is:

$$\alpha_{i,j|g} = \kappa \cdot \phi_{i,j|g} + 1. \tag{4.1}$$

Here, the concentration factor κ reflects the strength of belief in the respective hypothesis (the higher the concentration factor the more accurate a hypothesis has to be to yield high marginal likelihood values). Different settings for the concentration factor lead to different priors. In our approach, we compare hypotheses along a range of different concentration factors, i.e., strengths of belief in the respective hypothesis.

Consider the heterogeneous hypothesis $H_{\text{het}} = (\boldsymbol{\gamma}_{\text{halftimes}}, (\boldsymbol{\phi}_{\text{offense}}, \boldsymbol{\phi}_{\text{defense}}))$ from Figure 4.2b as an example. It features two groups (the first and second half of a soccer game), and for each group $g \in \{\text{1st half}, \text{2nd half}\}$ it defines specific beliefs in certain transition probabilities, via the matrix entries $\phi_{i,j|g}$. For each group, a matrix of prior parameters $\boldsymbol{\alpha}_g$ is determined according to Equation (4.1). The offense hypothesis for the

first half suggests transition probabilities $\boldsymbol{\phi}_{s_1|\text{1st half}} = (0, 0, 3/4, 1/4, 0)$ for the first row of the transition probability matrix. Choosing an arbitrary concentration factor of $\kappa = 10$, we therefore obtain a Dirichlet prior with parameters $\boldsymbol{\alpha}_{s_1|\text{1st half}} = (1, 1, 8.5, 3.5, 1)$.

Probabilistic Assignments. For *probabilistic* group assignments, i.e., $0 < \gamma_{g|t} < 1$, we need to adapt these basic priors to account for misassignments of groups. For example, consider a scenario in which the dataset is divided into two groups that behave completely different. Then, if some transitions cannot be assigned to groups with certainty, the model will randomly associate some transitions which behave like the first group with the second group, and vice versa. Thus, given uncertain group assignments, the behavior expected from a set of transitions assigned to one group is actually a mixture of behavioral traits of both groups. Consequently, we compute the number of pseudo-observations of the Dirichlet priors for a group g as a mixture of hypotheses that is determined by the group assignment probabilities of all transitions. For that purpose, for each transition t_k, we compute the probability that the model assigns t_k to group g although it actually belongs to group g' (i.e., $\gamma_{g|t_k} \cdot \gamma_{g'|t_k}$). This probability is then used as a weight for the respective belief matrix $\boldsymbol{\phi}_{g'}$. Formally:

$$\alpha_{i,j|g} = \kappa \cdot \left(\frac{1}{Z_i} \cdot \sum_{t_k \in D} \left(\sum_{g' \in G} \gamma_{g|t_k} \cdot \gamma_{g'|t_k} \cdot \phi_{i,j|g'} \right) \right) + 1, \tag{4.2}$$

where $1/Z_i$ represents a normalization factor to ensure that the transition probabilities from each state to the other states in the mixture sum up to 1. Note that for deterministic group assignments, the formula simplifies to Equation (4.1).

4.2.4. Model Inference

Similar to HypTrails (Section 3.3.2.2), MixedTrails uses the notion of Bayes factors (see Section 3.3.1) for comparing hypotheses. Thus, for deriving relative plausibilities, we use the MTMC model to determine the evidence (marginal likelihood) for each heterogeneous hypothesis given data (cf. Section 4.2.1). The marginal likelihood can be understood as an average over the likelihood of all parameter settings weighted by their prior probability (given by the hypothesis). This is formally expressed as an integral over all parameter settings $\boldsymbol{\theta}$:

$$\Pr(D|H) = \int \underbrace{\Pr(D|\boldsymbol{\theta}, \boldsymbol{\gamma})}_{\text{likelihood}} \; \underbrace{\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})}_{\text{prior}} \; d\boldsymbol{\theta} \tag{4.3}$$

In the remainder of this section, we elaborate on how to compute the marginal likelihood for our MTMC model given some observed data and any hypothesis (homo- and heterogeneous). We start by deriving an analytical solution. However, the resulting formula is computationally intractable for non-trivial datasets. Thus, we show that for the special case of hypotheses with deterministic group assignments, the calculation can be substantially simplified. Additionally, for the general case, we explain how it can be efficiently approximated by using a sampling approach.

Analytical solution. When ignoring the group assignment probabilities $\boldsymbol{\gamma}$ in Equation (4.3), the marginal likelihood of the MTMC model is equivalent to the homogeneous Markov chain model for which an analytical solution exists [454]. However, in our setting, we need to aggregate over all possible instantiations $\omega \in \Omega$ of group assignments Ω: Each instantiation ω maps each transition t to a group $\omega(t)$. The probability p_ω of an instantiation ω is determined by the group assignment probabilities specified in the hypothesis, i.e., $p_\omega = \prod_{t \in D} \gamma_{\omega(t)|t}$. For a fixed assignment to groups, we can then determine the overall marginal likelihood as the product of marginal likelihoods of the individual groups. For each group, the marginal likelihood can be calculated analytically as a combination of beta functions over the hyperparameters for that group, and over the observed counts in the data according to the fixed group assignment (see Singer et al. [454] for details). Overall, we obtain the following formula (for an in-depth derivation see Appendix C):

$$\Pr(D|H) = \sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{s_i \in S} \frac{B(\boldsymbol{T}_{s_i|g,\omega} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})}, \tag{4.4}$$

where $\boldsymbol{T}_{s_i|g,\omega}$ stands for the vector of transitions counts from s_i to all other states within group g for a given group assignment ω.

Thus, the marginal likelihood of MTMC can be seen as a weighted average over the marginal likelihood of all possible group assignments ω. Unfortunately, this solution is computationally intractable for real world datasets because the number of different group assignments $|\Omega|$ grows exponentially with each additional *transition* $t \in D$.

However, we can substantially decrease the computational costs for the important special case of deterministic group assignments, i.e., where the group assignment probabilities are either zero or one. Then, there is only one valid instantiation of the group assignments, i.e., all but one weight p_ω are zero, and the formula from Equation (4.4) simplifies to:

$$\Pr(D|H) = \prod_{g \in G} \prod_{s_i \in S} \frac{B(\boldsymbol{T}_{s_i|g} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})} \tag{4.5}$$

Thus, in this case, the marginal likelihood is equivalent to the product over the marginal likelihoods across all groups. This can be calculated much more efficiently as the computational complexity only linearly depends on the number of states and groups. The formula also allows for leveraging existing parallelized approaches like SparkTrails [42].

Approximation. For the general, probabilistic case, calculating the marginal likelihood of an MTMC model analytically with Equation (4.4) is computationally intractable. Therefore, we show how we can efficiently approximate it by direct sampling. According to the formula, the overall marginal likelihood is a weighted average over the marginal likelihoods of all group assignments Ω. To approximate this, we sample from the space of all group assignments Ω according to their respective probability p_ω and calculate the average marginal likelihood given these sampled group assignments $\Pr(D|\boldsymbol{\alpha}, \omega)$. Since for individual transitions the process of choosing groups is independent from each other, a single group assignment can be sampled by drawing the group z_k for each transition $t_k \in D$

according to its group assignment distribution $z_k \sim Cat(\boldsymbol{\gamma}_{t_k})$ (also see the generative process in Section 4.2.2). The sampling procedure follows the intuition that factors with small group assignment probabilities contribute less to the overall marginal likelihood. Formally, we can compute the approximated marginal likelihood from a list of sampled group assignments Ω' as:

$$\Pr(D|H) \approx \frac{1}{|\Omega'|} \sum_{\omega \in \Omega'} \underbrace{\prod_{g \in G} \prod_{s_i \in S} \frac{B(\boldsymbol{T}_{s_i|g,\omega} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})}}_{\Pr(D|\boldsymbol{\alpha},\omega)} \tag{4.6}$$

In our experiments, we found that the results are stable for very small numbers of iterations (less than 50) if the number of transitions is sufficiently high. This allows to run our experiments in Section 4.3 in only a few hours on a regular desktop machine.

4.2.5. Visualizing and interpreting results

In this section, we describe our recommended way of performing experiments, visualizing results, and interpreting them. To this end we use the soccer example from Figure 4.1 and investigate which strategies the soccer team has used. For instance, they may have passed the ball randomly, or they may have played by a more intricate strategy. More specifically, given the observed transitions from Figure 4.1 (a-c), we aim to compare the plausibility of the different beliefs in transition probabilities from Figure 4.1 (d-g) utilizing the marginal likelihood as elaborated in Section 4.2.4. In particular, we study the four hypotheses *uniform*, *offense*, *left-flank*, and *defense*, as well as a *data* hypothesis. The latter uses the actual observed transition probabilities as belief; thus it is only used for comparison. We consider these beliefs for three group assignments:

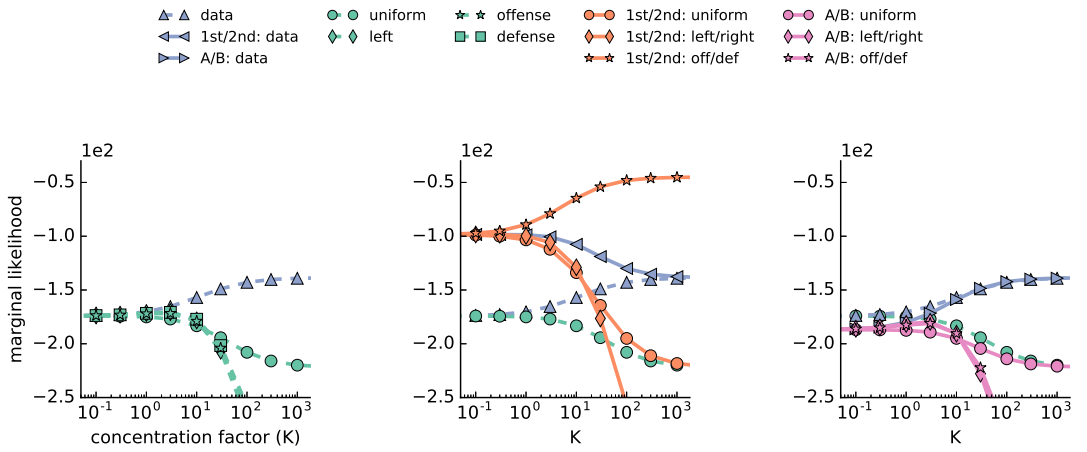
(a) a homogeneous one (all transitions are in one group),

(b) a group assignment defined by the halftime of the passes/shots, and

(c) a completely random group assignment.

The hypotheses are formulated analogously to the examples covered in Section 4.2.3. The results are shown in Figure 4.3 (a-c). In each plot, the x-axis denotes increasing values of the concentration factor κ, which expresses an increasingly strong belief in the hypotheses. The y-axis shows the marginal likelihood on a logarithmic scale; each line represents one given hypothesis; solid lines refer to heterogeneous hypotheses and dashed lines to homogeneous hypotheses. In general, higher values indicate more plausible hypotheses.

Relativity. An essential issue for interpreting the results from MixedTrails (or any method using Bayes factors) is that results are *relative*. Which means that even if one hypothesis outperforms all other hypotheses under consideration, this does not necessarily imply that it models the data well. However, the goal of our approach is to *compare* existing hypotheses from literature, domain experts, ideas, or intuition. The goal is not to find models which perform well for prediction or similar tasks. Nevertheless, it may



(a) Homogeneous hypotheses.

(b) Heterogeneous hypotheses based on a split according to halftimes.

(c) Heterogeneous hypotheses based on a random split.

Figure 4.3.: Results for the illustrating example. This plot shows the MixedTrails results for the illustrating soccer example, i.e., marginal likelihood values of different hypotheses on a logarithmic scale for increasing concentration factors κ (i.e., strengths of belief). We observe that among the hypotheses without grouping, the uniform hypothesis performs best (a). However, far more plausible explanations can be obtained by heterogeneous hypotheses that assume different behavior in both halftimes (b). Finally, randomly splitting the data into arbitrary groups (A/B) leads to less plausible explanations (c).

be desirable to validate the hypotheses with regard to their generative quality. For this, we suggest the comparison with the uniform hypothesis (as we do in this example) or a hypothesis with a flat (uninformed) prior ($\kappa = 0$). The former assumes all *transitions* to be equally likely, while the latter is equivalent to assuming that all transition probability *distributions* are equally likely. Also, additional baselines can arise naturally in specific application domains. For example, if analyzing navigation behavior between web pages, a baseline could be that only transitions to *linked* pages are equally likely, and not to all web pages in the dataset (cf. Dimitrov et al. [144]). We consider the relative order of hypotheses as still viable and interesting if the hypotheses are better than such a baseline hypothesis because they cover at least some aspects of the transition processes. At the same time, if all hypotheses perform worse than the flat prior ($\kappa = 0$), then the data may be too complex for the chosen hypotheses, or the facilitated background data is not sufficient to explain the underlying processes.

Significance. With regard to the significance of differences, we refer to Kass and Raftery's established interpretation table [273]. The table states that conclusions should only be drawn for sections of the marginal likelihood plots where the values are farther apart than 10 (also see Section 3.3.1 for more information). In these cases, the change of the posterior is to be interpreted as "decisive". Consequently, in this thesis, we only draw conclusions from such decisive results when applying MixedTrails.

General properties of curves. Different values along the x-axis enable interpretation beyond providing a relative order of hypotheses: For the left-hand side of the plots (values of κ close to zero) the influence of the transition probabilities of a hypothesis is very weak and the marginal likelihood depends mostly on the group assignment. Thus, the higher the marginal likelihood for $\kappa = 0$, the more a heterogeneous hypothesis can benefit if it models the transition probabilities in each group correctly.

For growing values of κ, the Bayesian framework increasingly takes into account the quality of the chosen transition probabilities for the corresponding group assignments. At first it allows for a large tolerance, i.e., it integrates over variations of the specified transition probabilities. Then, it consecutively decreases this tolerance, requiring that the transition probabilities are very precise. For very high values of κ, the marginal likelihood converges towards the likelihood of the hypothesis. Consequently, the marginal likelihood of heterogeneous hypotheses that assume identical transition behavior in all groups converges towards their homogeneous counterparts (cf. *uniform* and *1st/2nd: uniform* in Figure 4.3b). This is because there is no difference between a homogeneous and a heterogeneous hypothesis if the transition probabilities in each group describe the same generative process.

Overall, the relation of hypotheses along increasing concentration factors gives intricate information about the influence of the different components of the compared hypotheses. For more information and an illustrating example on the interpretation of marginal likelihood curves in the context of homogeneous hypotheses, also see Section 3.3.2.

Results on homogeneous hypotheses. Figure 4.3a shows results for the homogeneous hypotheses. As expected, the data "hypothesis", which is inferred from the actual observed transitions, achieves the highest marginal likelihood values for all κ. Apart from that, the uniform hypothesis explains the observed transitions best. The left-flank, the offense, and the defense hypothesis exhibit strongly decreasing marginal likelihoods for an increasing concentration factor, which indicates that these hypotheses are not supported by the observed data. These results can also be obtained by applying HypTrails.

Results on heterogeneous hypotheses: the split. Our approach MixedTrails enables us to also compare more fine-grained, *heterogeneous* hypotheses. Figure 4.3b features four heterogeneous hypotheses (solid lines) that assign the data deterministically into two groups, i.e., the first and the second halftime. Additionally, it shows the homogeneous data hypothesis and the uniform hypothesis for comparison (dashed lines). For a concentration factor $\kappa = 0$ the marginal likelihood depends only on the group assignment. Therefore, hypotheses with the same group assignment probabilities start at the same marginal likelihood level. Now, since our dataset indeed features different behavior in both halftimes as the group assignment of our heterogeneous hypotheses suggests, their marginal likelihood is higher compared to the homogeneous hypotheses at $\kappa = 0$. This indicates how strongly the split divides transitions into differing processes, before delving deeper into the plausibility of the expressed hypotheses with an increasing concentration factor κ.

Results on heterogeneous hypotheses: the curves. For higher values of κ, the marginal likelihoods diverge: The offense/defense hypothesis — in the first halftime players behave as the offense belief suggests, and in the second halftime as the defense

belief suggests (see Figure 4.2) — is fully supported by the observed data and thus yields the highest values for all κ. In comparison to the homogeneous hypotheses, this curve can be interpreted as: *"This hypothesis features a good group assignment and the transition beliefs reflect the behavior in the observed data well."* If we assign the same belief in transition probabilities to both halftimes, e.g., uniform probabilities, or the globally observed transition probabilities (data), then smaller values are obtained, indicating that these transition beliefs differ from the observed data. Additionally, for very large values of κ, the scores converge with the ones from the respective homogeneous hypothesis because the corresponding heterogeneous hypothesis does not define different transition probabilities for each group, which eventually nullifies the effect of the split. Finally, if we use transition beliefs that are not actually supported by the data for both groups, e.g., a left-flank and right-flank preference in the two halftimes, then the marginal likelihood curve rapidly declines. The respective curve can — in comparison to the other curves — be interpreted as: *"The hypothesis uses a good group assignment, but the transition beliefs are not reflected in observed data."*

Results for a random split and summary. Figure 4.3c shows the same four hypotheses, but assigns transitions to two arbitrary groups randomly (A/B). Since a random group assignment increases the model complexity, but does not allow for a better model of transition behavior, all hypotheses start with a lower value than the homogeneous hypotheses on the left-hand side of the plot. For larger values of κ, we can see the same convergence behavior as before, but, overall, the marginal likelihoods of the heterogeneous hypotheses are substantially lower and also rank lower than their homogeneous counterparts. This is expected of hypotheses that introduce groups without explaining the transition probabilities in each group significantly better than without groups. Overall, these examples give a broad overview of possible MixedTrails results. More examples are covered in Section 4.3.

4.3. Experiments

In this section, we demonstrate the applicability and benefits of our approach with experiments on synthetic data. An open source implementation in Python[2] as well as the datasets[3] are freely available. Conclusions from the experimental results drawn in the text rely on results that are "decisive" with respect to the established interpretation table given by Kass and Raftery [273], cf. Section 4.2.5. For an application of MixedTrails on real-world data please see Section 7.4.3.

4.3.1. Deterministic group assignments

We consider three synthetic examples in order to showcase the properties of MixedTrails in a controlled setting. For each example, we generate a transition dataset according to a predefined mechanism and compare the plausibility of several homogeneous and

[2] `http://dmir.org/mixedtrails`
[3] The scripts for generating the synthetic data are included in the code.

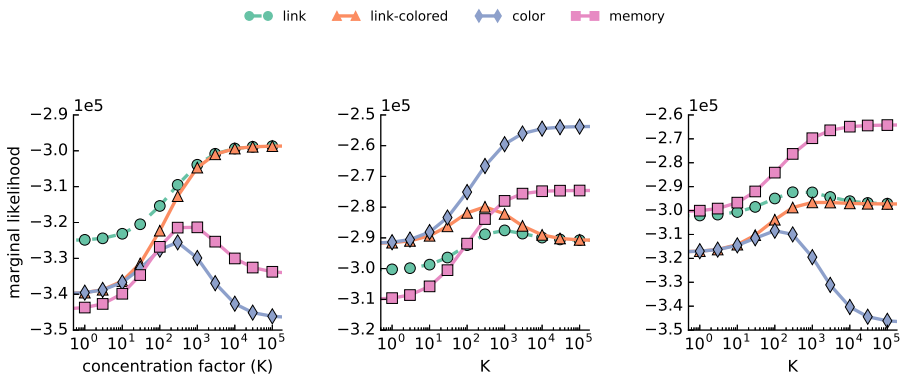
heterogeneous hypotheses. We show that those hypotheses that best capture the known mechanism generating the synthetic data are indeed reported as the most plausible ones.

Datasets. The synthetic transition datasets are based on a random Barabási-Albert preferential attachment graph [35] with 100 nodes and 10 edges for each new node. Each node has a random color $c \in \{\text{red}, \text{blue}\}$ assigned with a probability of $p_c = 0.5$. From this graph, we derive three different transition datasets generated by 10,000 random walkers with different characteristics. Just like each state, each walker also has a color $c \in \{\text{red}, \text{blue}\}$ assigned randomly with $p_c = 0.5$. Each walker chooses her first node randomly and navigates through the network generating transitions depending on different mechanisms which we describe next. The walkers stop after ten steps.Note that the parameters in this study have been chosen arbitrarily. Other settings (e.g., altering the number of walkers, the number of steps, or color probabilities) yield qualitatively similar results. However, reducing the size of the datasets too much will eventually cause the evidence for the correct hypotheses to be less prominent.

For the first dataset D_{link}, we consider *link* walkers that choose the next node uniformly from all adjacent nodes, independent of the walker color. This corresponds to a transition probability matrix $\boldsymbol{\theta}_{\text{link}}$ equal to the (row-wise) normalized adjacency matrix of the underlying graph. For the second dataset D_{color}, walkers of the "red" ("blue", respectively) group exclusively behave according to a probability matrix $\boldsymbol{\theta}_{\text{red}}$ ($\boldsymbol{\theta}_{\text{blue}}$) which adapts $\boldsymbol{\theta}_{\text{link}}$ such that transitions to red (blue) nodes are ten times more likely. The third dataset D_{mem} is generated by "memory walkers" that dynamically choose their next state based on their history, i.e., they use a different transition matrix dependent on the colors of the states they have already visited (including the current state). In particular, if they have visited more red than blue nodes, they use the matrix $\boldsymbol{\theta}_{\text{red}}$, and if they have visited more blue than red nodes, they use the matrix $\boldsymbol{\theta}_{\text{blue}}$. In case of a draw, they use the random transition matrix $\boldsymbol{\theta}_{\text{link}}$.

Hypotheses. For the three datasets we construct corresponding hypotheses: first, the homogeneous hypothesis $H_{\text{link}} = (\boldsymbol{\gamma}_{\text{one}}, \boldsymbol{\phi}_{link})$, which expresses the believe that there are no groups (cf. $\boldsymbol{\gamma}_{\text{one}}$) and all transition are randomly chosen from the available links, thus $\boldsymbol{\phi}_{link} = (\boldsymbol{\theta}_{link})$; secondly, the color-preference hypothesis $H_{\text{color}} = (\boldsymbol{\gamma}_{\text{color}}, \boldsymbol{\phi}_{\text{color}})$ maps each transition to a group based on the color assigned to its walker and uses the actual probability matrices for the transitions in the groups as belief matrices: $\boldsymbol{\phi}_{\text{color}} = (\boldsymbol{\theta}_{\text{red}}, \boldsymbol{\theta}_{\text{blue}})$; and thirdly, the memory hypothesis $H_{\text{mem}} = (\boldsymbol{\gamma}_{\text{mem}}, \boldsymbol{\phi}_{\text{mem}})$ reflects the generating mechanism in the third dataset: The transitions are assigned to groups according to the majority of node colors already visited, and the transition belief matrix is constructed as described in the generation of the third dataset: $\boldsymbol{\phi}_{\text{mem}} = (\boldsymbol{\theta}_{\text{red}}, \boldsymbol{\theta}_{\text{blue}}, \boldsymbol{\theta}_{\text{link}})$. To illustrate how our approach copes with groups that introduce unnecessary complexity, we add a fourth hypothesis $H_{\text{link-colored}} = (\boldsymbol{\gamma}_{\text{color}}, (\boldsymbol{\theta}_{link}, \boldsymbol{\theta}_{link}))$ that uses the grouping into "red" and "blue" walkers, but assumes the same movement behavior for both groups, i.e., equal transition likelihood for all links.

Results. Using MixedTrails, we compare these four hypotheses on all three datasets. The results are visualized in Figure 4.4. For the link dataset D_{link} (Figure 4.4a) we find that the homogeneous hypothesis reflects the data very well and thus achieves the



(a) Link dataset **(b)** Homophily dataset **(c)** Memory dataset

Figure 4.4.: Results for synthetic data with deterministic group assignments. We compare homogeneous (H_{link}) and heterogeneous hypotheses ($H_{\text{link-colored}}$, H_{color} and H_{mem}) on three synthetic datasets (D_{link}, D_{color} and D_{mem}). We observe that the hypotheses that are fitting the respective datasets work best, illustrating that the MixedTrails approach can identify the correct ordering of the defined hypotheses.

highest marginal likelihood (ML) values for all concentration factors. The differences for small concentration factors κ (left-hand side of the plot) indicate that the other group assignment probabilities used by the heterogeneous hypotheses do not introduce valuable information. At first, both heterogeneous hypotheses show increasing ML for increasing concentration factors κ since the hypotheses carry information with regard to the underlying transition processes, i.e., which network links are contained in the data. With increasing concentration factors κ, however, the emphasis on some specific links (i.e., to red or to blue nodes), which is not reflected in the data, leads to a drop of the ML. Furthermore, the memory hypothesis is closer to the data than the color hypothesis as it covers transitions to red and blue nodes in more equal proportions.

Next, we consider the color dataset D_{color} (Figure 4.4b). The ordering of the hypotheses on the left hand side of the plot indicates that the assignment of transition into groups (by walker color) adds valid information to the corresponding hypotheses. However, while the color preference hypothesis H_{color} models the transition behavior within the groups very well, the grouped link hypothesis $H_{\text{link-colored}}$ does not. This explains the diverging ML values for an increasing concentration factor. When comparing the simple link hypothesis H_{link} and the memory hypothesis H_{mem}, we observe that by introducing an incorrect grouping, the memory hypothesis starts at a lower ML values than the link hypothesis which does not introduce any groups. However, with increasing concentration factors, the memory hypothesis starts to perform better, since, in contrast to the link hypothesis, it does incorporate the red and blue transition behavior even if on differing (but somewhat color-consistent) transition groupings. Thus, overall, our model allows to establish the correct ordering of the hypotheses based on the processes used to generate the data.

Finally, we consider the memory dataset D_{mem} (Figure 4.4c). Here we can observe that — as expected — the memory hypothesis H_{mem} performs best for all values of κ. The

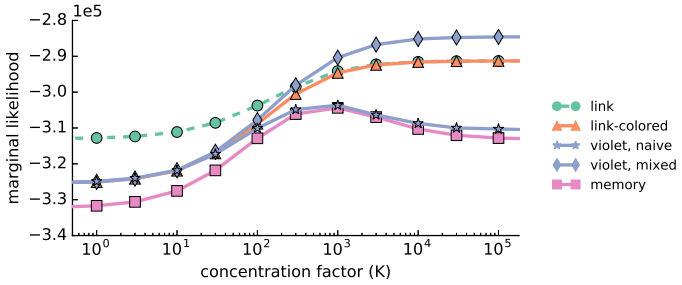


Figure 4.5.: Results for synthetic data with probabilistic group assignments. The *violet, mixed* hypothesis, using probabilistic group assignment probabilities, is the most plausible one for increasing concentration factors as it directly models the processes underlying the data. The *violet, naive* hypothesis illustrates the integral role of the mixing step, as skipping it significantly reduces the performance of a hypothesis even though the underlying processes were correctly understood. Further details are discussed in Section 4.3.2.

group assignment according to walker colors does not correlate with the actual groups in the data and thus leads to lower ML value for low values of κ compared to a homogeneous hypothesis. For high values of κ, we see that the color hypothesis H_{color} does not model the groups well compared to the hypotheses H_{link} and $H_{\text{link-colored}}$ that assume equal likelihood of all links.

Overall, MixedTrails yields results that are in line with the actual generation process of the datasets. Our approach thus allows to derive information about the quality of the group assignments as well as the transition behavior within the groups. The strongly diverging characteristics of the different hypotheses illustrates the flexibility of MixedTrails.

4.3.2. Probabilistic group assignments

So far, we have only considered *deterministic* group assignment probabilities in the experiments, i.e., assigning transitions to a single group by only using binary probabilities: $\gamma_{g|t} \in \{0, 1\}$. However, there is a wide variety of situations where it is useful to consider probabilistic group assignments or fuzzy walkers, e.g., when considering smooth behavior transitions between different times of a day, when transitions are assigned to groups by an uncertain classifier, or when walkers randomly choose between different movement patterns. Here, we explore probabilistic group assignments in a synthetic dataset. For a real world example of an uncertain classifier, see Section 7.4.3.

Dataset. We use the same underlying network as in the previous example to construct a dataset. However, instead of "red" and "blue" walkers, the sequences are now generated by walkers with "mix colors", called *violet walkers*, i.e., the walkers randomly choose to walk according to the red $\boldsymbol{\theta}_{\text{red}}$ or to the blue $\boldsymbol{\theta}_{\text{blue}}$ transition probability matrix at each step. For example, a violet walker w associated with a shade of violet $s_w = 0.3$ will choose to be a red walker for 30%, and a blue walker for 70% of her transitions. We create a dataset D_{violet} of 10,000 walkers that each perform 10 transitions. We assign a shade

of violet s_w to each walker w, which we draw from a Beta distribution $s_w \sim Beta(1,1)$. Before each transition of a walker, she randomly draws a color $c \in \{\text{red}, \text{blue}\}$ according to her shade of violet s_w using a Bernoulli distribution $c \sim Bernoulli(s_w)$. Then, she uses the respective transition matrix $\boldsymbol{\theta}_{\text{red}}$ or $\boldsymbol{\theta}_{\text{blue}}$ dependent on the chosen color c to determine her next destination. As in the previous experiment, altering the parameters of this study will not change the results of this study qualitatively. However, considering the probabilistic nature of our approach, reducing the size of the datasets too much will eventually result in random inconsistencies between runs and cause the evidence for the correct hypothesis to be less prominent.

Hypotheses. As hypotheses, we define H_{link}, $H_{\text{link-colored}}$ and H_{memory} analogously to Section 4.3.1. In addition, we introduce a hypothesis $H_{\text{violet}} = (\boldsymbol{\gamma}_{\text{violet}}, \boldsymbol{\phi}_{\text{violet}})$ specifically tailored to violet walkers. Thus, we define the group dependent transition probabilities as $\boldsymbol{\phi}_{\text{violet}} = (\boldsymbol{\theta}_{\text{red}}, \boldsymbol{\theta}_{\text{blue}})$. Now, violet walkers choose transition probability matrices *probabilistically* dependent on their shade of violet. Using our MTMC scheme, this can be modeled by setting the corresponding group assignment probabilities according to a walker's shade of violet s_w: $\gamma_{g|t_w} = (s_w, 1 - s_w)$. That is, each transition t_w by walker w has a probability of s_w to be a red transition and a probability of $1 - s_w$ to be from the blue transition probability matrix.

Results. The results are shown in Figure 4.5. The first observation is that the violet hypothesis H_{violet} (mixed) works best for increasing concentration factors. Note that we consider two variants of the violet hypothesis, one (*violet, mixed*) elicited using the mixing method proposed in Section 4.2.3 and one (*violet, naive*) elicited as if it was a deterministic hypothesis. The results show that the mixing step is an integral part of MixedTrails, as skipping it significantly reduces the performance of the heterogeneous hypothesis even though the underlying processes were correctly understood.

As for the other hypotheses, the *link* hypothesis works best. This is because, generally, a perfectly violet walker ($s_w = 0.5$) behaves exactly like a link walker. This also explains the differing results for lower concentration factors: The grouping introduced by the violet hypothesis injects complexity which is not splitting transitions in a manner that can easily be explained. Thus, for low concentration factors, which imply a large uncertainty in the hypothesis, this reduces the plausibility of the more complex hypothesis. However, with growing concentration factors the better modeling of the transition probabilities justifies the added complexity making the violet (*mixed*) hypothesis the most plausible one.

With regard to the increased complexity, the colored (heterogeneous) link hypothesis (*link-colored*) has the same disadvantage as the violet hypothesis; consequently, it is inferior to the homogeneous link hypothesis. The memory hypothesis has the lowest plausibility as it does not reflect the generative process of the dataset and introduces three groups instead of just two.

Overall this example shows that, by using MixedTrails, heterogeneous data can be modeled accurately and that the mixing procedure for eliciting probabilistic hypotheses as introduced in Section 4.2.3 is an integral part of the approach.

4.4. Discussion

With MixedTrails, we have proposed a powerful approach to formulate and compare hypotheses about heterogeneous sequence data. In this section, we discuss some alternative choices as well as possible misunderstandings and shortcomings of our method.

Comparison, prediction, and conception. MixedTrails is a method for hypothesis *comparison* (Section 1.1). This is also sometimes called a *deductive* approach in certain contexts [63, 240, 491] — meaning that it requires a set of predefined hypotheses based on *ideas* and *theories* from the application domain as input and compares them using observed data. While the corresponding results also give an indication of the predictive potential of hypotheses, we do not fit them to the data. For utilizing the data to learn models that excel at *prediction*, a multitude of other, more specialized methods are available [e.g., 172, 302, 545]. Note, that these methods usually do not yield directly interpretable results. If they do [e.g., 172], they can be used for hypothesis *conception* (Section 1.2.2). This is sometimes also called an *inductive* setting [63, 491] — taking the opposite approach than MixedTrails: such methods use observations to extract patterns or regularities from which *new* hypotheses or theories can be derived. We develop one of such approaches in Chapter 5, i.e., SubTrails for discovering subgroups with exceptional transition behavior. There are also other specialized approaches useful for conceiving novel hypotheses, e.g., methods for segmentation, labeling, or clustering [64, 177, 404, 494, 504]. However, in this thesis, we focus on subgroup discovery due to its inherently interpretable nature.

Extensions and alternative approaches. While MixedTrails provides a very flexible and easy to understand framework for specifying and comparing hypotheses, there is a variety of possible extensions and alternative approaches. For example, in this work, we employ priors for transition probabilities, but specify group assignment probabilities directly and fixed, which somewhat forces the user to be very specific with regard to group assignments. In contrast, using a flat prior over group assignments, the user could compare hypotheses that introduce groups of transition probabilities without having to specify which transition belongs to which process. Also, MixedTrails can not directly express dependencies between the groups of the transitions within a sequence as for example possible in Markov switching processes such as the Hidden Markov model (cf. for instance the concept of "stickiness" as considered by Fox et al. [177] and Wetzels et al. [521]). That is, while we can construct hypotheses in a way such that group assignment probabilities are derived by Hidden Markov structures, hidden state dependencies can not be explicitly modeled. We could resolve this by using more complex models for sequential data. This, however, would come at the cost of substantially increased efforts for specifying model parameters in the hypotheses, especially considering the wide range of incorporated background knowledge. Overall, MixedTrails tries to balance the amount of parameters required to formulate a hypothesis against expressiveness. Nevertheless, we acknowledge the potential of formulating more complex dependencies with the help of more complex models, especially when considering the possibility of flat/uninformed priors over certain parameter groups, but leave further studies to future work.

MixedTrails vs. separate HypTrails comparisons. A simplistic alternative to our approach could be to apply the original HypTrails method for homogeneous data separately to the groups of a hypothesis. This, however, is limited to deterministic group assignments and does not allow to compare hypotheses with different group assignments (or no group assignments at all). In addition, MixedTrails provides the theoretical background on how to aggregate results for the individual groups, i.e., by multiplying their marginal likelihood.

Using different strengths of belief. We are using different strengths of belief (i.e., concentration factors κ) in order to study different properties of our hypotheses. Calculating the marginal likelihood for very large concentration factors κ approximates the likelihood of the model for fixed parameters, which is commonly used to compare parameter settings in Frequentist statistics (e.g., via a likelihood ratio test). However, by also investigating lower concentration factors, we obtain additional information on the quality of the group assignments (cf. Section 4.2.5). Furthermore, our approach enables the observation of the dynamics for growing concentration factors, which allows us to judge whether a hypothesis covers predominant factors of the underlying processes generating the sequential data. Thus, we believe that the analysis based on different concentration factors can yield a more detailed comparison of hypotheses than other, one dimensional measures, such as the model likelihood, which is included in our approach as a special case and shown on the right-hand side of our result plots.

Nevertheless, we acknowledge that it may be useful to derive a single number by which hypotheses can be compared. To achieve this we could either set a fixed κ according to some background information or, in a more Bayesian way, we could treat the concentration parameter κ as a free parameter and marginalize over it. This, however, would require specifying a prior over this free parameter, which is inherently a difficult choice. As a simple solution, we propose to compute the average marginal likelihood over a set of κ values. This is equivalent to a prior that regards these values as equally likely. Overall, summarizing result curves into a single value in this way requires additional task-dependent choices and comes with a loss of information with regard to the result on the one hand, but allows for a more compact representation of results on the other hand. Developing guidelines for choosing appropriate priors over κ remains an open issue for future work.

Efficiency and convergence. In the general case, the marginal likelihood of the MTMC model has to be approximated. While the method from Section 4.2.4 has converged quickly ($\ll 50$ iterations) so that we were able to calculate our results on regular consumer hardware in a few hours, parallelizations along the lines of [42] may be useful for larger datasets. We have also experimented with other methods for approximating the marginal likelihood such as by Chib [109], but have found irregularities in the convergence behavior. Further studies may address both, the parallelization of our method and exploring other approximation schemes.

Multiple comparisons. Our approach enables the comparison of multiple hypotheses against each other. In that direction, it can also be checked whether one of the hypotheses performs better than a simple baseline hypothesis (such as the uniform hypothesis). If

many hypotheses are tested in this way, then the multiple comparison problem should be taken into account. That is, even if hypotheses are generated purely at random, some of them would appear to be statistically significantly better than the baseline, cf. Benjamini and Hochberg [54]. Although our approach is in principle affected by this problem, we see this issue as non-crucial in our setting as (i) the main goal of our approach is not to show whether one of our hypotheses can beat a baseline, but to compare hypotheses against each other (pairwise) and (ii) we use only a comparatively small set of hand-elicited hypotheses in our comparisons. Apart from that, there is intense discussion how multiple comparisons are to be viewed from a Bayesian perspective, see for example [197, 216]. Nonetheless, exploring the challenges of multiple comparisons is an issue that we will study more in-depth in future work.

4.5. Related work

MixedTrails is based on HypTrails introduced by Singer et al. [453] (also see Section 3.3.2). HypTrails as well as our own approach, MixedTrails, build on the concept of Markov chains. Corresponding related work on the application of Markov chains to human navigation behavior is covered in Chapter 2 and Section 3.2.

To the best of our knowledge Markov chains and their extensions (as covered in Section 3.2.2.1) have not been employed for the comparison of hypotheses so far. This specifically includes variants of the mixed Markov model [461]. Additionally, the expressiveness of most these models is limited [e.g., 223, 405, 419, 461], i.e., some hypotheses formulated using MixedTrails can not be expressed with these models.

Another set of Markov chain extensions related to our approach is the class of Markov switching processes [177, 411], which model observations dependent on hidden Markovian dependency structures. For more examples, please see Section 3.2.2.1. There are also methods based on, or related to, these methods which are used for prediction, clustering or segmentation [171, 181, 212, 347], including, e.g., Bayesian non-parametric methods [177, 482] which adjust their complexity based on the data. However, such methods fit models to the data, i.e., they learn model parameters. Sometimes these model parameters can be used to *find new* hypotheses, but the corresponding process is usually tedious as it often requires understanding possibly arbitrarily complicated probability distributions. Also, while, e.g, Hidden Markov models were applied to compare streaky behavior with a baseline model [521], to best of the authors knowledge, there are no general approaches to apply Markov switching processes for *formulating and comparing* existing hypotheses in the context of background data.

For a broad overview on work about model comparison as applied by MixedTrails, we refer to Section 3.3.3 for a more detailed discussion. With regard to model comparison in the context of Markov chains, statistical methods for comparing the fits of varying Markov order were summarized in [454]. This includes likelihood ratio tests, information-theoretic AIC, BIC, and DIC approaches, or the Bayes factor. MixedTrails focuses on comparing fits by using marginal likelihoods and Bayes factors [474]; these have the advantage of an automatic built-in Occam's razor balancing the goodness of fit with complexity [273].

For a more detailed discussion on alternative methods for model comparison, we refer to Section 3.3.3. Additionally, instead of only using a flat Dirichlet prior (as often the case in Bayes model comparison), we also utilized the sensitivity of the marginal likelihood on the prior for comparing theory-induced hypotheses within the Bayesian framework With this, we followed the HypTrails approach (cf. [453] and Section 3.3.2) which was inspired by, e.g., [287, 423, 496]. To the authors' knowledge, there exist no previous approaches for the comparison of hypotheses about transition behavior that differentiate between several groups contained in the data. Our contribution (in the form of MixedTrails) is in line with a general trend towards Bayesian methods for data analysis [50, 288].

4.6. Conclusion

With MixedTrails, we introduced a Bayesian method for comparing hypotheses about the underlying processes of heterogeneous sequence data. MixedTrails incorporates i) a method for formulating heterogeneous hypotheses using ii) the *Mixed Transition Markov Chain* (MTMC) model, which enables specifying individual hypotheses for very flexible subsets of transitions, i.e., with regard to certain user groups, state properties, or the set of antecedent transitions. Furthermore, iii) we introduced methods for eliciting hypotheses as parameters for this model, iv) showed how to calculate the marginal likelihood, and v) provided some guidance on how result plots can be interpreted to compare the corresponding hypotheses. The benefits of our approach were demonstrated on synthetic datasets and will be further exemplified on real-world data throughout this thesis (Sections 7.4.3 and 11.2). Overall, MixedTrails enables us to cope with one of the major challenges of understanding human navigation behavior identified in Chapter 1, i.e., formulating and comparing hypotheses incorporating the inherent heterogeneity of human navigation.

In the future, we may explore our method in additional real-world applications, such as investigating the movement of (groups of) Flickr users (beyond tourists and locals, cf. Section 7.4.3), or studying groups of editors on Wikipedia. Furthermore, more complex priors or hierarchical models may allow for more powerful ways of expressing hypotheses.

5. SubTrails: Mining subgroups with exceptional sequential behavior

As many studies have found, human navigation behavior is a heterogeneous process (cf. Sections 2.1.5 and 2.2.5), e.g., Marchionini [340] find differences in navigation behavior between younger and older users of a full-text electronic encyclopedia. While the MixedTrails approach, proposed in Chapter 4, allows to formulate and compare heterogeneous hypotheses about human navigation behavior, i.e., incorporating multiple sub-processes to explain a given set of observations, it requires that interpretable subsets of the data already exist for which navigational hypotheses can be formulated. Applying exceptional model mining can alleviate this issue: In this chapter, we propose an approach called SubTrails for mining descriptive subgroups (e.g., "male tourists from France") with exceptional transition behavior. This gives insights into the underlying heterogeneous processes of human navigation, and thus supports the conception of novel hypotheses (cf. Section 1.2.2). This chapter is based on our previously published article on SubTrails [312].

5.1. Introduction

Exceptional model mining [149, 307], as reviewed in Section 3.4, is a framework that identifies patterns which contain unusual interactions between multiple target attributes. In order to obtain operationalizable insights, it focuses on the detection of *easy-to-understand* subgroups, i.e., it aims to find exceptional subgroups with descriptions that are directly interpretable by domain experts.

Problem setting. While we have introduced a method for comparing hypotheses about heterogeneous navigation data with our MixedTrails approach in Chapter 4, coming up with a set of heterogeneous hypotheses to compare is not an easy task. In particular, either subsets of the observed data (e.g., younger vs. older students) have to be selected by hand, thus, resulting in a tedious process of finding and checking navigational characteristics for interesting subsets. Or, if discovered in an unsupervised fashion (e.g., clustering), the subsets are usually not straight forward to interpret because often descriptive attributes are not part of the discovery process. Applying exceptional model mining to the observed data can address this problem. In particular, it can be used to automatically identify subgroups of people (such as "male tourists from France") or sub-segments of time (such as "10 to 11 p.m.") that exhibit unusual movement characteristics, e.g., tourists moving between points-of-interest or people walking along well-lit streets at night. Similarly, this method can discover subgroups of web-users with unusual navigation behavior. Also, there are many application scenarios beyond navigation analysis, such as discovering companies with unusual development over time Judge and Swanson [262].

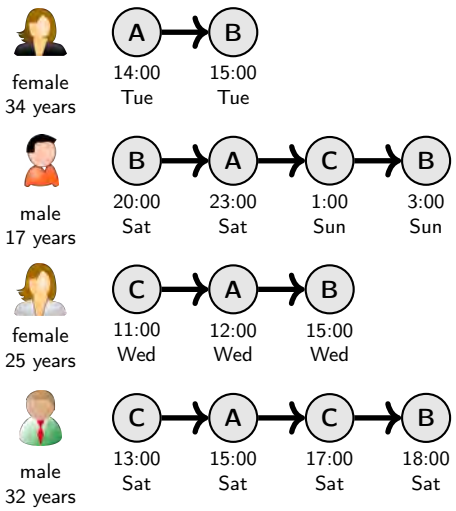
Approach. To enable the application of exceptional model mining to mining subgroups with exceptional transition behavior, we introduce *first-order Markov chains as a novel model class for exceptional model mining*. To apply exceptional model mining with this model, we derive an interestingness measure that quantifies the exceptionality of a subgroup's transition model. It measures how much the distance between the Markov transitions matrix of a subgroup and the respective matrix of the entire data deviates from the distance of random dataset samples. This measure can be integrated into any known search algorithm. We also show how an adaptation of our approach allows to find subgroups specifically matching (or contradicting) given hypotheses about transition behavior [cf. 44, 453, 503]. This enables the use of exceptional model mining for a new type of studies, i.e., the detailed analysis of such hypotheses. We demonstrate the potential of the proposed approach on several synthetic datasets. For an application on real-world data featuring human navigation behavior, we refer to several of our case studies (cf. Sections 7.4.2 and 11.3).

Structure and references. In this section, we heavily rely on Markov chains and the framework of exceptional model mining. For background on both concepts please refer to Sections 3.2 and 3.4, respectively. The main approach for mining subgroups with exceptional transition behavior is introduced in Section 5.2. Section 5.3 presents experiments and results on synthetic data. For applications on real-world data featuring human navigation behavior we refer to several of our case studies (cf. Sections 7.4.2 and 11.3). Finally, we discuss related work in Section 5.4, before we conclude in Section 5.5.

5.2. The SubTrails approach

Given a set of state sequences and additional information on the sequences or parts of sequences, our main goal is to discover subgroups of transitions that induce exceptional transition models. We formalize this as an exceptional model mining task.

For this purpose, we first derive a dataset D of transitions with model attributes A_M and describing attributes A_D (see Section 5.2.1). These allow to form a large set of candidate subgroup descriptions. For each corresponding candidate subgroup g, we then determine the corresponding set of transitions and compute its transition count matrix $\boldsymbol{T}_g$. By comparing this matrix to a reference matrix $\boldsymbol{T}_D$ derived from the entire data D, we can then calculate a score according to an interestingness measure q (see Section 5.2.2). In order to detect the subgroups with the highest scores, standard exceptional model mining search algorithms are utilized to explore the candidate space (see Section 5.2.3). The automatically found subgroups then should be assessed by human experts (see Section 5.2.4). In a variation of our approach, we do not use the transition count matrix of the entire data $\boldsymbol{T}_D$ for comparison with the subgroup matrices $\boldsymbol{T}_g$, but instead employ a matrix $\boldsymbol{T}_H$ that expresses a user-specified hypothesis. This allows for finding subgroups that specifically match or contradict this hypothesis (see Section 5.2.5).



(a) Sequence data with background knowledge

A_M A_D

Source State	**Target State**	Gender	Age	Hour	Weekday	# Visits of user
A	B	f	34	14	Tue	2
B	A	m	17	20	Sat	4
A	C	m	17	23	Sat	4
C	B	m	17	1	Sun	4
C	A	f	25	11	Wed	3
A	B	f	25	12	Wed	3
C	A	m	32	13	Sat	4
A	C	m	32	15	Sat	4
C	B	m	32	17	Sat	4

(b) Transition dataset

$$\begin{pmatrix} 0 & 2 & 2 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix}$$

(c) Transition count matrix for the entire dataset ($\boldsymbol{T}_D$)

$$\begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

(d) Transition count matrix for *Gender=f*

$$\begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

(e) Transition count matrix for *Weekday=Sat*

Figure 5.1.: Subgroups of sequential behavior. Sequential data with background information (a) is initially transformed to a transition dataset with transition model attributes A_M and descriptive attributes A_D (b). To discover interesting subgroups, transition matrices for the entire dataset (c) and for the candidate subgroups, e.g., *Gender=f* (d) or *Weekday=Sat* (e), are computed and compared with each other.

5.2.1. Data representation

We consider sequences of states and additional background information about them (cf. Section 3.1). Since we will perform exceptional model mining on a *transition level*, we split the given state sequences in order to construct a tabular dataset, in which each instance corresponds to a single transition (similar to Section 4.2.1). For each instance, the source and target state represent the values of the model attributes A_M from which the model parameters, i.e., the transition matrix of the Markov chain model, are derived. Each instance is also associated with a set of describing attributes A_D based on the given background information.

Figure 5.1 (a-b) illustrates such a preparation process for a simple example. It shows sequences of states (e.g., certain locations) that users have visited and some background knowledge, i.e., some user information and the time of each visit (Figure 5.1a). This information is integrated in a single data table (Figure 5.1b). It contains two columns for the transition model attributes A_M, i.e., for the source and the target state of each transition. Additional describing attributes A_D capture more information on these transitions. This includes information specific to a single transition such as the departure time at the source state but also information on the whole sequence that is projected to all of its transitions, e.g., user data or the sequence length. Example subgroup descriptions that can be expressed based on these attributes are *"all transitions by female users"*, *"all transitions on Saturdays"*, or combinations such as *"all transitions between 13:00h and 14:00h from users older than 30 years that visited at least three locations"*. As different types of information can be considered for the construction of the descriptive attributes, the approach is very flexible.

5.2.2. Interestingness measure

We aim to find subgroups that are interesting with regard to their transition models. For quantifying interestingness, we employ an *interestingness measure* q that assigns a score to each candidate subgroup. The score is based on a comparison between the transition count matrix of the subgroup $\boldsymbol{T}_g$ and a reference transition count matrix $\boldsymbol{T}_D$ that is derived from the overall dataset. In short, the interestingness measure that we propose expresses *how unusual the distance between the transition matrix of a subgroup and the reference matrix is in comparison to transition matrices of random samples from the overall dataset*. For that purpose, we first define a distance measure on transition matrices. Then, we show how this distance can be compared against transition matrices built from random dataset samples. We describe those two steps in detail before discussing more specific issues.

Distance measure and weighting. First, we compute the reference transition count matrix $\boldsymbol{T}_D = (d_{i,j})$ for the overall dataset D as exemplified in Figure 5.1c. To evaluate a subgroup g, all instances in the tabular dataset that match its subgroup description are identified and a transition count matrix $\boldsymbol{T}_g = (g_{i,j})$ is determined accordingly (see, e.g., Figure 5.1d and Figure 5.1e). Then, a distance measure is employed to measure the difference of transition probabilities in these matrices. After normalizing both

matrices $\boldsymbol{T}_D$ and $\boldsymbol{T}_g$ by row (yielding transition probability matrices $\boldsymbol{\theta}_D = (d_{i,j}/\sum_j d_{i,j})$ and $\boldsymbol{\theta}_g = (g_{i,j}/\sum_j g_{i,j})$, cf., Section 3.2), each row i represents a conditional categorical probability distribution for the next state given state s_i. In literature, several methods have been proposed to compare such distributions. Here, we focus on the *total variation distance* δ_{tv}, also called *statistical distance* or (excluding the constant factor) *Manhattan distance*. For one row, this is computed as the sum of absolute differences between the normalized row entries, i.e., between transition probabilities:

$$\delta_{tv}(g, D, i) = \frac{1}{2} \sum_j \left| \frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{d_{i,j}}{\sum_j d_{i,j}} \right| \tag{5.1}$$

We then aggregate this value over all states (matrix rows). Since in our setting differences in states with many observations in the subgroup should be more important than those with less observations, we weight the rows with the number of transitions $w_i = \sum_j g_{i,j}$ from the corresponding source state s_i in the subgroup:

$$\omega_{tv}(g, D) = \sum_i \left(w_i \cdot \sum_j \left| \frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{d_{i,j}}{\sum_j d_{i,j}} \right| \right) \tag{5.2}$$

The factor $\frac{1}{2}$ can be omitted as it is constant across all subgroups. States that do not occur in a subgroup are weighted with 0 and can be ignored in the computation even if transition probabilities are formally not defined in this case.

As an example, consider the transition matrix for the entire example dataset (Figure 5.1c) and the one for the subgroup $Gender = f$ (Figure 5.1d). The weighted total variation for this subgroup is computed as follows: $\omega_{tv}(Gender = f, D) = 2 \cdot (|\frac{0}{2} - \frac{0}{4}| + |\frac{2}{2} - \frac{2}{4}| + |\frac{0}{2} - \frac{2}{4}|) + 0 \cdot \text{NA} + 1 \cdot (|\frac{1}{1} - \frac{2}{4}| + |\frac{0}{1} - \frac{2}{4}| + |\frac{0}{1} - \frac{0}{4}|) = 3$.

Of course, there are also alternatives to the total variation distance measure that we can use, e.g., the *Kullback-Leibler divergence* $\delta_{kl}(g, D, i) = \sum_j g_{i,j} \cdot \log \frac{g_{i,j}}{d_{i,j}}$ or the *Hellinger distance* $\delta_{hell}(g, D, i) = \frac{1}{\sqrt{2}} \sqrt{\sum_j (\sqrt{g_{i,j}} - \sqrt{d_{i,j}})^2}$. However, for SubTrails, we focus on the weighted total variation as it naturally extends existing approaches for interestingness measures from classical pattern mining: it can be considered as an extension of the *multi-class weighted relative accuracy* measure for multi-class subgroup discovery [2]. Additionally, it can also be interpreted as a special case of *belief update* in a Bayesian approach as it has been proposed by Silberschatz and Tuzhilin [449] for traditional pattern mining. We provide a proof for this in Appendix D. Despite this focus, we also conducted a large set of experiments with all three distance measures in parallel with overall very similar results.

Comparison with random samples. The measure ω_{tv} describes a weighted distance between transition matrices. Yet, it is heavily influenced by the number of transitions covered by a subgroup. For example, small subgroups might be over-penalized by small weighting factors w_i, while very large subgroups can be expected to reflect the distribution of the overall dataset more precisely. Thus, using ω_{tv} directly as an interestingness measure does not consistently allow for identifying subgroups that actually influence transition behavior in presence of noise attributes, cf. Section 5.3.2.

To account for these effects, we propose a sampling-based normalization procedure. First, we compute the weighted distance $\omega_{tv}(g, D)$ of the subgroup g to the reference matrix as described before. Then, we draw a set of r random sample transition datasets $R = \{R_1, \ldots, R_r\}$, $R_i \subset D$ from the overall dataset D without replacement[1], each containing as many transitions as the evaluated subgroup g. Now, we compute the weighted distances $\omega_{tv}(R_i)$ for each of these samples, and build a *distribution of false discoveries* (cf. Duivesteijn and Knobbe [151]) from the obtained scores. In particular, we compute the mean value $\mu(\omega_{tv}(R_1, D), \ldots, \omega_{tv}(R_r, D))$ and the sample standard deviation $\sigma(\omega_{tv}(R_1, D), \ldots, \omega_{tv}(R_r, D))$ for the distances of the random samples. A subgroup is considered as interesting if the distance of the subgroup strongly deviates from the distances of the random samples. We quantify this by a (marginally adapted) *z-score*, which we will use as the interestingness measure q in our approach:

$$q_{tv}(g, D) = \frac{\omega_{tv}(g, D) - \mu(\omega_{tv}(R_1, D), \ldots, \omega_{tv}(R_r, D))}{\sigma(\omega_{tv}(R_1, D), \ldots, \omega_{tv}(R_r, D)) + \epsilon}, \tag{5.3}$$

with ϵ being a very small constant to avoid divisions by zero. Thus, $q_{tv}(g, D)$ quantifies how unusual the difference of the transition matrix of the subgroup g and the reference matrix is *compared to a random set of transitions drawn from the overall data that contains the same number of transitions.*

Stratification of samples. When drawing random samples equally across all states, high scores q_{tv} can exclusively be caused by a peculiar distribution of source states in a subgroup. However, this is not desirable when studying transition behavior. Consider, e.g., a dataset D, where transitions for all but one source state (matrix rows) are deterministic (the transition probability is 1 for a single target state), and all source states have the same number of observed transitions. Then, random transition samples R_i will be drawn mostly from the deterministic states and thus, will consistently have very small weighted distances $\omega_{tv}(R_i, D)$. Now, if any subgroup g only contains transitions from the non-deterministic source state, a random deviation from the underlying transition probabilities is likely. Yet, even if this deviation and thus the distance $\omega_{tv}(g, D)$ is small on an absolute scale, this distance would still be higher than the ones of the random samples. As a consequence, g appears as an exceptional subgroup with respect to its transition probabilities, even if only the distribution of source states differs.

To address this issue, we adapt our sampling procedure: we do not use simple random sampling, but instead apply stratified sampling w.r.t. the source states of the transitions. Thus, we draw the random samples $R_1, \ldots, R_r$ in such a way that for each source state in the data, each random sample contains exactly as many transitions as the evaluated subgroup. Note, that we do *not* stratify with respect to the target states since a different distribution of these states signals different transition behavior.

Significance. To ensure that our findings are not only caused by random fluctuations in the data, the z-score q_{tv} which we employ as our interestingness score can be used as a test statistic for a z-test on statistical significance. Yet, this test requires a normal

[1]The rationale for using sampling *without replacement* is that the subgroup itself also cannot contain multiple instances of the same transition.

distribution of the weighted distances $\omega_{tv}(R_i, D)$ obtained from the samples. Although in many practical situations the distribution of the sampled distances is *approximately* normally distributed, this does not necessarily hold in all cases. We thus propose a two-step approach to assess statistical significance of the results. First, we use a normality test such as the *Shapiro-Wilk-Test* [444] on the set of distance scores obtained for the sample set R. If the test does not reject the assumption of normality, a p-value can be directly computed from the z-score. If normality is rejected, a substantially larger set of random samples can be drawn to compute the *empirical p-value* of a specific subgroup [204], i.e., the fraction of samples that show a more extreme distance score than the subgroup. Although this is computationally too expensive to perform for every single candidate subgroup, it can be used for confirming significance for the most interesting subgroups in the result set.

For both methods one must consider the *multiple comparison problem* [243]: if many different subgroups are investigated (as it is usually done in pattern mining), then some candidates will pass standard significance tests with unadapted significance values by pure chance. Therefore an appropriate correction such as *Bonferroni correction* [154] or *layered critical values* [514] must be applied.

Estimate the effect of limited sample numbers. Determining the interestingness score $q_{tv}(g, D)$ requires to choose a number of random samples r. While fewer samples allow faster computation, results might get affected by random outliers in drawn samples. To estimate the potential error in the score computation caused by the limited number of samples, we employ a *bootstrapping approach* [159]: we perform additional sampling on the weighted distances of the original samples $S = \{\omega_{tv}(R_1, D), \ldots, \omega_{tv}(R_r, D)\}$. From this set, we repeatedly draw (e.g., $10,000$ times) *"bootstrap replications"*, i.e., we draw r distance values by sampling *with* replacement from S and compute the subgroup score q_{tv} for each replication. The standard deviation of the replication scores provides an approximation of the standard error compared to an infinitely large number of samples, cf. [160]. In other words, we estimate how precise we compute the interestingness score q_{tv} with the chosen value of r compared to an infinite number of samples. If the calculated standard error is high compared to the subgroup score, re-computation with a higher number of samples is recommended.

5.2.3. Subgroup search

To detect interesting subgroups, we enumerate all candidate subgroups in the search space in order to find the ones with the highest scores. For this task, a large variety of mining algorithms has been proposed in the pattern mining literature featuring exhaustive as well as heuristic search strategies, e.g., depth-first search [282], best-first search [515, 569], or beam-search [301, 493]. For this study, we do not focus on efficient algorithms for exceptional model mining, but apply a depth-first mining algorithm as a standard solution.

Candidate evaluation in our approach is computationally slightly more expensive than for traditional subgroup discovery. That is, the runtime complexity for determining the score of a single subgroup in our implementation is $O(r \cdot (N + S^2))$ for a dataset with N

transitions, S different states, and a user chosen parameter of r samples: selecting the set of instances from a subgroup as well as drawing a stratified sample requires $O(N)$ operations per subgroup and sample. The transition matrices for each of these transition sets can also be built in linear time. The weighted distance for each of the r samples and the subgroup can then be determined in $O(S^2)$ as a constant number of operations is required for each of the S^2 matrix cells.

A typical problem in pattern mining is redundancy, i.e., the result set often contains several similar subgroups. For example, if the subgroup *male* induces an exceptional transition model and thus achieves a high score, then also the subgroup *males older than 18* can be expected to feature a similarly unusual model and receive a high score—even if age does not influence transition behavior at all. A simple, but effective approach to reduce redundancy in the result set is to adapt a *minimum improvement constraint* [39] as a filter criterion. To that end, we remove a subgroup from the result set if the result also contains a generalization, i.e., a subgroup described by a subset of conditions, with a similar (e.g., less than 10% difference) or a higher score.

5.2.4. Subgroup assessment

Automatic discovery algorithms with the proposed interestingness measure can detect subgroups with exceptional transition models. Yet, to interpret the results, manual inspection and assessment of the top findings is crucial as this allows users to identify in what aspects the found "interesting" subgroups differ from the overall data. For that purpose, a comparison between the subgroup transition matrix and the reference matrix is required. Yet, manual comparison can be difficult for large matrices (state spaces). Therefore, we recommend to assess subgroups with summarizing *key statistics*, such as the number of transitions in a subgroup, the weighted distance ω_{tv} between subgroup and reference transition matrices, the unweighted raw distance $\Delta_{tv} = \sum_i \delta_{tv}(g, D, i)$, or the distribution of source and target states. Additionally, *exemplification*, e.g., by displaying representative sequences, and *visualizations* are helpful tools for subgroup inspection. See Sections 7.4.2 and 11.3 for two visualiation examples in the context of music listening behavior and geo-spatial navigation, respectively.

5.2.5. User-defined hypotheses

In addition to comparing subgroups to the overall dataset, our approach can also detect subgroups that specifically contradict or match a user-defined hypothesis. Following the concepts of Singer et al. [453], we can express such a hypothesis as a *belief matrix* $\boldsymbol{T}_H = (h_{i,j})$, where higher values $h_{i,j}$ indicate a stronger belief in transitions from state s_i to state s_j. An example of a hypothesis considering the example dataset of Figure 5.1 could be stated as: $\left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{smallmatrix}\right)$. This hypothesis formalizes a belief that users from state A (first row) will always go to state B, and users from the states B and C will proceed to any of the three states with equal probability.[2]

[2]Note that it is equivalent to formulate a transition count matrix $\boldsymbol{T}_H = (h_{i,j})$ or a transition probability matrix $\boldsymbol{\theta}_H = (h_{i,j}/\sum_j h_{i,j})$ because the weighted total variation ω_{tv} (cf., Section 5.2.2) normalizes the

Now, given a belief matrix $\boldsymbol{T}_H$ of a hypothesis, the interestingness score of a subgroup is computed analogously to the original case, but instead of using the transition matrix derived from the overall dataset $\boldsymbol{T}_D$ as reference, we use the belief matrix $\boldsymbol{T}_H$ of the hypothesis for the computation of the weighted distance ω_{tv}. A subgroup g (exceptionally) contradicts the hypothesis, if its transition matrix $\boldsymbol{T}_g$ has a significantly larger distance to the hypothesis matrix $\boldsymbol{T}_H$ than the stratified random samples of the dataset. To find subgroups that *match* a hypothesis specifically well instead of contradicting it, the inverted interestingness measure $-q_{tv}(g, D)$ can be used instead.

5.3. Experiments

Here, we demonstrate the potential of our approach with synthetic data. For empirical data illustrating possible application scenarios and findings, please see Sections 7.4.2 and 11.3. Using the synthetic data, we show that our approach is able to recover (combinations of) conditions that determine the transition probabilities in presence of noise attributes. For computing the interestingness measure, we used $r = 1,000$ random samples. Our implementation (an extension of the VIKAMINE data mining environment [24]) and the synthetic datasets are publicly available.[3]

5.3.1. Random transition matrices

We start with a synthetic dataset directly generated from two first-order Markov chain transition matrices each representing a navigational sub-process. Transitions from both matrices combined will make up the overall observed behavior. We aim to discover subgroups with solely transitions from one or the other, pinpointing the corresponding navigational sub-processes.

Experimental setup. We created two 5×5 matrices of transition probabilities by inserting uniformly distributed random values in each cell and normalizing the matrices row-wise. Then, for each generated instance, one of the matrices was chosen based on two attributes, a ternary attribute A and a binary attribute B. If both attributes take their first values, i.e., $A = A1$ and $B = B1$, then transitions were generated from the first matrix, otherwise from the second matrix. For each combination of values, we generated $10,000$ transitions, resulting in 60,000 transitions overall. For each transition, we additionally generated random values for 20 binary noise attributes, each with an individual random probability for the value *true*. We employed our approach with a maximum search depth of two selectors to find subgroups with different transition models compared to the overall dataset. Our approach should then detect the subgroup $A = A1 \wedge B = B1$ as the most relevant one.

Results. The top-5 result subgroups are displayed in Table 5.1. It shows the number of covered transitions (instances), the score of the interestingness measure q_{tv} including

entries of the reference matrix $\boldsymbol{T}_D$ (i.e., $\boldsymbol{T}_H$ for this hypothesis based variation) and the weights w_i only depend on the values of the transition count matrix $\boldsymbol{T}_g$ of the subgroup g.

[3] `http://florian.lemmerich.net/paper/subtrails.html`

Table 5.1.: Top subgroups for random transition matrix data. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv}, and the unweighted total variation Δ_{tv}.

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
A = A1 ∧ B = B1	10,000	113.01 ± 2.74	5,783	1.54
A = A1	20,000	67.23 ± 1.60	4,634	0.60
B = B1	30,000	45.52 ± 0.94	3,480	0.33
B = B2	30,000	44.69 ± 1.08	3,480	0.51
A = A3	20,000	32.05 ± 0.77	2,378	0.53

the standard error of its computation estimated by bootstrapping (±), the weighted total variation ω_{tv} between the subgroup and the reference transition matrix, and its unweighted counterpart Δ_{tv}. The result tables for the following experiments will be structured analogously.

We observe that our approach successfully recovered the subgroup of transitions that were generated from a different probability matrix, i.e., the subgroup $(A = A1 \wedge B = B1)$. This subgroup receives the best score q_{tv} by a wide margin. The subgroup with the next highest score $(A = A1)$ is a generalization of this subgroup. Since it contains transitions generated from both matrices in a different mixture, it also features indeed an unusual transition model compared to the entire dataset. In the same way, the next subgroups all feature the attributes A and B that actually influence the transition behavior, and none of the noise attributes. These top subgroups all pass a Bonferroni-adjusted statistical significance test as described in Section 5.2.2 with an empirical p-value of $p \ll 10^{-10}$, while all subgroups containing only noise attributes (not among the shown top subgroups) do not pass such a test with a critical value of $\alpha = 0.05$.

5.3.2. Random walker

Our second demonstration example features a set of transitions generated by a random walker in a network of colored nodes.

Experimental setup. First, we generated a scale-free network consisting of 1,000 nodes (states) with a Barabási-Albert model [35]. That is, starting with a small clique of nodes, new nodes with degree 10 were inserted to the graph iteratively using preferential attachment. Then, we assigned one of ten colors randomly to each node. On this network, we generated $200,000$ sequences of random walks with five transitions each, resulting in $1,000,000$ transitions overall. For each sequence, we randomly assigned a walker type. With a probability of 0.8, the walk was purely *random*, i.e., given the current node of the walker, the next node was chosen with uniform probability among the neighbouring nodes. Otherwise, the walk was *homophile*, i.e., transitions to nodes of the same color were twice as likely. For each transition, the resulting dataset contains the source node, the target node, the type of the respective walker (random or homophile), and additionally the values for 20 binary noise attributes, which were assigned with an individual random probability each.

Table 5.2.: Top subgroups for random walker data. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv}, and the unweighted total variation Δ_{tv}.

(a) Comparison to the overall dataset.

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	35.67 ± 0.78	51,929	125.96
Type = Random	799,085	34.34 ± 0.80	51,929	31.73
Noise9 = False	681,835	2.25 ± 0.06	51,358	36.27
Noise9 = True	318,165	2.23 ± 0.06	51,358	77.94
Noise2 = False	18,875	1.80 ± 0.05	14,844	394.51

(b) Comparison to the *homophile* hypothesis, contradicting.

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Random	799,085	26.88± 0.57	1,554,130	981.38
Noise4 = True	519,130	2.28 ± 0.06	1,008,912	981.25
Noise2 = False	18,875	2.25 ± 0.06	37,057	987.49
Noise1 = True	469,290	2.00 ± 0.05	912,032	981.26
Noise19 = True	342,765	1.93 ± 0.05	666,229	981.28

(c) Comparison to the *homophile* hypothesis, matching.

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	12.10 ± 0.27	389,841	981.04
Noise4 = False	480,870	2.69 ± 0.07	934,190	981.20
Noise19 = False	657,235	2.27 ± 0.06	1,276,868	981.20
Noise1 = False	530,710	1.99 ± 0.05	1,031,101	981.20
Noise0 = True	523,410	1.74 ± 0.05	1,016,899	981.21

With this data, we performed three experiments. In the first, we searched for subgroups with different transition models compared to the entire data. In the second and third experiment, we explored the option of finding subgroups that contradict — respectively match — a hypothesis. For that purpose, we elicited a hypothesis matrix $\boldsymbol{T}_H = (h_{i,j})$ that expresses belief in walkers being *homophile*, i.e., transitions between nodes of the same color are more likely. Towards that end, we set a matrix value $h_{i,j}$ to 1 if i and j belong to the same color and $h_{i,j} = 0$ otherwise. Edges of the underlying network were ignored for the hypothesis generation.

Results. Table 5.2 presents the results for the three experiments. As intended, exceptional model mining identified the subgroups that influence the transition behavior as the top subgroups for all three tasks. In the first experiment (see Table 5.3a), both subgroups described by the *Type* attribute are top-ranked. For the second experiment (see Table 5.3b), the subgroup *Type=Random* receives the highest score. By construction, this subgroup should indeed expose the least homophile behavior since any subgroup described by noise attributes contains transitions from homophile as well as non-homophile walkers. Its complement subgroup *Type=Homophile* does not contradict our hypothesis and thus does not appear in the top subgroups. By contrast and as expected, the subgroup

Type=Homophile receives the highest score in the third experiment that searches for subgroups matching the homophile hypothesis, while *Type=Random* is not returned as a top result, cf. Table 5.3c. For all three experiments, the statistical significance of the top subgroups described by the *Type* attribute was decisive ($p \ll 10^{-10}$), while the top findings for the noise attributes were not significant at the Bonferroni-adjusted level $\alpha = 0.05$.

In additional experiments (no result tables shown), we employed the weighted distance ω_{tv} directly as an interestingness measure. By doing so, we were not able to recover the relevant subgroups as they were dominated by several random noise subgroups. This shows the necessity of a comparison with random samples.

We also experimented extensively with different parameterizations (e.g., different walker type probabilities or different numbers of node colors). Consistently, we were able to identify the two subgroups *Type=Random* and *Type=Homophile* as the top subgroups.

5.4. Related work

Mining patterns in sequential data has a long history in data mining. However, large parts of research have been dedicated to the tasks of finding frequent sub-sequences efficiently, see for example [8, 366, 551]. Also see Section 3.4.2.2 for more information. Other popular settings are sequence classification [315, 540] and sequence labeling [299]. However, unlike SubTrails, these methods do not aim to detect subgroups with unusual transition behavior.

Our solution is based on exceptional model mining as introduced in Section 3.4. This data mining task aims at finding descriptions of data subsets that show an unusual statistical distribution of arbitrary target concepts. While many model classes have been studied (e.g., classification [307] and regression models [150], or Bayesian networks [152]). No models featuring sequential data have been explored for exceptional model mining so far. Also, see Section 3.4.2 for a general overview of applications of subgroup discovery and EMM.

We presented an approach to detect subgroups with exceptional transition models, i.e., subgroups that show unusual distributions of the target states in first-order Markov chain models. The results from our approach may correlate with subgroups that could also be obtained by multi-class subgroup discovery [2] that investigates the distribution of target states. However, such a static analysis aims to achieve a different goal than our analysis of behavior dynamics and will not capture all subgroups with exceptional transition models. For example, in the random walker synthetic dataset (see Section 5.3.2) the distribution of target states is approximately uniform for all subgroups by construction, also for the ones that influence the transition behavior. As a consequence and in contrast to our method, a static analysis could not recover the exceptional subgroups. Furthermore, the task of finding subgroups that match or contradict a hypothesis of dynamic state transitions (e.g., as demonstrated in the Flickr example, see Section 7.4.2) cannot be formulated as a more traditional subgroup discovery task.

Our interestingness measure is inspired by previous methods. The weighted distance measure can be considered as an adaptation of the multi-class weighted relative accuracy [2]

or as a special case of the Bayesian belief update [449]. The randomization/sampling processes to capture significant differences of subgroups also builds upon previous approaches. In that direction, Gionis et al. [204] utilized swap randomization to construct alternative datasets in order to ensure the statistical significance of data mining results. For subgroup discovery, analyzing a distribution of false discoveries obtained by randomization was proposed to assess subgroups and interestingness measures [151]. We extended these methods to exceptional model mining with complex targets and used it directly in the interestingness measure for the subgroup search.

For modeling sequential processes, Markov chains were used in various forms and in a wide variety of applications ranging from user navigation [402, 454] to economical settings and meteorological data [186] (also see Section 3.2.2). The *mixed Markov model* extension [405] of classical Markov chains features separate transition matrices for "segments" of users, but these segments are not interpretable, i.e., have no explicit descriptions. The work maybe closest to ours is [426], where the authors detected outliers of user sessions with respect to their probability in a Markov-chain model; outliers were then manually categorized into several interpretable groups. By contrast, our solution allows to identify descriptions of groups that show unusual transition behavior automatically from large sets of candidate subgroups.

Also, compared to HypTrails [453] (Section 3.3.2) and MixedTrails (Chapter 4) which allow to compare hypotheses about Markov chain models, SubTrails — as an exceptional model mining approach — not only enables us to find interpretable sub-processes in human navigation behavior but also let's us identify (sets of) conditions under which a given hypothesis is matched or contradicted.

5.5. Conclusion

With SubTrails we proposed a pattern mining approach to exploring heterogeneous aspects of human navigation behavior, supporting the process of hypotheses conception (cf. Section 1.2.2). In particular, we introduced first-order Markov chains as a novel model class for exceptional model mining in sequence data with background knowledge. This enables a novel kind of analysis: it allows to detect interpretable subgroups that exhibit exceptional transition behavior, i.e., induce different transition models compared to the entire dataset. In addition, we presented a variation of the standard task that compares subgroups against user-defined hypotheses, enabling a detailed analysis of given hypotheses about transition behavior. We illustrated the potential of our approach by applying it to several, advanced synthetic scenarios, i.e., SubTrails successfully recovered exceptional transitions from artificial noise attributes. For insights gained by applying SubTrails to real-world data please see Sections 7.4.2 and 11.3. Overall, SubTrails presents a novel approach to gain an understanding of the underlying heterogeneous processes of human navigation, and will ultimately enable to formulate and compare more intricate hypotheses by incorporating the corresponding heterogeneous aspects. Thus, it addresses one of the major challenges of analyzing heterogeneous navigation behavior as outlined in Section 1.2, namely *hypothesis conception*. An example of this process can be seen in

Chapter 7, where subgroups discovered by SubTrails are used as indicators to formulate new heterogeneous hypotheses about photowalking behavior.

In the future, we aim to improve and extend our approach in several directions. First, the proposed interestingness measure is currently based on individual transitions. As a consequence, a few very long sequences (e.g., of very active users) can strongly influence the results. To avoid dominance of such sequences, weighting of the transition instances according to the overall activity could be applied in future extensions [cf. 22]. In addition, we intend to investigate ways of speeding-up the mining process, e.g., by optimistic estimate pruning [532] or by using advanced data structures [310], and aim to apply sophisticated options to reduce redundancy, cf. [311, 318, 319]. Finally, we would like to generalize the proposed model class to Markov chains of higher order or even more advanced sequential models that potentially also take indirect state transitions into account.

6. Analysis tools

Besides our methodological contributions in Chapters 4 and 5, we have developed several tools to support the process of understanding of human navigation behavior. This ranges from introducing an efficient implementation of relevant algorithms to developing several data collection, analysis, and visualization systems. In particular, we present three analysis tools: SparkTrails, VizTrails, and the EveryAware platform. *SparkTrails* (Section 6.1) is a distributed implementation of the HypTrails approach [453] based on the MapReduce paradigm [137] for comparing hypotheses about human navigation behavior (cf., Section 1.2.1). It allows to efficiently handle real-world scenarios with large state spaces as often encountered when studying human navigation behavior. *VizTrails* (Section 6.2) supports the process of hypothesis conception (cf. Section 1.2.2) by providing visualizations of geo-spatial navigation data. It facilitates deeper insights into the corresponding trajectories by enabling interactive exploration of aggregated statistics and providing geo-spatial context. The *EveryAware* system (Section 6.3) takes a more holistic approach and provides a platform for collecting mobile sensor data in a participatory setting. That is, it enables user-driven campaigns by collecting, analyzing and visualizing data such as air quality or noise pollution in a geo-spatial context while explicitly supporting mobile, personal devices, and subjective information, such as emotions or perceptions. Similar to VizTrails the explorative nature of EveryAware aids the processes of hypothesis conception (cf. Section 1.2.2). In the following, we present each of these systems in detail.

6.1. SparkTrails: A MapReduce implementation of HypTrails for comparing hypotheses about human trails

This section presents a distributed and parallel implementation of HypTrails (cf. Section 3.3.2) which enables the comparison of hypotheses about the underlying processes of human navigation on large scale datasets and state spaces (cf., Section 1.2.1). Many of our case studies in Part III rely on this approach. The content of this section follows our previously published work on SparkTrails [42].

6.1.1. Introduction

As reviewed in Section 3.3.2, HypTrails [453] (cf., Section 3.3.2) is a Bayesian approach for formulating and comparing hypotheses about the underlying processes of human navigation behavior. However, especially recently, real-world datasets of geo-spatial as well as online navigation are often very large. This requires approaches for analyzing such data to operate efficiently. While a standard implementation of HypTrails is available, it exposes performance issues when working with large-scale data.

To address this, we take advantage of the structural properties of HypTrails and propose a fast, scalable, and distributed implementation, called *SparkTrails*, based on the MapReduce paradigm [137]. We implement our method on Apache Spark and evaluate our approach on several large-scale datasets observing greatly improved performance and the ability to scale freely. The implementation is publicly available and open source.[1]

In the following, we first examine the computational structure of HypTrails (Section 6.1.2) and exploit our findings in the subsequent section in order to derive the algorithmic details of our SparkTrails approach (Section 6.1.3). Afterwards, we evaluate SparkTrails on several large-scale datasets in Section 6.1.4, and conclude in Section 6.1.5. With regard to related work, we note that HypTrails [453] is a relatively novel method, and thus, no other studies on its performance or corresponding parallelized or distributed implementations exist so far.

6.1.2. Computational structure of HypTrails

As detailed in Section 3.3.2, HypTrails is a Bayesian approach for formulating and comparing a set of hypotheses about the underlying processes of human navigation behavior based on a set of observations. In this context, observations are represented as a path dataset D which is converted to a transition count matrix $\boldsymbol{T} = (n_{i,j})$ where $n_{i,j}$ corresponds to the transition count from state i to state j (cf. Definitions 1 and 3). Given this data, hypotheses are compared based on their marginal likelihood $P(D|H)$. In particular, hypotheses are formulated as stochastic matrices $\boldsymbol{\phi} = (\phi_{i,j})$ with each entry representing the transition probability from one state to the other. These stochastic matrices are transformed into parameters of a Dirichlet distribution by employing a concentration factor κ (cf. Section 3.3.2.3). This results in a parameter matrix $\boldsymbol{\alpha} = (\alpha_{i,j})$. Now, let Γ denote the gamma function, then the overall formula to calculate the marginal likelihood $\Pr(D|H)$ of a hypothesis H with a given parameter matrix $\boldsymbol{\alpha}$ is:[2]

$$\Pr(D|H) = \prod_i \underbrace{\frac{\Gamma(\sum_j \alpha_{i,j})}{\prod_j \Gamma(\alpha_{i,j})} \frac{\prod_j \Gamma(n_{i,j} + \alpha_{i,j})}{\Gamma(\sum_j (n_{i,j} + \alpha_{i,j}))}}_{\text{evidence } \Pr_i(D|H) \text{ for an individual state } i} \tag{6.1}$$

This formula has to be calculated several times for each hypothesis depending on the number of concentration factors used to construct the evidence curves HypTrails uses for comparing hypotheses (cf. Figure 3.6 in Section 3.3.2.1). Note that the number of terms in Equation (6.1) grows quadratically with an increasing number of states. Thus, real world examples with large state spaces can — in addition to memory issues — lead to very long runtimes.

[1] `http://dmir.org/sparktrails`

[2] See Section 3.3.2.2 for an explanation on why $\boldsymbol{\alpha}$ can encode hypotheses.

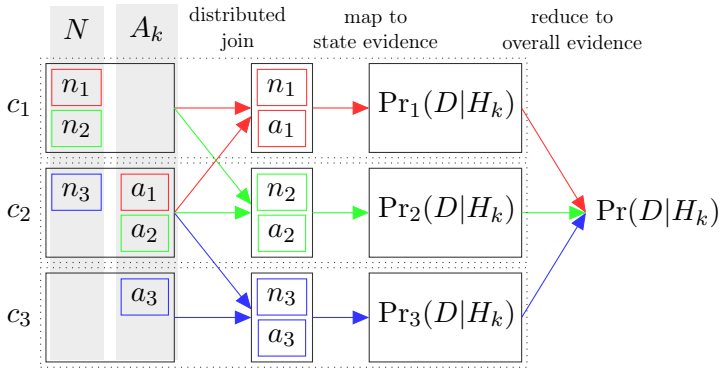


Figure 6.1.: SparkTrails concept. A schematic distributed calculation of HypTrails for three states. c_i are computational nodes where the rows of the observation matrix $\boldsymbol{T}$ and the elicited hypothesis matrix $\boldsymbol{\alpha}$ are stored in a distributed fashion. After joining these two matrices by row, each computation node calculates the evidence for one (or more) state. The resulting state evidences are then merged into the overall evidence.

6.1.3. Distributed implementation

To be able to cope with large state spaces, for HypTrails, we implement SparkTrails where we employ a MapReduce approach in order to distribute the workload as well as the memory requirements of the HypTrails method across several computational nodes. In this section we introduce the general idea and discuss several optimizations.

Main idea. The overall evidence calculated by HypTrails corresponds to the product of the evidences of each individual state (cf. Equation 6.1). Thus, we calculate these state evidences individually in a distributed fashion and merge the results into the overall evidence.[3] The process, as illustrated in Figure 6.1, can be embedded into the MapReduce paradigm as indicated by the distributed join as well as the map and reduce steps.

Row Sparsity. Observations are often sparse resulting in many states with no outgoing transitions. For these states all $n_{i,j}$ in $\Pr_i(D|H)$ are 0. Hence the components of the nominator and the denominator cancel each other out yielding an evidence of 1. Consequently these states can be left out of the distributed join (for $\boldsymbol{T}$ *and* $\boldsymbol{\alpha}$). This greatly reduces the amount of data being shuffled between nodes.

Column Sparsity. We further exploit the observation sparsity by working with sparse row vectors. For each state evidence calculation $\Pr_i(D|H)$, this lets us reduce the number of Γ values to calculate by two times the number of states transitions $n_{i,j}$ which have not been observed. This is because $\Gamma(n_{i,j} + \alpha_{i,j})$ and $\Gamma(\alpha_{i,j})$ cancel each other out if $n_{i,j} = 0$.

Belief. Since HypTrails calculates evidence values for several concentration factors κ, we would need to run it for each corresponding parameter matrix $\boldsymbol{\alpha}$ separately. However, we can distribute the transition probability matrix instead of the parameter matrix and

[3] To avoid underflow, we actually calculate the *logarithm* of the marginal likelihood (evidence) so we can sum values instead of multiplying them.

Table 6.1.: Runtimes of SparkTrails. The data is based on real-worl data from Wikipedia (w_{self}, w_{nw}) and Flickr (f) as well as several synthetic examples (s_1, s_2, r). In all cases we observe a strongly reduced runtime for the distributed algorithm (*spark*). Also, runtimes scale almost linearly when increasing the number of computation nodes (e).

	w_{self}	w_{nw}	f	s_1	s_2	r
Python	9.0m	20.1m	1.4m	-	-	-
Spark ($e = 4$)	0.4m	1.7m	3.4m	2.5h	9.7h	18.3h
Spark ($e = 8$)	0.2m	0.9m	1.7m	1.2h	4.8h	8.9h
Spark ($e = 16$)	0.1m	0.7m	1.2m	0.7h	2.7h	5.2h

move the elicitation process into the state evidence calculation.[4] This results in evidence vectors, one entry for each κ, avoiding expensive distributed joins.

More. Our implementation features additional optimizations, such as exploiting the row sparsity property mentioned above *for hypotheses* as well, taking advantage of their structural properties to avoid data shuffling, speeding up the distributed join via pre-sorting or even consider coordinate-wise instead of row-wise calculations in case of (unlikely) memory issues. See the code base for details.[5]

6.1.4. Experiments

For evaluation we calculate the evidence for 10 different concentration factors κ on synthetic as well as real-world data including Wikipedia navigation [537] and photo trails in Los Angeles (cf., Chapter 7). We test our distributed implementation based on Apache Spark and an optimized version of the original Python implementation. Table 6.1 lists the results for the multiplication based hypothesis elicitation variant (cf. Sections 3.3.2 and 6.1.2).

SparkTrails runs on a YARN cluster with 6 worker nodes à 6 physical Intel Xeon cores, 128GB RAM and 5 executors. The Python code is not parallelized and uses a 2.1GHz AMD Opteron CPU and 256GB RAM. For Python, the larger state spaces did not fit into memory accounting for missing runtimes, and we have not included the time to load data into memory (∼20 minutes for w_{nw}). For *SparkTrails* this time is included.

For Wikipedia, observations are transitions between articles from the clickstream dataset (Feb. 2015) by Wulczyn and Taraborelli [537]. The hypothesis w_{self} is based on the observed transitions themselves representing the optimal hypothesis. The alternative hypothesis w_{nw} is based on the link network[6] representing the hypothesis that people choose from available links uniformly. While the overall state count is larger than 45 million, the observations and the network are very sparse resulting in small runtimes. For photo trails (f), we consider transitions between geo-spatial grid-cells extracted from photo sequences on Flickr; the hypothesis is based on distance. The small runtime for Python can be explained by a small state count (∼84k), sparse observations and a dense

[4] If we choose an elicitation process which can be applied for each state independently. See Section 3.3.2.3 for more details on elicitation approaches.

[5] `http://dmir.org/sparktrails`, accessed: December 2017

[6] Based on an XML dump of the English Wikipedia from the 04.03.2015.

hypothesis. However, when considering the time to load data into memory (∼13m), SparkTrails is still a lot faster. To test our approach on dense data as well, we created a full transition matrix and used it as both, observations and hypothesis, with 93k (s_1) and 186k (s_2) states. Finally, we test on a randomly sampled matrix with 0.01% of all entries being set for 26 million states (r).

Additional information on the datasets as well as the different implementations can be found online.[7] Overall, we observe that our approach, *SparkTrails*, can handle larger datasets, yields dramatically smaller runtimes, and scales well with an increased number of computational nodes.

6.1.5. Conclusion

We proposed a distributed implementation of HypTrails (see Section 3.3.2). Our experiments showed that this implementation can handle large-scale data efficiently and outperforms non-distributed methods by a large margin. Furthermore, our approach scaled almost linearly with the number of computation nodes and thus, can handle very large observation datasets and hypotheses. In Chapter 7, we use SparkTrails to calculate the results for our study on human navigation based on Flickr data. Future work may include efficient methods for creating large hypotheses or adapting our implementation for possible extensions to HypTrails such as MixedTrails (cf. Chapter 4).

6.2. VizTrails: An information visualization tool for exploring geographic movement trajectories

In this section, we introduce VizTrails, a tool for visualizing geo-spatial navigation behavior in the context of various background information. This helps to better understand the underlying processes and supports the procedure of conceiving hypotheses about human navigation (cf. Section 1.2.2). The content of this section follows our previously published work on VizTrails [45].

6.2.1. Introduction

As listed in Section 2.1, many practitioners and researchers have studied human movement trajectories in cities through a variety of data sources including mobile phone data, GPS and Wifi tracking, location-based social media platforms, online photo sharing sites, and others. Our work in Chapter 7 extends this line of research by studying the underlying processes of a set of trails derived from human navigation behavior in the form of urban photo trails. To this end, we applied the HypTrails approach [453] (cf., Section 3.3.2) which allows to formulate and compare different hypotheses about the production of such trails. However, the process of formulating hypotheses is rather abstract because most of the time generalized mathematical formulas need to be used to efficiently formulate transition probabilities for each state combination (cf., Section 7.3).

[7] `http://dmir.org/sparktrails`

To support the process of formulating hypotheses and to mitigate its abstract nature, we have implemented a visualization tool called VizTrails[8], which allows us to better understand how geo-spatial navigational data (focusing on photo trails) materializes. It further enables us to gain further insights on how the specific hypotheses we formulate explain the corresponding paths. VizTrails achieves this by showing aggregated information for grid cells (or any other spatial discretization, e.g., tracts) on a map featuring interactive visualization of statistics, such as the number of users passing through cells, the in- and out-degree from and to other cells, or the cells commonly visited next. Also, among other features, VizTrails enables overlaying the map with content from arbitrary SPARQL queries for relating the observed trajectory statistics with geo-spatial context. VizTrails is designed for minimizing the required pre-processing steps.

Overall, VizTrails facilitates deeper insights into geo-spatial trajectory data by enabling interactive exploration of aggregated statistics in the context of additional geo-spatial context. Thus, it supports the process of formulating novel hypotheses about human navigation behavior (cf., Section 1.2.2). In the following, we present VizTrails including an overview of its architecture (Section 6.2.2), as well as several of its visualization features (Section 6.2.3). We follow up with a brief overview on related work in Section 6.2.4, and give a conclusion on Section 6.2.5.

6.2.2. Architecture

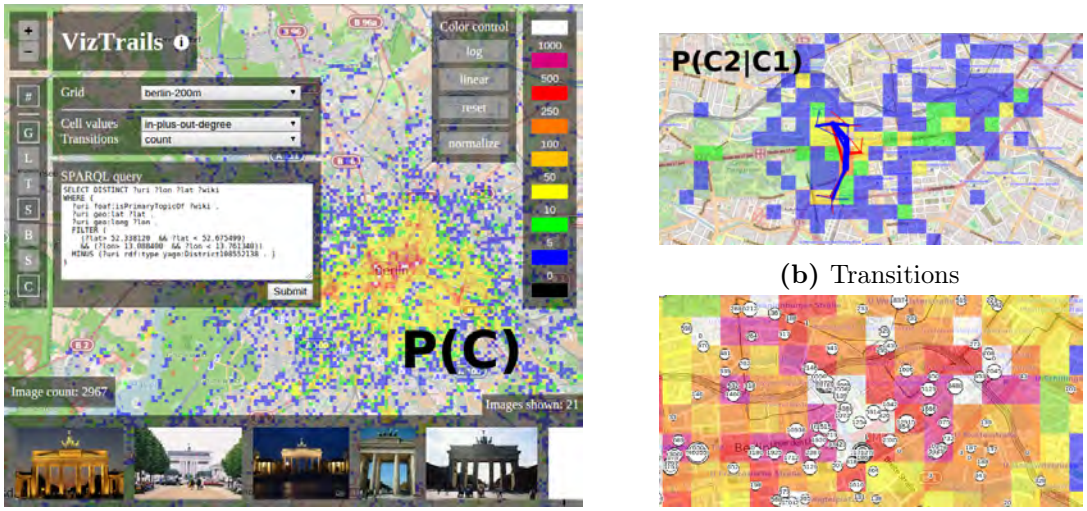
VizTrails is a web application based visualization system. It consists of two independent layers: the REST-layer for serving statistics on human navigation data and the UI-layer for visualizing the provided data.

The REST-layer is connected to a database and provides endpoints for accessing data points, user trajectories, grid cells, cell transitions, and more. It is built to be modular, i.e., the underlying database is easily exchangeable. Thus, it can not only serve data from relational databases like MySQL or PostgreSQL, but can also directly access data from distributed NoSQL databases like HBase or Cassandra. This is especially useful when large amounts of trajectory data are processed via parallel computation frameworks like Hadoop or Spark which directly write to such distributed data storage systems.

The UI-layer is browser-based. It pulls the data from the endpoints provided by the REST-layer and visualizes it via HTML, JavaScript, and corresponding frameworks like jQuery or OpenLayers. As a primary goal of VizTrails, the UI-layer enables data exploration in real-time. Since the listing of available grids and transitions is directly coupled with the REST-layer, new grid and transition types are immediately available in the user interface. This allows for a smooth workflow from generating and analyzing data towards visualizing it.

[8] `http://dmir.org/viztrails`

[9] Map data copyrighted OpenStreetMap contributors and available from `https://www.openstreetmap.org`.



(a) Grid View

(b) Transitions

(c) POIs

Figure 6.2.: VizTrails' visualization components. In (a) we show the general grid view visualizing different values for individual grid cells providing a general overview of some global statistics. In this case, photo counts (e.g., $P(C)$) in Berlin are depicted. (b) demonstrates how transitions from or to a cell are visualized when clicking on that particular cell (e.g., $P(C2|C1)$). This allows to explore how people move from or to different places. Third, (c) shows how entities from DBpedia and their respective view counts on Wikipedia are visualized providing trajectories with spatio-semantic context. These different visualization modes aid in exploring data about human movement trajectories in an intuitive and explorative way.[9]

6.2.3. Visualizations

We visualize geo-spatial trajectory data by discretizing an area defined by a bounding box into grid cells (or any other spatial discretization, e.g., tracts) as depicted in Figure 6.2a. Trajectories are then projected onto this grid. This allows us to visualize aggregated statistics on the set of all trails that contain a location within this grid cell. These include single cell statistics, cell transitions, and the respective geo-spatial context. In the following, we describe these visualizations in the same order.

Cell frequencies. For an overview of the general spatial distribution of the recorded data points, we color each grid cell according to the number of data points in that cell. The color as well as the value intervals associated with each color can be freely chosen. In addition to the number of data points in each cell, this visualization can be used to visualize any other scalar valued statistics depending on the values the discretization provides (in our case we also provide in- and out-degree for each cell). A dialog allows to choose from a number of different grids and associated values and updates as new grids are available in the database. Upon choosing a grid the map automatically pans and zooms to the appropriate extent.

Markov chain transitions. Now, in order to explore trajectories, the UI allows to visualize first-order Markov chain transitions. When clicking on a cell, cell colors change from a coloring based on overall statistics, to colors associated with the count of transitions starting at a point within the clicked cell. We also show lines for the most probable trails from (red) or to (blue) that cell. Thus, for example in the Flickr case, it can easily be judged where people will go from the current cell in order to take their next picture. Figure 6.2b shows the transitions from the "Brandenburg Gate" in Berlin. Here people mostly move towards three destinations, namely the "Reichstag" building, the "Potsdamer Platz" and the "Museum Island". Note, that this feature not only allows to visualize actual trajectories, but can also be used to contrast them with hypotheses about transitions as formulated, e.g., by Becker et al. [44] (also see Chapter 7).

Spatio-semantic context. In [44] (also see Chapter 7), we have found that the processes resulting in human trajectories are strongly connected with geo-spatial features such as points of interest and their corresponding popularity in the social and semantic web. In order to be able to directly correlate trajectories with such features, we provide the possibility to query and visualize geo-spatial entities from DBpedia [10] via SPARQL [11]. In addition, these entities can be weighted by the view counts of the respective Wikipedia articles[12] (if available), as shown in the example screenshot in Figure 6.2c.

Flickr. Although VizTrails can visualize arbitrary geo-spatial trails, our demonstration example features urban *photo* trails from the Flickr platform. As an additional feature for this dataset, we can also search for particular photo ids or show public photos that have been taken within a bounding box drawn on the map, cf. Figure 6.2a.

[10] `http://dbpedia.org`
[11] `http://www.w3.org/TR/rdf-sparql-query/`
[12] extracted from `http://dumps.wikimedia.org/other`

6.2.4. Related work

Geo-spatial visualizations are widely acknowledged as a part of analysis processes in which we can explore corresponding data, and build hypotheses [320]. In particular, Gahegan et al. [187] and MacEachren et al. [335] argued that visualizations can be tightly integrated into the knowledge discovery process for gaining insights into the underlying mechanics of the observed geo-spatial data. There are many corresponding methods [15, 103, 320, 327]. This includes approaches for general activity patterns [297], traffic data [327], or individual movement [268]. For temporal patterns, such as human navigation behavior, spatio-temporal visualizations are of special interest [140, 298]. However, instead of analyzing overall trajectories, in this thesis, we focus on aggregates of single transitions (between location or places). Thus, tools that focus on spatial interactions or flows can help to visualize the corresponding data. Guo et al. [222] and Chua et al. [118] proposed respective visualization tools. However, the former does not show interactions embedded in a map thus loosing the geo-spatial context, and the latter visualizes all interactions between all entities at the same time (represented as arrows or arcs) which causes cluttered visualizations where entities are dense. With VizTrails we opt for only showing interactions when a specific entity is selected. This looses information on the overall distribution of interactions, but allows for clean visualizations. Additionally, we embed background information into our visualizations, similar to Slingsby et al. [459] who visualized tags from platforms such as Flickr on maps. In contrast to Slingsby et al., we add contextual information queried from DBpedia via SPARQL queries and allow to show images taken by Flickr users for selected areas. While we do not claim to replace any of the existing tools, we believe that combining the aforementioned features, VizTrails is well tailored for our application scenario, i.e., gaining insights into geo-spatial human navigation in order to formulate and explore novel hypotheses about its underlying processes.

6.2.5. Conclusion

We introduced the interactive visualization tool called VizTrails that allows exploring human movement and corresponding trails. To this end, we used a general discretization approach to visualize a number of metrics as well as mutual transitions between areas of interest. VizTrails also allows to set these trails into geo-spatial context using semantic web data via SPARQL queries. Thus, it enables interactive exploration and facilitates deeper insights into spatial trajectory data. In Chapter 7, we use VizTrails to explore and visualize hypotheses about how people move through urban areas based on geo-tagged photos from Flickr. Overall, VizTrails supports the process of conceiving and exploring novel hypotheses as is one of the main challenges addressed by this thesis (cf., Section 1.2.2).

6.3. EveryAware: A platform for collecting, analyzing and visualizing data for mobile participatory sensing campaigns

Human navigation behavior can be observed in many different application scenarios. This also includes users navigating their environment in the context of participatory sensing campaigns. In this section, we introduce the EveryAware platform which is built for collecting, analyzing and visualizing data in this context. Its explorative nature aids the processes of hypothesis conception (cf. Section 1.2.2). In Chapter 8 and Section 11.1, we study data collected using this platform. For the introduction of EveryAware in the following sections, we follow our previous published work (cf. Becker et al. [43]).

6.3.1. Introduction

In the context of the *Internet of Things*, many new applications have been designed for mobile devices enabling people to record environmental as well as personal data by making use of cheap, embedded and (specifically) mobile sensors, such as microphones, cameras, accelerometers, gyroscopes as well as temperature, pressure, air quality, or heart rate sensors. In combination with GPS receivers this tremendously growing number of measurement possibilities enables — among other things — to study highly contextual navigational processes of human behavior as we study in this thesis.

In particular, Cuff et al. suggested that there is a wide range of applications in which people can be engaged in *mobile* sensing expecting a rapidly growing field and a multitude of applications on an urban level [131]. This is in particular true for the field of *citizen science* where volunteers contribute for the benefit of human knowledge and science [224]. Methods and techniques of flexibly acquiring and handling this data play a central role in understanding human behavior and paving the way towards behavioral shifts within large citizen populations.

Thus, the citizen science movement is especially supported by emerging web based platforms making it easy to collectively upload content and share data. The data in this context can be divided into two classes.

1. *Objective data*, which stems mainly from sensors and includes measurements like sound intensity or gas concentration.

2. *Subjective data*, which comprises reactions and perceptions of humans faced with particular environmental conditions.

Problem setting. While traditional *Internet of Things* approaches provide powerful functionalities to support (objective) sensor data, they hardly support the collection and augmentation with subjective information. However, beside collecting and handling measurements, data also needs to be understood and interpreted which is not an easy task. That is, objective data can change its interpretation entirely in different semantic contexts. For example high noise levels at a rock concert are perceived as enjoyable while a leaking water-tap can be considered as *noise pollution*. Therefore, on the way to the

Internet of Everything, the next step is an Internet of Things *and People* [492] not only working on objective data but incorporating people to add impressions, interpretations, and other subjective context.

Approach and benefits. EveryAware aims at providing a platform that links objective sensor data (such as air quality or noise pollution measurements) with subjective information (such as impressions, interpretations, or perceptions). In particular, we propose a highly efficient, generic data collection and processing framework featuring a powerful extension mechanism to allow for semantic data augmentation. With this, we aim to support data alignment and aggregation methods to create representative statistics and visualizations, in order to support advanced knowledge discovery algorithms to mine hidden patterns and relations [247]. Furthermore, EveryAware is built to incorporate geo-spatial information, thus, it allows to collect data about the highly dynamic behavior of users navigating their environment in the context of participatory sensing campaigns. This allows to study human navigation behavior in a novel scenario and a range of unexplored of contextual information. See Chapter 8 Section 11.1 for examples. The EveryAware platform is live[13] and the source code is available[14].

Structure. In the next section (Section 6.3.2), we present the two main parts of the EveryAware system, the *conceptual* and the *implementation* layer. In Section 6.3.3, we present two reference applications of EveryAware, i.e., WideNoise and AirProbe, as well as the currently developed module *Gears*. WideNoise and AirProbe are specialized applications to collect, explore, and analyze noise pollution and air quality, respectively. Gears on the other hand, while building on the same underlying architecture, is a module that aims to provide a *generic* framework for sensor data collection and visualization. After introducing these modules, we then critically discuss the current features of EveryAware (Section 6.3.4) and review existing data collection services in the context of the Internet of Things in Section 6.3.5. Finally we summarize our work including possible future directions in Section 6.3.6. Also, see Chapter 8 and Section 11.1 for studies analyzing human navigation behavior based on data collected using the EveryAware platform.

6.3.2. Architecture

On the conceptual level, the EveryAware platform has been designed to enable users to collect, visualize, and share personal sensor measurements (mainly focusing on environmental factors) and at the same time augment the collected data with arbitrary information explicitly supporting subjective context.

On the technical level, the data processing engine allows for the application of dedicated data mining and knowledge discovery algorithms in order to fully exploit the synergies of a central data storage and the wide variety of objective and subjective information. Our platform was co-developed with the Ubicon framework [20] and extends it to provide the functionality needed for our architecture.

In the following, we introduce both, the conceptual as well as the technical layer.

[13] `http://cs.everyaware.eu`, accessed: August 2017
[14] `http://dmir.org/everyaware-opensource`, accessed: August 2017

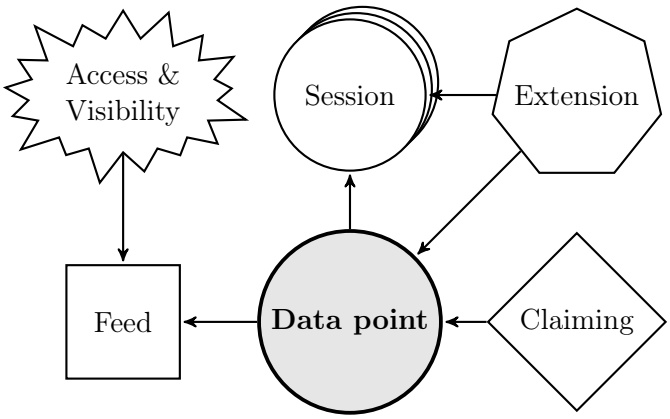


Figure 6.3.: Conceptual design of the EveryAware system.

6.3.2.1. Conceptual layer

The conceptual layer defines the basic entities and features of the EveryAware system revolving around the notion of *data points* (see Figure 6.3).

Data Points. Since EveryAware aims to support arbitrary applications and data types, the concept of data points are held as general as possible. They can consist of air quality data, noise pollution measurements, heartrate readings, or conceptually even images, videos, or other binary data. Processing and interpreting the actual content of data points (including subjective data) is handled later, i.e., by a data processor engine (Section 6.3.2.2).

However, EveryAware is supposed to be able to handle arbitrary data (e.g., supporting indexing, querying, basic analysis, etc.), even if no data processor module exists which is able to interpret a particular type of data point. To achieve this, each data point is augmented with a fixed set of *description attributes* in addition to the actual data. The description attributes are divided into three categories:

- (1) *Meta attributes* are attributes which allow to keep track of data independent information like received time, recording time, a device ID, or session IDs, etc.
- (2) *Geo attributes* emphasize the geo-spatial nature of personal sensor data (especially environmental data) and make it possible to record the location of the sample being taken including longitude and latitude as well as accuracy attributes, the location provider or other relevant information.
- (3) *Content attributes* describe the content and its format. They help the system to further process the data. These attributes include the data type (e.g., air, noise, image), the format (e.g., JSON, XML, PNG), version numbers, etc.

Feeds. A major aspect of collecting massive amounts of data points is how to manage them and how to control their accessibility. To this end we introduce the common notion of *feeds* which are used to organized data points. A data point is always part of one or

more feeds. Users can contribute to existing feeds or create their own feeds. While useful for organizing data points, feeds also allow to attach data points to real world entities such as major events like music festivals, places like the Eiffel Tower, or portable things like a smartphone. However, the most important feature of feeds is managing access restrictions. That is, feeds can be *open* or *closed* concerning read and write access, where *write access* refers to the possibility of adding new data points to a feed. Open feeds are accessible by everyone including anonymous users. Closed feeds are only accessible by a limited set of users (*members*). The access restrictions allow users to create feeds and share them with friends or other interested users without making their data publicly available.[15]

Sessions. Data points are often semantically related by their temporal context. Such contexts can be defined by the source of the data points, i.e., from turning on the source to turning it of, or — representing a more semantic context — they are explicitly defined by the user and represent something like "my way to work" or "a stroll in the park". To represent such temporal relations, we introduce *sessions* which basically are collections of data points from the same source limited to a fixed timespan.

Extensions. In order to make the data representation flexible and for inherently supporting the augmentation and annotation of the collected data (e.g., with additional information, subjective data, or semantic context), we introduce the concept of *extensions*. That is, we allow data points as well as sessions to be extended by other data points. Because of the generic definition of data points the information used for annotation is again very flexible.

One major application of the extension mechanism is tagging (i.e., adding specific words further describing the tagged entity). In particular, sessions and data points can be tagged by extending them (by referring to the respective data point or session ids) with data points containing a set of tags. Because tagging data can have different formats, a dedicated data processor module is required (see Section 6.3.2.2).

Using the extension scheme, it is also possible to update data points as well as sessions after they have been sent without losing the original data. Since, generally, no raw data is deleted, this also allows to always access the version history of a data point.

Claiming. Often, a significant hindrance for users to participate in data collection campaigns is the requirement to create an account and sign up on the specific platform. EveryAware aims to tackle this issue by introducing the notion of *claiming*. In particular, claiming allows anonymous contribution to the EveryAware system while giving the user the possibility to claim data points as soon as she decides to register an account. This makes the contribution of data to the EveryAware convenient and provides some level of anonymity since no previous registration process is required. Currently this functionality is implemented by using device ids, i.e., the ids of the device sending the data point. Knowing such a device id, the user can claim the corresponding data points. This still

[15] Note that Becker et al. [43] also introduced several levels of visibility for each data point in a feed (i.e., "details", "statistics", "anonymous", and "none"). However this concept was mostly dropped since it has proven to be impractical to use, i.e., for developers it was tedious to implement and for users the concepts were hard to grasp (also see Section 6.3.4).

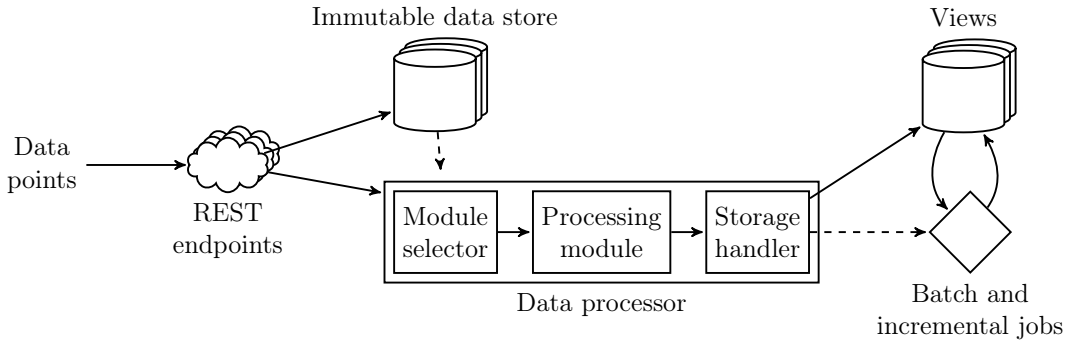


Figure 6.4.: Technical architecture of the EveryAware system.

does not provide a high level of anonymity. However, there are several alternatives to device ids discussed in Section 6.3.4.

6.3.2.2. Implementation layer

The goal of the EveryAware platform is to provide a reliable platform for collecting, analyzing and visualizing a wide range of data types, specifically supporting environmental data, health sensors, subjective data and more. For this, the EveryAware builds on a flexible data processing framework which we explain in this section.

Layout. The overall data flow of the EveryAware processing framework is shown in Figure 6.4. Data points come in through several (possibly distributed) REST endpoints and are directly stored in an *immutable data store*. At the same time they are forwarded to the *data processor* engine which interprets the raw data from the data points and writes them into several basic *views* which can then be queried by users. To support advanced visualizations, the basic views are augmented through a set of *batch and incremental jobs*.

Immutable data store. The highest priority of the EveryAware platform is to reliably store the received data points. To ensure this, computational overhead such as syntactic checks is kept at a minimum. The received data is directly written to an (possibly distributed) *immutable data store*. Processing the data is externalized and is handled by the data processor (covered later). This gives us the following advantages:

- *High performance and availability:* Any computation in the endpoint can introduce sources of error and impede the performance of the data reception process. This is especially a problem for applications with high transmission frequencies and complex data structures.

- *Flexibility:* We can accept literally any data, since the endpoint does not restrict the type of data sent via the REST endpoint. Different data types are handled by a dynamically manageable set of processing modules in the data processor.

- *Robustness:* Storing the data in its raw form enables us to recreate the processed content at any point in time.

Data processor. The data processor directly receives data from the REST endpoint (or can poll it from the immutable storage if required). The data is then parsed, interpreted and augmented using a chain of processing modules. The resulting information is stored in different *views* which are query-able from data access endpoints. The pipeline from receiving the data, over processing it, to storing the results in a view, is managed by three subsequent components: the *module selector*, a set of *processing modules*, and a set of *storage handlers*.

- The *module selector* selects a *processing module* from a priority-chain. The matching module is selected based on its data type defined by the data points descriptions attributes or by deriving the data type from the raw content.
- The second component is the selected *processing module*. It extracts the actual data from the raw content (e.g., a JSON file) and possibly augments the data with additional information, calculates statistics, or handles missing information. As covered later, there are dedicated modules for parsing noise pollution data, air quality measurements, or subjective information (e.g., tags).
- Then, the data is passed to a *storage handler* which stores the results in dedicated *views*. The storage handler may also pass the processed data directly to incremental jobs for augmenting the views (as covered below). The data processor also tracks which data points have been processed.

The modular approach allows to simply exchange processing modules (e.g., when an updated version needs to be deployed). To ensure the consistency of the data, the processing state for affected input data is reset. The data processor engine then processes the marked data and replaces the results in corresponding view tables.

Furthermore, like the immutable input store, the data processor is designed to work in distributed environments. This allows for replication of the REST endpoints enabling effective processing of large amounts of incoming data.

This architecture has several advantages: The priority-chain-approach in the *module selector* allows for flexibility in extending the data processor engine with additional processing modules on demand. The modularized approach in general makes it easy to deploy updates without risking to break the existing data, where keeping track of the *processing state* of data points is the key to flexible module extensions and guarantees robustness against processing failures. And finally, supporting distributed processing ensures a scalable architecture than can cope with increasing amounts of incoming data.

Batch and incremental jobs. The data processor mainly processes and interprets the incoming data. However, while it can handle certain data augmentation steps, it is not designed for large batch processes which are often required for rich visualizations, complex analysis tasks, or machine learning approaches. Because of this, EveryAware also employs a set of *batch and incremental jobs* which are responsible for augmenting the basic views generated by the data processor. These jobs can get data from the views as a whole, but they can also be served data point by data point by the data processor. Thus, the general structure of such jobs usually follows the basic ideas of the Lambda Architecture [343]

proposed by Marz. That is, they consist of a batch layer, a speed layer and a serving layer. As an example we consider our cluster preparation module which pre-calculates clusters, e.g., to be visualized on a map (see for example Figure 6.6b). This is necessary because, for example for our AirProbe application covered in Section 6.3.3, the amount of points is too large to cluster on the fly. The batch component of the cluster job reads data from the basic views provided by the data processor, clusters them and prepares them for retrieval. The speed component of the cluster job receives updates directly from the data processor or continuously polls the basic views and incrementally integrates the new points into the existing clusters. These combined clusters are then served to the user via dedicated API endpoints. Integrating such jobs into the general EveryAware framework makes it possible to provide a very flexible and modular way to add new analysis methods or visualizations.

6.3.3. Applications

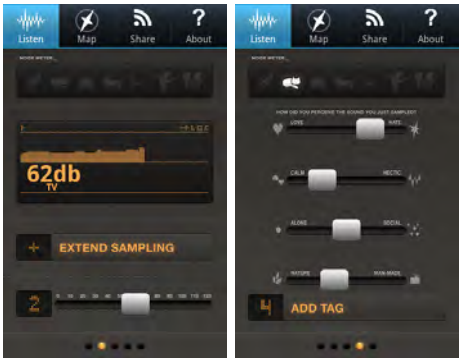
The EveryAware platform provides a powerful framework for working with arbitrary data, including participatory sensing or quantified self applications. Among other things, this also enables novel studies on human navigation behavior in the participatory sensing domain. In this section we give two examples of applications from the participatory sensing domain for measuring noise pollution and assessing air quality, i,e., WideNoise and AirProbe. In addition we give an outlook to the currently developed module *Gears* which — in contrast to the specialized applications WideNoise and AirProbe — aims at a more generic data representation for collecting, analyzing and visualizing data.

6.3.3.1. WideNoise and AirProbe: Noise pollution and air quality

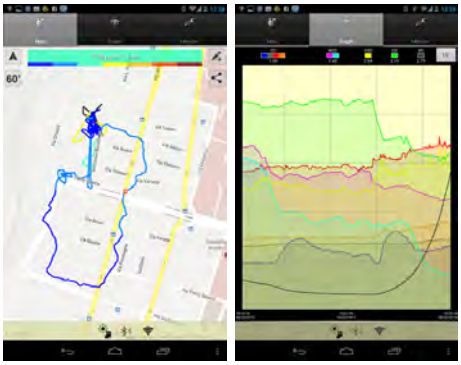
Two major applications which have been implemented using our platform are WideNoise and AirProbe. Both have been developed as part of the EU research project EveryAware[16] and have been mentioned before by Atzmueller et al. [20]. WideNoise is an application for measuring noise pollution. It was originally developed by WideTag[17] and further enhanced during the EveryAware project. AirProbe, developed by CSP[18] as part of the EveryAware project, monitors air quality. Both applications have a smartphone interface (see Figure 6.5) as well as a web interface (see Figures 6.6b and 6.7). The smartphone gathers data and transmits it to our server where it is augmented and aggregated by the data processor to be visualized on the web frontend. WideNoise measures noise levels using the built-in microphone of the smartphone (cf., Figure 6.5a), while AirProbe records the measurements (such as NO_2, CO, O_3, VOC[19], temperature and humidity) using an external sensorbox [162] as shown in Figure 6.6a. The following paragraphs will focus on WideNoise and AirProbe and their common as well as distinguishing features.

Subjective data. Noise pollution and air quality are both interpreted in highly subjective

[16] `http://everyaware.eu`, accessed December 2017
[17] `http://widetag.com/`, accessed March 2013 (not accessible any more)
[18] `http://csp.it/`, accessed: December 2017
[19] Volatile organic compounds



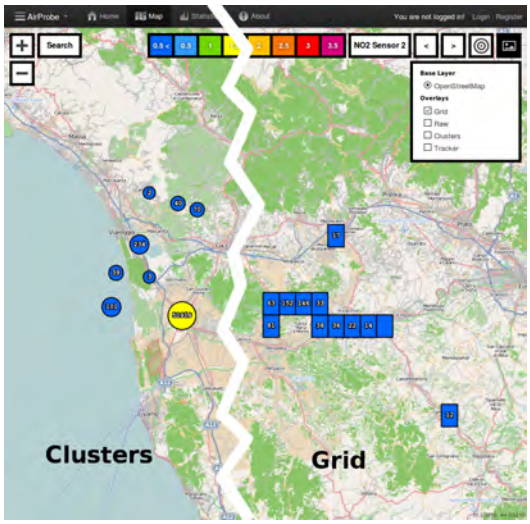
(a) *WideNoise*: the recording view as well as the perception dialog.



(b) *AirProbe*: the current session overview and the sensor value view.

Figure 6.5.: Screenshots of the WideNoise and AirProbe Android applications. For WideNoise (a) the first screenshot shows the dialog used to measure noise levels (in dB) over a short period of time. There the user can also guess the noise level before being presented with the result in order to learn to judge noise levels. The second screenshot shows a dialog for adding perceptions to the measured noise thus enabling the collection of subjective data. For AirProbe (b) the first screenshot shows the current session the user is recording where the buttons at the top-right corner allow the user to add custom tags for adding information about the context of the measurements. The second screenshot shows the most recent measurements from different sensors being recorded.

(a) The AirProbe sensor box



(b) A screenshot of the *map page* of AirProbe. The left side shows the cluster view, the right side shows the grid view.

Figure 6.6.: AirProbe sensorbox and map view of the EveryAware web application. The sensor box is used for collecting air quality data and a map visualizes various statistics.

(a) A session of recorded air quality measurements. Here a student walks from the computer science building to the cafeteria at the University of Würzburg.

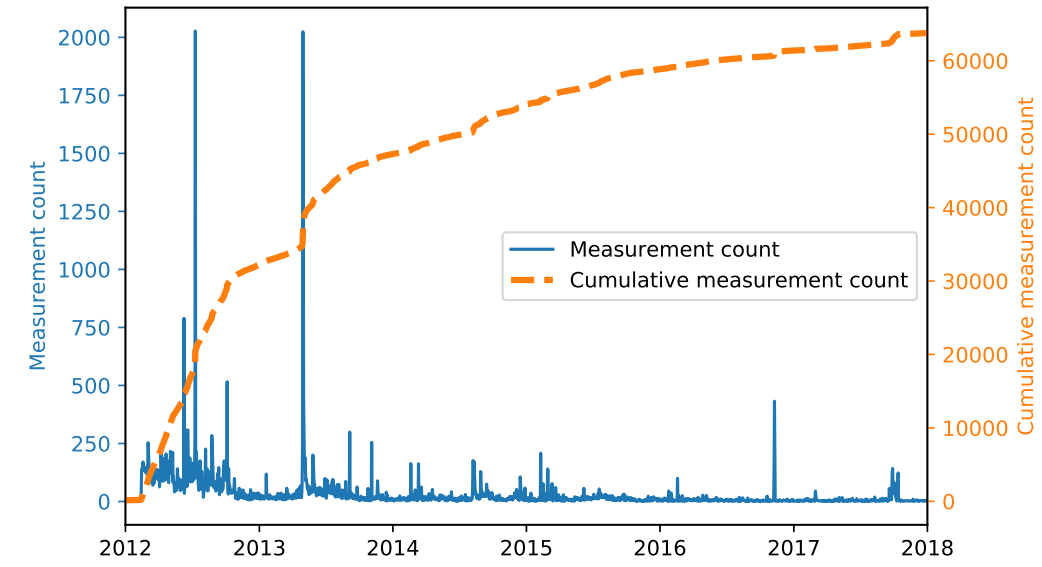
(b) A heatmap visualizing the number of measurements in Kassel. The data was recorded by a group of users during the participatory sensing campaign APIC, cf., Sirbu et al. [457].

Figure 6.7.: Mobility related visualizations on the EveryAware web application. The session view (a) allows for exploring individual traces, and the map view (b) shows aggregated spatial coverage information.

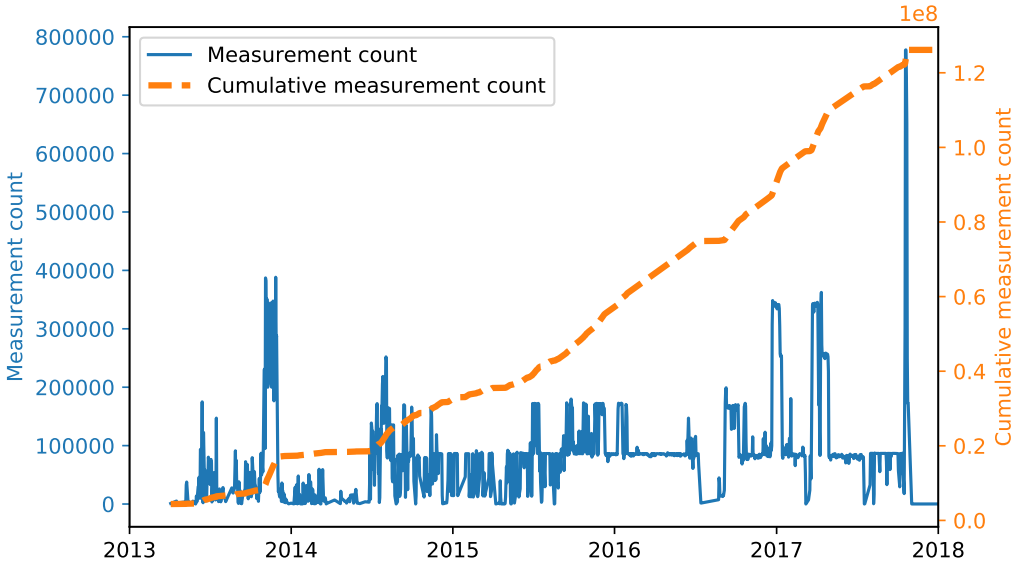
contexts. For example, high noise levels are not perceived as pollution when users are attending a rock festival. Similarly users may not perceive high ozone levels as bad when they are sunbathing. Thus, both, WideNoise and AirProbe support extending the objective sensor readings with subjective data as explicitly supported by the EveryAware system. In both cases, this subjective data is expressed as tags. Also, WideNoise allows to add noise estimates as well as user perceptions to the noise samples. The corresponding functionality is shown in Figure 6.5.

Visualization and mobility. After the data is received by the EveryAware system, the data processor parses and aggregates the data by applying several dedicated processing modules. The results are statistics and corresponding visualizations on a global scale as well as on the user level. One major visualization is the world map as shown in Figure 6.6b. It displays, for example, a clustered view of the recorded data (providing corresponding detail information on demand [447]) as well as a tag cloud characterizing the summarized data by its semantic context. Two more important views allowing to visualize human mobility in the context of participatory sensing campaigns can be seen in Figure 6.7. It shows a view for reviewing existing sessions a user has recorded in Figure 6.7a, and a view which shows a map where the spatial coverage is visualized as a heatmap Figure 6.7b collected by a group of participants during the participatory sensing challenge APIC [457].

Basic statistics. WideNoise has been used in a variety of campaigns. This includes for example a case study in the Heathrow airport area to monitor noise pollution caused by



(a) WideNoise



(b) AirProbe

Figure 6.8.: Number of measurements over time for the EveryAware applications WideNoise and AirProbe. For WideNoise, several major peaks are visible. These were due to specific case studies (e.g., fist large peak), or news paper articles (e.g., second large peak). For AirProbe, the first large peak corresponds data from the APIC challenge (Chapter 8). The other peaks stem from various internal case studies.

air traffic. In Figure 6.8, we show data from the years 2012 and 2018, where we collected more than 64,000 noise samples recorded by over 18,500 devices from all over the world. The AirProbe system was built for measuring the air pollution in a major case study across several cities. For more details, see Chapter 8. There we have already collected close to 120 million air quality samples from only about 82 devices. The large difference in numbers is due to the discrete nature of WideNoise, while AirProbe applies a continuous sampling scheme. Also see Figures 6.8a and 6.8b for sample growth rates. The peaks are mostly due to events, media announcements, or dedicated case studies. The large number of samples makes the AirProbe application our main benchmark application. In particular for AirProbe we had to carefully optimize our system in order to provide a seamless user experience: for example when new data is sent to the server, updates should optimally be visible in less than eight seconds [279].

6.3.3.2. Gears: Towards processing generic data

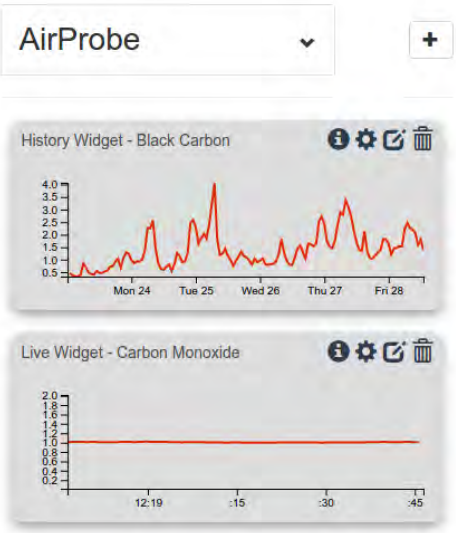
WideNoise and AirProbe have demonstrated the flexibility of EveryAware to handle diverse application scenarios by providing a general framework for collecting, analyzing and visualizing environmental data in combination with subjective impressions and perceptions. However, both applications work with very specific types of data, i.e., noise pollution and air quality measurements. Thus, the corresponding modules are mostly optimized for the respective data type. Consequently many integrated analysis tools and visualizations are not directly portable between the applications. To solve this the EveryAware platform currently moves towards implementing a generic data type which features a flexible structure in order to support a wide range of different applications. For this we implement a module similar to WideNoise and AirProbe, we call *Gears*. That is a dedicated data processing module takes care of parsing and interpreting the data. In addition, a corresponding visualization framework is developed that — due to the underlying generic data format — supports a broad spectrum of data types. In the following we briefly introduce the data format and its technical implementation as well as the visualization component. Both are visualized by Figure 6.9. Note that the *Gears* system is still under development. However, a beta version is already available.[20]

The generic data format. The goal of a generic data format is to support a wide variety of different application scenarios in the domain of participatory sensing. To achieve this, we introduce a format based on the notions of *sources*, *channels* and *channel components* which emphasizes our focus on sensing campaigns. The most basic concept are *sources* which represent sources of data such as air quality boxes, weather stations, a fitness band, and more. Each source is made up of *channels* which roughly represent the individual sensors of the corresponding source, such as an accelerometer, a CO_2 module, or a heart rate sensor. Now, each channel can have several *(channel) components*. For example, an accelerometer always measures the acceleration in three dimensions: x, y, and z. Each channel component can either hold numbers or text. These data points are processed by a dedicated data processor module (cf. Section 6.3.2.2) and saved in a view

[20] `http://cs.everyaware.eu/event/gears`

```
{
    timestamp: 12312335354345,
    channels: {
        accelerometer: {
            x: 0.5533,
            y: 0.12223,
            z: 0.001
        },
        co2: {
            value: 0.65
        },
        tags: {
            value: "smog,traffic"
        }
    }
}
```

(a) The *Gears* data format



(b) The *Gears* dashboard

Figure 6.9.: Example of the Gears data format and web dashboard. The listing on the left (a) shows an example data point (e.g., from some mobile sensor box) using the *Gears* data format in its JSON representation. The screenshot on the right (b) depicts the *Gears* dashboard for visualizing generic data. Here, two widgets have been configured showing black carbon and carbon monoxide readings from corresponding sensor sources. One is a live widget (updating as new data is coming in) and one is a history widget (used for exploring existing data).

table analogously to the data points from WideNoise and AirProbe. Thus, all the features mentioned in Section 6.3.2.1 are applicable to generic data points as well. This includes, e.g., meta, geo, and content attributes, as well as the concepts of sessions and extensions.

An example of a single data point from an air quality sensor box can be seen in Figure 6.9a. The example also defines a channel for tags demonstrating how the same format can be used to also add subjective or contextual data to the recorded sensor values. There also is wide variety of other application domains for this data format. For example, we have implemented a frontend for defining and distributing questionnaires and surveys. Also, the dashboard visualization framework, seen in Figure 6.9b, uses the *Gears* data format in combination with feeds (as introduced above) to manage dashboards and widgets for visualization.

Widget based visualizations. The generic data format *Gears* also allows to build visualizations which can be applied to a wide variety of different data types. In particular we are working on libraries allowing to efficiently implement widgets which can be used on arbitrary data within the *Gears* environment. To fully exploit theses widgets we further provide a fully customizable dashboard system as visualized in Figure 6.9b. It shows two widgets, a history widget for exploring past data, and a live widget for monitoring currently incoming data. Additionally, the widgets can also be used in a standalone mode, e.g., to build static statistics pages. This system allows to design highly reusable

components and a flexible user experience. We aim to enable users to easily run campaigns like WideNoise or the AirProbe challenge without having to set up their own systems.

6.3.4. Discussion

The EveryAware system is a flexible platform for handling many different types of data and aims at explicitly supporting subjective information, such as perceptions or personal impression, e.g., in the form of tags. It was originally introduced in 2013 by Becker et al. [43] and has been successively improved since then. Some concepts have proven impractical and other components have been further developed or replaced by more efficient modules. Thus overall, EveryAware is a quickly changing and evolving system driven by a relatively small number of people. Here, we discuss some issues and future directions which can help to further improve EveryAware.

Access and visibilities. Our original proposal of EveryAware [43] specified visibility levels for each data point in each feed. These visibility levels were supposed to allow users to set the granularity of the data they share in order to protect their privacy. For example, it was possible to restrict the "visibility" of data points to aggregated views, e.g., as part of a mean value over all users. This way the user was able to obfuscate more detailed information such her daily route from home to work or other data from which personal information can be derived. While the idea is still valid, there are two technical issues with this approach: First, each data point had to be handled separately, making the index structures to query data points elaborate and implementing analysis or visualization algorithms tedious. Secondly, enforcing the visibility restrictions is not an easy problem, e.g., because every API endpoint needs to be carefully designed and for every new algorithm these restrictions have to be ensured. Nevertheless, Mineraud et al. [359] judged our four-level visibility design to be crucial for providing the "necessary flexibility to maximize the re-usability of the data by remote third-party services". Thus, while dropping the point-wise support of these levels and currently moving towards a simpler, strictly feed-based access management, we aim to reintroduce our four-level visibility approach based on feeds. In combination with a general data format like *Gears* as introduced in Section 6.3.3.2 and dynamic feed to data point relations, it may be possible to enforce these visibility levels.

Claiming. The claiming procedure is a concept which can significantly reduce the threshold for users to share data through the EveryAware platform by postponing the registration procedure. However, the current implementation has several issues. First we still lack an efficient method for incorporating the information about claimed data points into our already processed views. Especially for incrementally calculated aggregated statistics the claiming procedure may trigger long running batch jobs. Second, using device ids to authorize claiming (as is currently the case) is insecure because it is not a protected resource and users can still be identified. Generating a random key instead of using the device id can help to reduce the risk of someone claiming the data of someone else. However, in both cases, the individual user is still traceable because the random id stays the same. To solve this, the random key can be encrypted together with a random

seed. While in principle this allows users to anonymously contribute data to EveryAware, processing a claiming request will require decrypting the random keys of all unclaimed data points. Future research may yield solutions to some of the mentioned problems.

Extensions and sessions. Extensions have been proven useful in our scenario, e.g., in order to add tags after the actual data point has already been sent. However, the current approach is limited to extending single data points or sessions. While we will keep basic support for this feature in order to allow for arbitrary data to be extended, we are currently exploring an alternative approach using the generic data format *Gears* which allows to track tagging and other annotation data via dedicated channels or sources. Using dedicated channels our sources would replace some of the meta headers of the conceptual layer (cf. Section 6.3.2.1) moving them from the headers to the actual data. While this makes the overall procedure more flexible, it possibly bloats the data with meta information and will require specialized data processor modules.

Data processor. The data processor was originally designed to poll new data points from the immutable data store. We are currently replacing it by an actor-based processor based on a publish-subscribe architecture allowing for a more dynamic and scalable implementation. Also, this enables us to serve data back to the user in real-time.

Work in progress and future work. Generally, the EveryAware platform keeps evolving as we adapt to more and more application scenarios. In this process the underlying concepts sometimes need to change rapidly. Nevertheless, we believe that with the *Gears* module in combination with our flexible data model based on feeds, we will be able to support a wide range of applications without having to implement specialized solutions.

6.3.5. Related work

The Internet of Things is defined by the connectedness of devices. Thus, data exchange must be as simple as possible. For this end a large number of data storage and distribution platforms have emerged Mineraud et al. [359]. Some examples are Xively[21] (formerly Cosm and before that Pachube), Ubidots[22], Exosite[23], and ThingSpeak[24].

These platforms provide powerful frameworks to work with devices, sensors, and timeseries data. Most of them support flexible paradigms for defining devices with different data types, channels, or sensors and provide dashboard systems to visualize the recorded data. However, even though the data structures of devices can be customized, the mentioned IoT platforms all impose a specific model for the data points of connected devices. The EveryAware system takes a different approach where data points are kept on a more abstract level allowing for any content to be handled by a flexible data processor engine (cf. Section 6.3.2.1). Theoretically, the model of any platform mentioned above can be emulated using the EveryAware system if an appropriate data processor module

[21] `https://xively.com`, accessed: August 2017
[22] `http://ubidots.com`, accessed: August 2017
[23] `https://exosite.com`, accessed: August 2017
[24] `https://thingspeak.com/`, accessed: August 2017

is implemented. For an example, see the Gears module as described in Section 6.3.3.2. Conceptually, this paradigm even allows to collect images or video footage.

Furthermore, the Internet of Things has focused on *things* and the corresponding data but not on how people interact with these things [492] or what the data actually means to them. This mindset is mirrored by the providers mentioned above, thus, hardly any functionality is present to share information about the collected data or to add subjective impressions. With EveryAware we try to work towards a future *Internet of Things and People* by allowing data points to be extended with any kind of data. This helps to put data into a meaningful context including tags and other subjective information people may think of. To the best of our knowledge, this approach is novel in context of the Internet of Things. Our platform also allows the application of various data mining and knowledge discovery tools. This aspect addresses challenges as mentioned for example by Hotho et al. [247] and further distinguishes our platform from the mentioned providers.

6.3.6. Conclusion

The EveryAware platform provides a robust, efficient, and flexible system for collecting, analyzing, and visualizing sensor data specifically supporting subjective data such as user impressions or perceptions. We introduced two specialized applications, i.e., WideNoise for noise measurements and AirProbe for air quality sensing, as well as the currently developed module *Gears* for generic data. This demonstrates that EveryAware can support a variety of different application domains, e.g., in the context of participatory sensing. Specifically, the mobile aspect of this kind of data opens up new possibilities to study human mobility (e.g., Figure 6.7) in combination with a rich ecosystem of contextual information. Furthermore the explorative nature of EveryAware aids the processes of hypothesis conception (cf. Section 1.2.2). For a case study analyzing data from EveryAware, also see Chapter 8 and Section 11.1.

As future work, we mainly aim to push the generic module *Gears* and plan to add new and more advanced analysis algorithms incorporating methodology from data mining and machine learning. Also there are still some challenges to be addressed including, for example, in-depth evaluation of distribution techniques, data storage and processing engines, enabling users to customize or add certain data processor modules, or developing a trigger system based on user-defined events. While this will still involve many hours of design and implementation effort, we believe that a well designed open source system in this area deployed and maintained by the academic community can greatly improve the current situation of ad-hoc solutions, commercial closed-source systems, and privacy critical collection campaigns.

Part III.

Case studies

7. Photowalking urban environments

In this chapter, we investigate human navigation behavior in urban areas based on photos from the social photo-sharing platform Flickr. In the process we apply most of our proposed methods from Part II. That is, we construct a pipeline consisting of HypTrails (Section 3.3.2) for comparing hypotheses about overall navigation processes, SubTrails (Chapter 5) for discovering subgroups with exceptional transition behavior, and MixedTrails (Chapter 4) for comparing hypotheses featuring heterogeneous processes human navigation. Furthermore, we use SparkTrails (Section 6.1) for speeding up calculations and VizTrails (Section 6.2) for visualizations. Furthermore, we showcase how to prepare data and formulate hypotheses in a setting with spatially continuous observations. Thus, overall, this case study illustrates how our methodological work provides a full-featured toolbox for exploring human navigation behavior. The work presented in the following is mainly based on our previously published work on phototrails [44] and contains results from Becker et al. [41] and Lemmerich et al. [312].

7.1. Introduction

Also already mentioned in Chapter 1, understanding the way people navigate urban areas represents an important problem that has implications for a range of societal challenges such as city planning and evolution, public transportation or crime. Recent research in computational social science has studied human movement trajectories in cities through a variety of data sources including mobile phone data [214, 463], GPS tracking [527], Wifi tracking [418], location-based social media platforms [116], online photo sharing sites [135, 205, 206] and others.[1] Such studies have provided a number of insights into human movement trajectories. For example, past work has indicated that human mobility exhibits regularities [214, 463] and spatio-temporal patterns [116]. Research has also shown that we can successfully leverage these patterns for certain tasks such as constructing high quality travel itineraries [135].[2] Yet, little is known about how the corresponding trails materialize, i.e., what factors play a role when people move through urban spaces. Thus, in the following, we extend the stream of research on human movement by studying the underlying navigational processes of *how* urban photo trails are produced. A better understanding of this process is relevant for a series of practical problems, such as local recommendations of picturesque locations, studying touristic movement patterns, or movement of people in urban environments in general.

[1] For more information on data sources in the context of human mobility research, also see Section 2.1.2.
[2] For more information on studies and results about geo-spatial navigation, also see Section 2.1.

(a) Berlin heatmap **(b)** Transition probabilities **(c)** Proximity hypothesis

Figure 7.1.: Analyzing human navigation behavior in urban areas based on discretized sequences of photos from Flickr. In (a), we visualize a cell-based grid layout of Berlin with photo frequencies visualized in a heatmap format, as derived from our data at interest. (b) depicts an example of transition probabilities between cells. In particular, it depicts the cell where the "German Bundestag" is located, and visualizes the transition probabilities to subsequent cells, i.e. cells where people take photos after they photographed the Bundestag. For instance, with a probability of 0.07, people take a picture at the "Brandenburg Gate" after they have taken one at the Bundestag. As we are interested in gaining insights into the processes producing these trails, we formulate hypotheses based on belief in transitions between cells (represented as parameters of a Markov chain). In (c), we depict the transition probabilities from the Brandenburg Gate to other cells based on an exemplary proximity hypothesis which represents the belief that people successively take their next photo close to their last one.

Problem setting. In the following,tial photo data from four cities (Berlin, London, Los Angeles and New York) retrieved from the social photo-sharing platform Flickr[3]. In particular, we analyze *urban photo trails* defined as a sequence of spatial positions in a city over a period of time as, e.g., obtained from the geo-temporal meta data of photos.

On this data, we first assess the plausibility of different potential explanations (hypotheses, cf. Section 3.3.2.1) for the overall movement patterns we observe and compare the corresponding hypotheses across cities. For example, we compare a *proximity hypothesis* — which represents a belief that people frequently take subsequent photos in geographically close regions of a city — with a *points of interest (POI) hypothesis* that represents the belief that humans take subsequent photos of POIs.

Secondly, we investigate whether there are several sub-processes responsible for the high-level patterns we observe. In particular, besides studying the overall movement behavior, i.e., assuming a homogeneous behavioral process, we also investigate whether we can find subsets of our data (such as trajectories from tourists or locals) which exhibit exceptional behavioral patterns. Using these patterns, we then aim to find heterogeneous explanations for the observed urban photo trails, taking several sub-processes into account.

Approach. To tackle these challenges, we resort to several approaches reviewed or proposed in this thesis including *HypTrails* [453] (see Section 3.3.2), for comparing hypotheses about overall navigation processes, SubTrails (Chapter 5), for discovering subgroups with

[3] `https://flickr.com`, accessed: December 2017

exceptional transition behavior, and MixedTrails (Chapter 4), for comparing hypotheses featuring heterogeneous processes of transition behavior. To apply these methods, we map each photo of a trail to a discrete area within a specific city (i.e., using grid cells as depicted in Figure 7.1a or census tracts). For all our approaches we investigate transition probabilities between these areas as shown in Figure 7.1b. For formulating hypotheses (e.g., Figure 7.1c), we utilize general information (like geo-spatial distance), data extracted from the social semantic web (e.g., points of interests in cities from Wikipedia, DBpedia, and YAGO), and corresponding usage statistics (view counts).

Contribution and findings. The main contribution of this case study is an in-depth analysis of human navigation behavior in the context of urban photowalking on a homogeneous as well as a heterogeneous level.

This encompasses a systematic evaluation of homogeneous hypotheses for explaining how urban photo trails are produced in four different cities. We find interesting commonalities, in particular, that the partial ordering of evidence for different hypotheses is quite stable across the cities we have investigated. Furthermore, information extracted from social media — in the form of concepts and usage statistics from Wikipedia — allows for finding advanced explanations for human movement trajectories. Most prominently, our results suggest that humans seem to prefer to consecutively take photos at proximate POIs that are also popular on Wikipedia. In addition, we also observe differences between cities: For example, proximity is less relevant for Los Angeles, which is a plausible finding given the unique topology of the city among the studied datasets.

Finally, we also study the heterogeneous nature of the overall navigation process by first discovering subsets of our data with exceptional navigation characteristics and secondly combining them into heterogeneous hypotheses accounting for these characteristics.

The findings of our work can enable photo sharing websites to offer localized recommendations of picturesque photo spots according to actual tourist trajectories, city planners to explore human movement patterns of its inhabitants in general, or tourist organizations to facilitate and optimize tours.

Structure. Starting with Section 7.2, we describe the Flickr data we use in our experiments. In Section 7.3, we present our hypotheses. With regard to experiments, Section 7.4.1 summarizes our results on the overall data, Section 7.4.2 explores transition subsets with exceptional navigation behavior, and Section 7.4.3 compares heterogeneous hypotheses which assume that the overall navigation behavior is a combination of tourists and locals which exhibit specific behavioral characteristics. Finally, we discuss our work in Section 7.5, give an overview of related work in Section 7.6, and conclude in Section 7.7.

7.2. Data

In this section, we frist describe the process of collecting photos from the social photo-sharing platform Flickr[4] which we use to analyze urban human navigation behavior in this case study. Then, we describe the transformation procedure of these photos into the

[4] `https://flickr.com`, accessed: December 2017

Table 7.1.: Data collection parameters for the Flickr data. Bounding boxes and center coordinates used for collecting our data and creating hypotheses.

	Berlin	London	Los Angeles	New York
min lon.	13.088400	-0.5103	-118.6682	-74.2589
min lat.	52.338120	51.2868	33.7037	40.4774
max lon.	13.761340	0.3340	-118.1552	-73.7004
max lat.	52.675499	51.6923	34.3368	40.9176
center lon.	13.383333	-0.1280	-118.2450	74.0071
center lat.	52.516667	51.5077	34.0535	40.7146

required representation of trails and state transitions for which we employ two different discretization methods. Finally, we give a quick overview on how we derive popularity scores for the POIs in each city based on Wikipedia.

7.2.1. Data collection

Our datasets[5] contain meta data — i.e., user, temporal, and geo-spatial (latitude and longitude) data — about images uploaded to the Flickr platform. In particular, we focus on pictures taken in the cities of Berlin, London, Los Angeles, and New York between January 2010 and December 2014. For each city, we define a bounding box, see Table 7.1. We acquired corresponding data by crawling Flickr's public API. Since our analysis requires an exact position, we remove pictures with less than street-level accuracy (level 16 on the Flickr scale[6]).

For our analyses, we interpret the sequence of all photos of a single user as a *photo trail* ordered by the time each photo was taken, regardless of the time difference between the photos (also see Section 7.5 for a discussion on the influece of large time differences).

7.2.2. Discretization

The methodology covered in this thesis (Part II) builds upon discrete state spaces to analyze human navigation behavior (cf. Section 3.1). Thus, we need to discretize the continuous state space defined by the geo-spatial context of the collected photos. Depending on the discretization mechanism (yielding a discrete set of states), each photo in a photo trail is mapped to such a state according to its geo-reference. In this case study, we employ two different discretization approaches: a grid-based approach, used for fine-grained *homogeneous* movement analysis, and a more semantic discretization based on census tracts for studies on *heterogeneous* aspects of human navigation. Both methods are introduced in the following two sections.

[5]Dataset access can be requested via e-mail: becker@informatik.uni-wuerzburg.de.

[6]`https://www.flickr.com/services/api/flickr.places.findByLatLon.html`, accessed: December 2017

Table 7.2.: Basic dataset statistics of the Flickr datasets. For the four cities we investigate, this table shows details about the crawled photos from Flickr as well as the points of interests queried from DPpedia. In addition to the covered time spans and photo counts, we also list statistics about derived information such as the number of cells or trails.

	Berlin	London	Los Angeles	New York
years	2010-11	2010-14	2010-14	2010-14
photos	60,978	794,535	300,373	714,549
cells	43,052	66,444	84,014	58,065
trails	4,364	35,101	15,357	31,246
covered cells	6,343	23,694	25,834	15,232
avg. trail length	13.97	22.64	19.56	22.87
POIs	1,085	7,228	1,462	6,002
avg. view count	1,240	1,272	3,654	1,511

7.2.2.1. Grid cells

In our first (and more extensive) set of experiments on *homogeneous* navigation behavior, we aim to explain how people move between places in a city, such as venues, sights, or train stations. Thus, we choose a grid-based approach with a cell size of 200m x 200m. From our experience, this cell size is small enough to distinguish places close to each other and large enough to (i) aggregate movement at a single place as well as (ii) to reduce the sensitivity due to GPS inaccuracies. Figure 7.1 shows cells of such a grid on Berlin to give an idea about the chosen granularity.

Furthermore, in our first experiments we focus on the *sequential* characteristics of different places. Thus, we remove all self-transitions (i.e., transitions from one grid cell to itself) from the photo trails in order to account for people taking several photos at one place. Basic statistics of the processed datasets are summarized in Table 7.2. For the four cities we investigate, this table lists details about the covered time spans and photo counts, as well as statistics about derived information such as the number of cells or trails. For example, we have crawled 60,978 photos in Berlin from 2010 to 2011 resulting in 4,364 trails and 6,343 covered cells out of the 43,052 cells covering Berlin. In Section 7.3, we formulate *homogeneous* hypotheses to explain this data and compare them via HypTrails [453] (cf. Section 3.3.2) in Section 7.4.1.

7.2.2.2. Tracts

We also study *heterogeneous* aspects of human navigation in the context of photowalking. To this end, we employ a more coarse level than the previously introduced fine-grained grids. That is, we focus on data from Manhattan and apply a discretization based on census tracts (administrative units). We map each photo according to its geo-location to one of these 288 census tracts (cf. Gambs et al. [190]) that we use as state space. Then, for each user, we build a sequence of different tracts she has taken photos at (again, excluding self-transitions). The final dataset contains 386,981 transitions overall. To this

dataset we apply SubTrails (cf. Chapter 5) to find subgroups with exceptional transition behavior in Section 7.4.2 and employ MixedTrails Chapter 4 in Section 7.4.3 to study heterogeneous hypotheses accounting for the discovered subgroups.

7.2.3. Points of interest

We work with hypotheses that utilize information about points of interest (POIs) in a city (see Section 7.3). We query these POIs from the social semantic web, in our case DBpedia [306], and YAGO [338].[7] For each city, the POIs are filtered by bounding box according to the properties *geo:lat* and *geo:long*. Also, area concepts such as "Germany" or "Berlin", which do not correspond to actual locations (*rdf:type* equal to *yago:District108552138*), are removed. See When formally referring to the set of all POIs for each city we use the letter Q in Section 7.3.

Additionally, we quantify the importance of a POI in some hypotheses. As an approximate measure of importance we take page view counts of the Wikipedia articles describing the POIs. For that purpose, we extracted view counts from data available at the Wikimedia download page[8] —in this work, we use the view counts for January 2012. Table 7.2 shows the number of POIs per city and their average view count.

7.3. Hypotheses about urban navigation

A major part of this case study is concerned with explaining fine-grained human navigation behavior on the overall data (assuming a homogeneous navigation process which does not distinguish between sub-processes, such as different user groups, cf. Section 7.4.3). To this end, we employ the HypTrails approach [453] (cf. Section 3.3.2) which allows to compare hypotheses about human trails on a discrete state space. Hypotheses are expressed by transition probability matrices $\boldsymbol{\phi}$ that reflect beliefs about transitions between such states. This section describes how several intuitions about photo trails can be expressed as such matrices.

7.3.1. Basic concepts

For our geo-spatial setting, we defined states by discretizing the continuous geo-spatial area as described in Section 7.2. To formulate hypotheses for HypTrails, we now formulate the beliefs about transitions between each ordered pair of states. That is, given a state s_i of a user's last photo, we specify the probabilities $\Pr(s_j|s_i)$ for her next photo to be taken at every other state s_j.

For example, if a hypothesis assumes that a user, who took her last photo at state s_1, will take the next photo at state s_2 with probability 0.5, then we set $\Pr(s_2|s_1) = 0.5$. We assign these transition probabilities between states as the values of the hypothesis matrix: $\boldsymbol{\phi} = (\phi_{i,j})$ with $\phi_{i,j} = \Pr(s_j|s_i)$. Please recall that the hypothesis matrix is a stochastic matrix since each row i of $\boldsymbol{\phi}$ sums to 1, i.e., $\sum_j \phi_{i,j} = 1$ (cf. Section 3.3.2). To simpify

[7] POIs were queried from `http://dbpedia.org/sparql` on the 15.03.2015.
[8] http://dumps.wikimedia.org/other/pagecounts-raw/

the formulas in this section, we do not directly express transition beliefs as probabilities, i.e., we skip normalization factors. Rather, we specify a belief function $\bar{P}(s_j|s_i)$ where the rows do not necessarily sum up to 1. This function can then be transformed into a probability distribution by multiplying it by a normalization factor $\frac{1}{Z}$ obtained by summing over all values of $\bar{P}$ with regard to the source state s_i:

$$\Pr(s_j|s_i) = \frac{1}{Z}\bar{P}(s_j|s_i), \text{ with } Z = \sum_{j=1}^{n} \bar{P}(s_j|s_i) \tag{7.1}$$

In this case study, we use Gaussian distributions for weighting transition probabilities. In this context, the elements of a hypothesis matrix $\boldsymbol{\phi}$ often take very small values. For computational reasons, we set the value for a belief in a transition $\phi_{i,j}$ to 0 if the transition probability falls below the threshold of 0.01. Furthermore, we set the beliefs in self-transitions to zero ($\phi_{i,i} = 0$) for all hypotheses because we are more interested in modeling actual *movement* without the influence of *stationary* processes. This is in accordance to the removal of self-transitions in the observed data for our experiments.

Next, we describe the homogeneous hypotheses that we compare for explaining human navigation behavior as observed through photo trails. We distinguish between global and local hypotheses. For *global* hypotheses the transition probabilities are the same independent of the source state, i.e., $\forall i, j, k : \Pr(s_j|s_i) = \Pr(s_j|s_k)$. For *local* hypotheses this does not hold, resulting in individual transition probabilities for each source state.

7.3.2. Uniform hypothesis

This global hypothesis believes that each transition is equally likely assuming that users take pictures uniformly at random anywhere regardless of their previous location:

$$\bar{P}_{\text{uniform}}(s_j|s_i) = 1 \tag{7.2}$$

As in the example in Section 3.3.2.1, we use the uniform hypothesis as a baseline hypothesis: an informative hypothesis should at least be more plausible than the uniform hypothesis in order to express valid notions about the processes underlying human photo trails.

7.3.3. Center hypothesis

Typically, the city center is the most lively part of a city. Thus, this hypothesis assumes that users always take their next picture near the city center regardless of the location of their last picture. To formalize this global hypothesis, we use the geographic center C of the city (as listed in Table 7.1) and lay a two-dimensional Gaussian distribution centered at this point over the corresponding discrete state space. Given the haversine distance $dist(C, j)$, cf. Sinnott [456], between the city center C and the central point of state s_j, we calculate the entries of the hypotheses matrices from the following distribution:

$$\bar{P}_{\text{center}}(s_j|s_i) = e^{-\frac{1}{2\sigma^2}dist(C,j)^2} \tag{7.3}$$

We parameterize the center hypothesis with the standard deviation σ (e.g., in kilometers). A small value of σ indicates that most pictures are taken very closely to the city center. When σ approaches infinity the hypothesis approximates the uniform hypothesis.

7.3.4. Points of interest (POI) hypothesis

Previous work on photo trails has shown that it is possible to automatically construct travel itineraries through a city by analyzing the behavior of Flickr users [135]. This suggests that humans favor points of interests — including not only tourist attractions, but also important public transportation spots or the locations of government institutions — when taking photos throughout major urban tourist areas. The POI hypothesis, which is global by nature, captures the intuition that people take a majority of pictures near such POIs. To model this, we first express an attraction force for each POI with a two-dimensional Gaussian distribution. Formally, for each state s_j and each POI $q \in Q$, we get an attraction value $G(q, j)$ that corresponds to the likelihood that the POI q is the cause for a picture at state s_j by factoring in the distance between the POI and the state:

$$G(q, j) = e^{-\frac{1}{2\sigma^2} dist(q,j)^2} \tag{7.4}$$

As before, $dist(q, j)$ describes the haversine distance between POI q and state s_j. Then, for each state, we aggregate the attraction values of all POIs $q \in Q$ in the respective city:

$$\bar{P}_{\text{poi}}(s_j|s_i) = \sum_{q \in Q} G(q, j) \tag{7.5}$$

In doing so, states that are near multiple POIs have a stronger attraction to users. Again, we have to choose an appropriate standard deviation σ; a small σ assumes that photos are taken directly at the point of interest, whereas a larger σ assumes that pictures are taken in the surroundings of a POI. Larger values of σ may represent the fact that people do not take pictures directly at a POI, e.g., to cover an architectural attraction fully in one picture. To find the POIs for each city, we utilize DBpedia as described in Section 7.2.3.

Weighted POI hypothesis. Each city contains a large amount of potential POIs. However, not all of these are equally important. In particular, studies as by Hasan et al. [232] find that popularity of places plays an important role when modeling human mobility. For example, the "Brandenburg Gate" is more likely to influence human trails in Berlin than the less known "Charlottenburg Gate". We capture this notion in a weighted POI hypothesis by approximating the importance of a POI q by the view count $views(q)$ of the Wikipedia article corresponding to this POI. If the view count of an article is very high (as e.g., for the "Brandenburg Gate"), we expect the respective POI to have a stronger influence on the sequence of image locations. We quantify this hypothesis by weighting each term of the POI hypothesis:

$$\bar{P}_{\text{weighted\_poi}}(s_j|s_i) = \sum_{q \in Q} (views(q) \cdot G(q, j)) \tag{7.6}$$

Since we expect view counts to follow a power law, we also apply a sub-linear weighting scheme to avoid overemphasizing the importance of very popular points of interest:

$$\bar{P}_{\text{log\_weighted\_poi}}(s_j|s_i) = \sum_{q \in Q} (\log(views(q)) \cdot G(q,j)) \tag{7.7}$$

7.3.5. Proximity hypothesis

The proximity hypothesis is motivated by findings of previous work [116, 214, 453]. It expresses the belief that the next image of a user will be taken nearby the last image. This is the first *local* hypothesis. To formalize this hypothesis, we consider the haversine distances $dist(i,j)$ between the center points of two states s_i, s_j. Then, we can again specify the respective transition probabilities by applying a two-dimensional Gaussian distribution:

$$\bar{P}_{\text{prox}}(s_j|s_i) = e^{-\frac{1}{2\sigma^2} dist(i,j)^2} \tag{7.8}$$

As before, a standard deviation σ must be specified; a small value of σ suggests a photo is more likely to be taken very close to a user's previous photo. An example for this hypothesis is depicted in Figure 7.1c where we visualize our beliefs in transitions from one state to other states (i.e., which is represented by one row $\boldsymbol{\phi}_{s_i}$ of the hypothesized transition probability matrix $\boldsymbol{\phi}$).

7.3.6. Mixture of hypotheses

Finally, we are interested in studying the effects of a mixture of two hypotheses. Technically, we mix two hypotheses by element-wise multiplication of the corresponding hypothesis matrices. In this case study, we focus on combining the intuition that people are likely to take pictures at POIs (or close to the city center) on the one hand, but at the same time stay close to their current location for their next photo on the other hand. We can capture this by combining the POI (or center hypotheses) with the proximity hypothesis This results in two local hypotheses as detailed in the following. Please note that other kinds of combinations are also conceivable.

Proximate weighted POI hypothesis. First, we are combining the POI hypothesis with the proximity hypothesis, i.e., we assume that people will move to a POI to take their next photo but, instead of moving to a random POI, they choose one close by:

$$\bar{P}_{\text{prox\_(log\_)weighted\_poi}}(s_j|s_i) = \bar{P}_{prox}(s_j|s_i) \cdot \bar{P}_{(log\_)weighted\_poi}(i,j) \tag{7.9}$$

Proximate center hypothesis. Similarly, the following formulation expresses the belief that the next picture is likely taken closer to the city center, but limits the area to move to a location close to the current one:

$$\bar{P}_{\text{prox\_center}}(s_j|s_i) = \bar{P}_{prox}(s_j|s_i) \cdot \bar{P}_{center}(i,j) \tag{7.10}$$

7.4. Results

In this section, we present our results of analyzing urban human navigation behavior in the context of photo trails from Flickr. This encompasses mainly experiments of homogeneous nature (i.e., without considering sub-processes like differently behaving user groups), but also includes several results on heterogeneous aspects. In particular, we focus on homogeneous processes in Section 7.4.1 using a fine-grained discretization based on grid cells. This is the main part of this study and encompasses experiments across several cities. Afterwards, we also study homogeneous explanations for urban navigation based on transitions between census tracts. In particular, we first explore subgroups with exceptional transition behavior in Section 7.4.2, and then use the accquired information to formulate and compare heterogeneous hypotheses in Section 7.4.3.

7.4.1. Modeling homogeneous behavior

In this section we study human navigation behavior through photo trails on a homogeneous level. In particular, in Section 7.3, we introduced a set of homogeneous hypotheses that express beliefs on where people take their next picture while moving through a city. In this section, we compare these hypotheses with each other based on empirical trails derived from four different cities — Berlin (Germany), London (United Kingdom), Los Angeles (USA), and New York (USA) (see Section 7.2) — by employing the HypTrails approach [453] as outlined in Section 3.3.2.

In the following experiments, we use the grid-based data as outline in Section 7.2.2.1. Also, note that we scale concentration factors κ with regard to the number of state spaces. That is, we calculate the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_{i,j})$ elicited from the hypothesis matrix $\boldsymbol{\phi} = (\phi_{i,j})$ as $\alpha_{i,j} = \kappa \cdot m \cdot \phi_{i,j}$, where m is the number of states and $\boldsymbol{\phi}$ represents transition probabilities $\phi_{i,j} = \Pr(s_i|s_i)$ between states. See Section 3.3.2.3 for details.

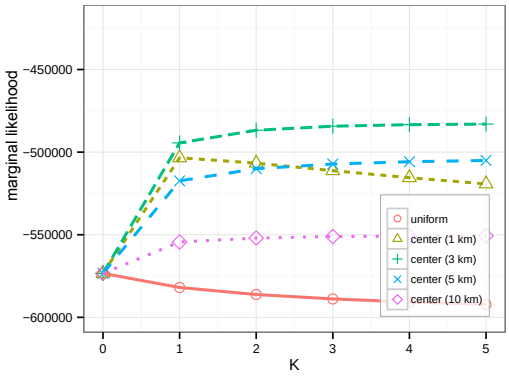
First, we focus on Berlin as a representative example in Section 7.4.1.1. We report in-depth experimental results for different parameter settings of each hypothesis. Afterwards, we report results for the other three cities in Section 7.4.1.2 focusing on the individually best parameter settings and highlight prominent differences between them.

7.4.1.1. Berlin

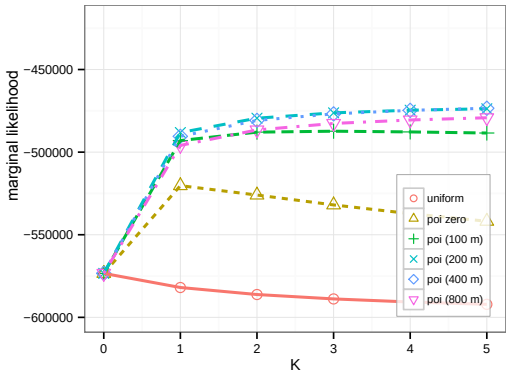
In this section, we thoroughly study each hypothesis and their different parameterizations in the same order as they have been introduced in Section 7.3, focusing on Berlin.

Center hypotheses. For Berlin, the most photos are clearly centered around the cultural center as shown in Figure 7.1a. Thus, we expect the center hypothesis — i.e., the belief that people move towards the city center and stay there for taking photos (see Section 7.3) — to be a better explanation of human photowalking behavior than our baseline (uniform) hypothesis. We use the center of Berlin from Table 7.1 and consider four different standard deviations σ: 1km, 3km, 5km and 10km. The results are depicted in Figure 7.2a.

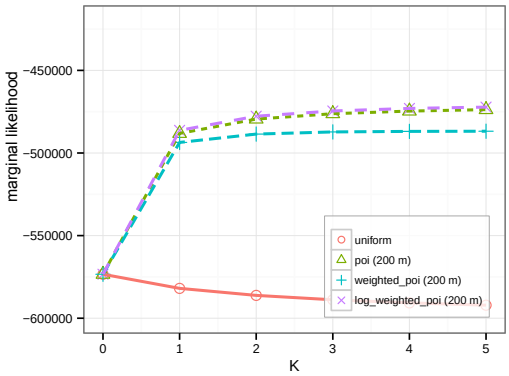
As expected, the results show that for all considered values of $\kappa > 0$ and all parameterizations of the hypothesis, the center hypothesis is more plausible than the uniform one



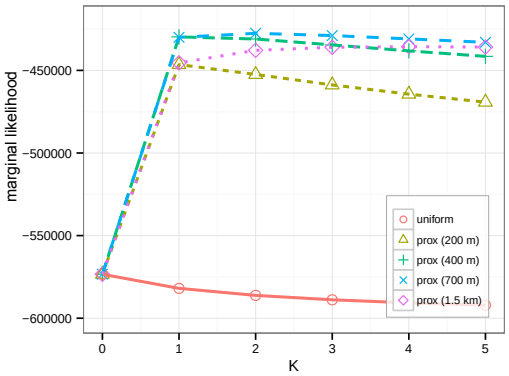
(a) Center hypotheses



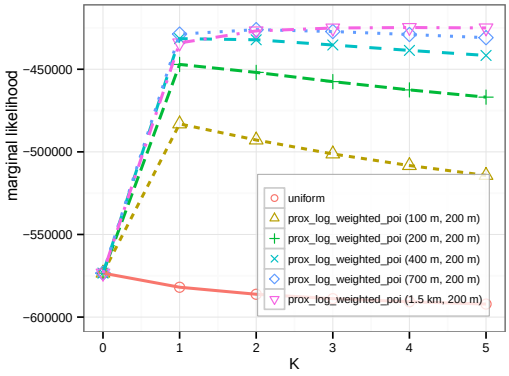
(b) POI hypotheses



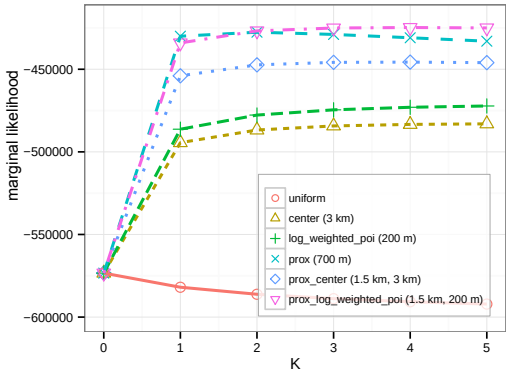
(c) Weighted POI hypotheses



(d) Proximity hypotheses



(e) Proximate weighted POI hypotheses



(f) Comparison between best hypotheses

Figure 7.2.: Comparison of homogeneous hypotheses on photo trails in Berlin. This figure visualizes the results for our hypotheses on human photo trails in Berlin. First, for each type of hypotheses at interest, we compare various parameter configurations (a-e). Then, in (f) we compare the best hypotheses from each set. Overall (f), a combination of proximity and weighted POIs provides the best hypothesis. This suggests that people prefer to subsequently take photos at popular, yet proximate POIs in a city (cf. Section 7.4.1.1).

(higher evidences). The best center hypothesis is based on $\sigma = 3$km and the worst on $\sigma = 10$km. Standard deviations of 1km and 3km are mediocre and cross for increasing κ. The initially high evidence values of 1km mean that this hypothesis covers an important aspect of the data. The quickly dropping values, however, are an indicator that it also fails to model important transitions outside the 1km radius. This is because with an increasing concentration factor κ, HypTrails [453] decreases the tolerance for a hypothesis (cf. Section 3.3.2). Contrary, for $\sigma = 5$km low values of κ show lower evidence, but it does not drop as quickly, eventually resulting in higher evidence values than for $\sigma = 1$km. This indicates that the 5km standard deviation covers most transitions, but fails to model the strong focus on the center aspect.

Overall, we find that the center hypothesis is a reasonable explanation for photowalking trails in Berlin. In detail, of the investigated standard deviations, 3km works best, while 1km is too specific and 5km is too broad.

Points of interest hypotheses. With regard to the POI hypothesis (see Section 7.3), we consider five different standard deviations: 0m (only considering the grid cell the POI is located in), 100m, 200m, 400m, and 800m. The results (see Figure 7.2b) suggest that the POI hypothesis provides good explanations about how people photowalk a city as all parameterizations indicate higher evidence compared to the baseline (uniform) hypothesis. In detail, the results show that the POI hypothesis focusing on a single state ($\sigma = 0$km) performs inferior to those POI hypotheses allowing their influence to spread. The two rather close-ranged spreads 200m and 400m perform the best, implying that people indeed move towards POIs. The worse performance of too narrow and too wide ranges is an indicator that people tend to visit places and take photos of the place at a close range, but not necessarily directly at the POI. For example, a minimum range might be required to capture a large building in one picture.

Weighted points of interest hypotheses. The weighted POI hypotheses models more popular POIs to have a stronger influence on transitions. For tractability, we focus on the best spreading parameter σ for the unweighted POI hypothesis from the previous paragraph, i.e. $\sigma = 200$m. Overall (see Figure 7.2c), the hypothesis that people prefer to take pictures at places with many *popular* POIs (here, derived from Wikipedia) provides a reasonable explanation for how people photowalk a city. By using online usage statistics from Wikipedia (view counts), we can strengthen the evidence of the hypothesis by a small — but significant — amount if we use logarithmic scaling.

Proximity hypotheses. For the proximity hypothesis (see Section 7.3), we use four different standard deviations σ: 200m, 400m, 700m and 1.5km. Overall, the results shown in Figure 7.2d demonstrate that the hypothesis that people prefer to consecutively take pictures in their proximity captures an important aspect of the production of human photo trails; $\sigma = 700$m produces the highest evidence for all considered values of $\kappa > 0$. For standard deviations of 200m and 400m, a similar situation occurs as for the center hypotheses with a standard deviation of 1km: They seem to concentrate their belief on a too narrow proximity leading to decreasing evidence values for higher values of κ. Contrary, the proximity hypothesis with $\sigma = 1.5$km is too broad, somewhat neglecting the centralized character of the proximity aspect.

Mixtures of hypotheses. To evaluate the mixture of the POI and the proximity hypothesis (cf. Section 7.3.6), we focus on the logarithmically weighted POI hypothesis with $\sigma = 400$m since it was one of the best performing hypotheses so far. This is combined with different standard deviations for proximity, i.e., 100m, 200m, 400m, 700m and 1.5km. The results shown in Figure 7.2e indeed demonstrate that adding the proximity aspect to the POI hypothesis strongly improves the evidence of the corresponding belief how people consecutively take pictures in a city. The best results can be achieved with larger standard deviations σ, i.e., $\sigma = 700$m and $\sigma = 1.5$km.

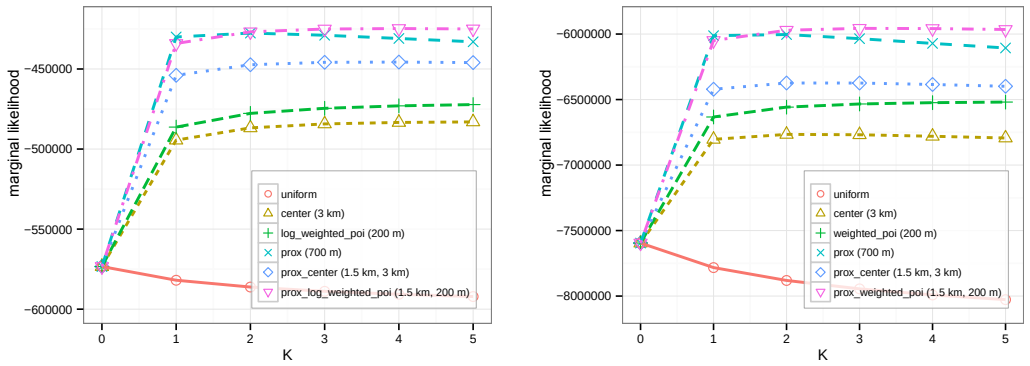
We also investigated different parametrization for the mixture of the proximity and the center hypothesis. The best parameter setting was a standard deviation of $\sigma = 3$km for the center and a standard deviation of $\sigma = 1.5$km for the proximity hypothesis. We depict the results for this hypothesis in the overal comparison in Figure 7.2f.

Comparison. For a direct comparison of the different hypotheses we are taking the most plausible ones (best parameters) of each set as elaborated beforehand. The results are shown in Figure 7.2f. We can see that the center and the *weighted* POI hypothesis perform quite similar which may be due to the larger number of (important) POIs in the city center. At the same time, the proximity hypothesis performs very well and combining it with the other hypotheses improves them strongly. Indeed, the combination of the proximity hypothesis and the *weighted* POI hypothesis provides the best explanation of how people move around Berlin while taking photos. This result suggests that information extracted from the social semantic web, in the form of concepts and usage statistics from Wikipedia, allows for finding advanced explanations for human movement trajectories.

7.4.1.2. Los Angeles, London and New York

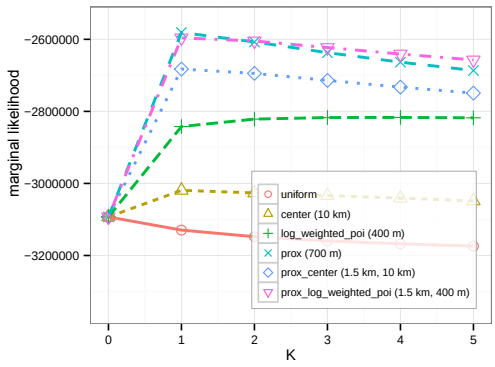
To further augment the results from Section 7.4.1.1, we analyze three more cities, namely, Los Angeles (USA), London (United Kingdom) and New York City (USA). We show similarities and highlight some differences between the cities. For a concise presentation, we focus on the best parameter settings for each hypothesis. The best parameters were determined separately for each city. Results are depicted in Figure 7.3. For most parts, all hypotheses perform very similar and the best parametrizations are consistent. This indicates that the hypotheses about photo trails in Berlin can be generalized to other cities quite well, implying that some basic patterns exist that even hold across countries.

However, there are two exceptions which are worth mentioning. First, in Los Angeles (see Figure 7.3c), the most plausible center hypothesis has a standard deviation of 10km instead of 3km. This indicates that LA either has a very large center or none at all—arguably, LA is a spread out city which may cause this divergence. Additionally, in LA higher standard deviations for the POI hypothesis, i.e., 400m instead of 200m, are favored compared to the other cities. Also, even the best performing hypotheses are strongly decreasing with increasing κ. This further supports the idea that LA is structurally different from the other cities. Second, the linearly weighted POI hypothesis in London is superior to the logarithmically weighted one. This may be due to different view count distributions and has to be further investigated in the future.

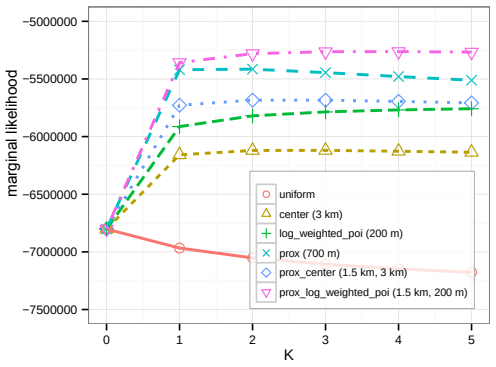


(a) Berlin

(b) London



(c) Los Angeles



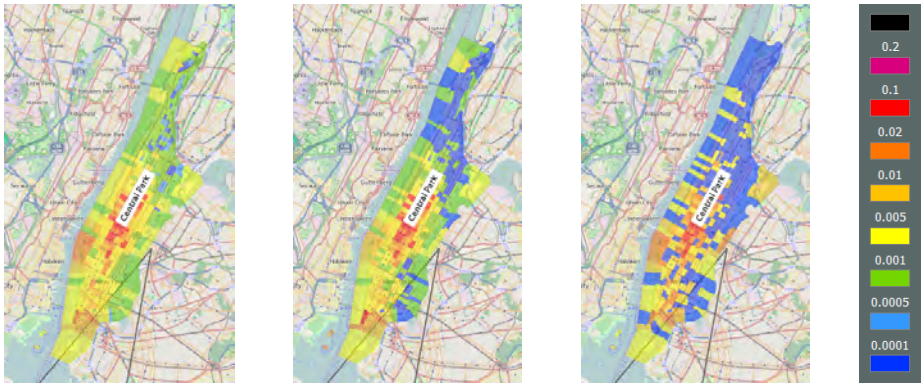
(d) New York City

Figure 7.3.: Comparison of homogeneous hypotheses on photo trails across cities. This figure visualizes the results for our studies on human photo trails derived from Berlin (a), London (b), Los Angeles (c), and New York City (d). We present a comparison of the best instances for each type of hypotheses for each city. We can identify similar explanations across cities, but there are some differences for LA and London (c.f., Section 7.4.1.2).

7.4.2. Subgroups with exceptional transition behavior

In the previous sections, we have studied the overall photowalking behavior of Flickr users in several cities. However, many related studies indicate that human navigation behavior is inherently heterogeneous (see Section 2.1.5). That is, there are multiple processes responsible for the observed overall data. To study this phenomenon, we use the photo trails over census tracts in Manhattan as described in Section 7.2.2.2 and apply the SubTrails approach as introduced in Chapter 5. In the following, we cover the corresponding experimental setup and report the results.

Experimental setup. For the following experiments, since SubTrails is not optimized for large scale state spaces, we use the census tract data described in Section 7.2.2.2 focusing on a more semantic variant of discretizing the continuous geo-spatial environment. To augment the data for discovering subgroups, we elicit a wide range of describing attributes



(a) All transitions **(b)** Transitions of tourists **(c)** Night transitions (22–23 h)

Figure 7.4.: Visualization of exceptional photo walking behavior. This figure shows transition probabilities from Central Park to other tracts in Manhattan for (a) the entire dataset, (b) the subgroup *tourists*, and (c) the subgroup *transitions during night time (22–23 h)*. Cool colors (blue, green) represent small, warm colors (orange, red) high transition probabilities, see the legend on the right hand side.

for each transition, i.e., the number of photos the respective user has uploaded from Manhattan, the number of views the source photo of the transition received on Flickr, as well as the month, the weekday and the hour this photo was taken. We add two more features based on the user's origin, that is, the tourist status and the nationality (country). We consider a user to be a tourist if the time from her first to her last photo does not exceed 21 days, cf. De Choudhury et al. [135]. Country information of a user was derived from the location field in her user profile by extracting the country using a combination of querying GeoNames[9] and specialized regular expressions. The country information was only available for about half of the users.

Based on these attributes, we use all attribute-value pairs for nominal attributes, and all intervals obtained by equal-frequency discretization into five groups for numeric attributes. Overall, this results in 163 selection conditions. With regard to the search space, we focus on subgroups with simple descriptions, i.e., no combinations of selection conditions are considered. For computing the interestingness measure, we use $r = 1,000$ random samples. We confirm our top results to be statistically significant on an $\alpha = 0.01$ level using the procedure presented in Section 5.2.2.

In a first experiment, we aim at discovering subgroups with different transition models compared to the entire data. Additionally, we investigate subgroups which match and contradict the *proximate POI* hypothesis (see Section 7.3.6), respectively. This hypothesis has been shown to be one of the best hypotheses for explaining movements in Section 7.4.1 (even if on a different state space).

Results. Table 7.3 reports our results: the most exceptional subgroups in comparison with the overall data (see Table 7.4a) describe transitions by users that take either very

[9] http://www.geonames.org/

Table 7.3.: Top subgroups of with exceptional photo walking behavior. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv} and the unweighted total variation Δ_{tv}.

(a) Comparison to the overall dataset

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Photos > 714	76,859	103.83 ± 2.41	42,277	106.68
# Photos ≤ 25	78,254	88.83 ± 2.07	37,555	141.78
Tourist = True	76,667	75.42 ± 1.79	33,418	148.64
Tourist = False	310,314	75.00 ± 1.60	33,418	16.92
Country = US	163,406	64.47 ± 1.39	44,822	70.97
# Photos = 228-715	77,448	46.10 ± 1.02	33,214	115.65
Country = Mexico	2,667	33.22 ± 0.82	3,575	122.83
# PhotoViews > 164	79,218	31.58 ± 0.74	31,461	107.84
# PhotoViews < 12	76,573	30.54 ±0.71	30,881	110.83

(b) Comparison to the *proximate POI* hypothesis, contradicting

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Photos ≤ 25	78,254	64.85 ± 1.37	110,124	221.07
# Photos = 26–81	77,003	23.41 ± 0.53	99,646	207.21
Hour = 22h–23h	14,944	18.26 ± 0.43	20,526	215.69
Hour = 23h–0h	11,726	17.42 ± 0.37	16,404	208.91
Hour = 21h–22h	17,806	16.52 ± 0.33	23,951	211.34
Tourist = False	310,314	16.09 ± 0.35	379,676	185.13
Hour = 0h–1h	9,693	15.12 ± 0.33	13,590	215.42

(c) Comparison to the *proximate POI* hypothesis, matching

Description	# Inst.	$-q_{tv}$ (score)	ω_{tv}	Δ_{tv}
# Photos > 714	76,859	58.59 ± 1.30	80,690	164.16
# PhotoViews < 12	76,573	21.56 ± 0.50	88,948	185.78
Hour = 12h–13h	25,022	14.04 ± 0.32	29,590	187.84
# Photos = 228–714	77,448	10.63 ± 0.23	91,877	193.57
Tourist = True	76,667	10.60 ± 0.24	91,214	197.79
Hour = 14h–15h	27,420	10.51 ± 0.25	33,028	194.40
Hour = 11h–12h	20,323	9.18 ± 0.21	24,613	196.99

many (more than 714) or very few (less than 25) photos. We explain this by the fact that users with overall fewer photos are more likely to travel a longer distance before taking another picture, resulting in more long distance transitions. The next two subgroups *Tourist=True* and *Tourist=False* suggest that tourists continue their trip to different locations than locals, e.g., as they are more interested in touristic attractions. Further top subgroups with deviating transition models involve the number of views pictures receive on Flickr and the country of origin.

Table 7.4b and Table 7.4c display the top subgroups that contradict the proximate POI hypothesis, respectively match it. We observe that users with small amounts of pictures and non-tourists do not move as the investigated hypothesis suggests (possibly hinting at the same user population). Also, night time mobility (roughly 21h – 1h, see the result table for exact subgroup ordering) does not match this hypothesis, maybe due to the closing of touristic attractions at night times. By contrast, tourists and users with many pictures as well as transitions at midday are especially consistent with the proximate POI hypothesis.

Although we discover these exceptional subgroups from the large set of candidates automatically, it has to be investigated post-hoc *how* the transition models deviate. In that direction, we studied the subgroup *Tourist=True* in detail. For that purpose, we first computed the source state with the most unusual distribution of target states, i.e., the row that contributes the highest value to the weighted total variation q_{tv}. For the tourist subgroup, this state (tract) corresponds to the central park. We then visualized the transition probabilities for this single state for the entire dataset and the subgroup in Figure 7.4 using the *VizTrails* visualization tool (see Section 6.2). It can be observed that tourists are less likely to move to the northern parts of Manhattan, but are more likely to take their next picture in the city center or at the islands south of Manhattan. For a second investigated subgroup, i.e., the subgroup of transitions between 22h and 23h, this effect is even more pronounced as almost no transitions from the central park to the northern or north-eastern tracts can be observed. Note, that this visualization only covers the transition probabilities from a single state, not the overall transition matrix used for detecting interesting subgroups.

7.4.3. Tourists vs. locals

Section 7.4.2 showed that there are subgroups whose transition behavior differs exceptionally from the overall data. One prominent example were the subgroups of tourists and locals. In this section, we study these two groups and their contribution to explaining navigation. In particular, we hypothesize that the navigation behavior of the population can be explained better when accounting for the inherent differences of tourists and locals. That is, we believe that tourists take their next picture at a much closer location than locals (and residents of a city). This can have several reasons including for example that tourists are commonly much more interested in their surroundings than locals, that they exhibit stronger affinity to points of interest (POIs), or that they generally take more photos. Similar to Section 7.4.2, we investigate this theory based on the photowalking trails over census tracts in Manhattan as described in Section 7.2.2.2 and apply the

MixedTrails approach as introduced in Chapter 4 to formulate heterogeneous hypotheses to explain the observed transition behavior.

Experimental setup. As a general model for tourists and locals we use a combination of spatial proximity and a preference for POIs (cf. Section 7.3.6) because in Section 7.4.1 this hypothesis was found to be one of the best explanations for the transitions of Flickr users. To account for the difference of tourists and locals, we build two different transition probability matrices that we call $\boldsymbol{\phi}_{\text{near}}$ and $\boldsymbol{\phi}_{\text{far}}$, which feature different parameterization of the proximate POI hypothesis (cf. Section 7.3.6). In particular, we fix the influence radius of POIs to $400m$ and set the standard deviation of the proximity factor to $2.5km$ ($\boldsymbol{\phi}_{\text{near}}$) and $5.0km$ ($\boldsymbol{\phi}_{\text{far}}$). Note that, here, we use larger radii than in Section 7.4.1 because, instead of considering a fine-grained grid-based discretization, we use the coarser census tracts as the underlying state space.

To classify users as tourists or locals, we use the time difference between her first and her last photo in the data, cf. De Choudhury et al. [135]. In that regard, we consider different *group assignments* (cf. MixedTrails, Chapter 4): (i) a baseline $\boldsymbol{\gamma}_{\text{one}}$ that puts all transitions into one group (regardless of being a local of a tourist), (ii) a deterministic grouping $\boldsymbol{\gamma}_{\text{det}}$ by defining tourists as users with a trail duration of 21 or less days, and (iii) a smooth distinction between tourists and locals around 21 days by using a sigmoid function $sig(t) = {}^{1}/_{1+e^{-t}}$ resulting in probabilistic group assignments $\boldsymbol{\gamma}_{\text{prob}}$.

We combine these three group assignments and transition probability matrices to form five (partially heterogeneous) hypotheses according to the MixedTrails paradigm:

- $H_{\text{near}} = (\boldsymbol{\gamma}_{\text{one}}, \boldsymbol{\phi}_{\text{near}})$
- $H_{\text{far}} = (\boldsymbol{\gamma}_{\text{one}}, \boldsymbol{\phi}_{\text{far}})$
- $H_{\text{det: tourist=near}} = (\boldsymbol{\gamma}_{\text{det}}, (\boldsymbol{\phi}_{\text{near}}, \boldsymbol{\phi}_{\text{far}}))$
- $H_{\text{prob: tourist=near}} = (\boldsymbol{\gamma}_{\text{prob}}, (\boldsymbol{\phi}_{\text{near}}, \boldsymbol{\phi}_{\text{far}}))$
- $H_{\text{prob: tourist=far}} = (\boldsymbol{\gamma}_{\text{prob}}, (\boldsymbol{\phi}_{\text{far}}, \boldsymbol{\phi}_{\text{near}}))$

For example, the last hypothesis $H_{\text{prob: tourist=far}}$ expresses a belief that there are two groups — locals and tourists — in the data, and the longer the sequence of a user is (in days), the more likely she is to be a local. Furthermore, this hypothesis assumes that tourists are more likely to have a longer distance to the next photo location than locals. We additionally added a homogeneous uniform hypothesis as a *baseline* that assumes that all transitions are equally likely and that no groups exist.

Results. Figure 7.5 shows the results. The uniform hypothesis is substantially less plausible than all *proximate POI* based hypotheses. Among the latter, we see that for smaller concentrations factors homogeneous groupings perform better, which indicates that in general the split into tourists and locals by itself does not produce particularly distinct movement behavior. However, for increasing concentration factors κ, it turns out that the hypothesis $H_{\text{prob: tourist=near}}$ works best, which uses a probabilistic group assignment in combination with the belief that tourist take their next photo at a more near-by location with a close POI while locals choose locations with higher distances more

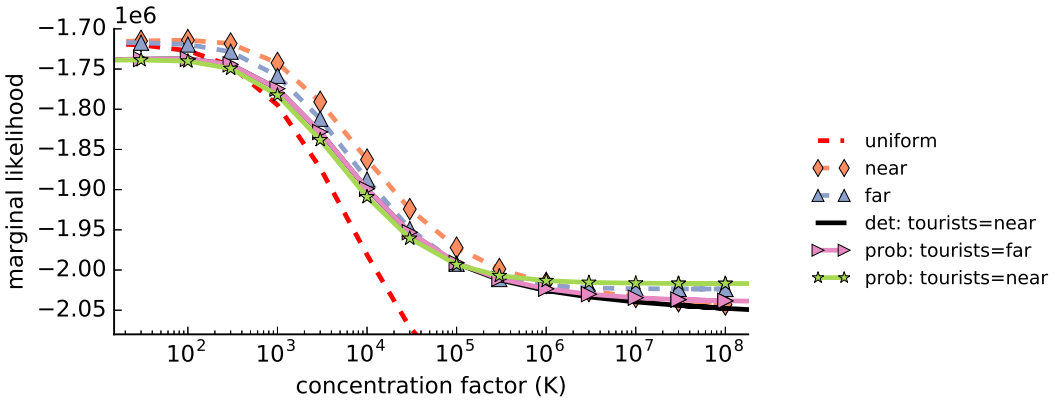


Figure 7.5.: Comparison of heterogeneous hypotheses on photo walking behavior. We model the navigation behavior between tracts in Manhattan based on photo trails on the social photo-sharing platform Flickr. Overall, we have not found hypotheses explaining the data well as indicated by the strongly decreasing marginal likelihoods. However, those we evaluated are better than the baseline, i.e., the uniform hypothesis. The best one (*prob: tourists=near*) assumes that tourists are more prone to move to *close by* tracts than locals. Here, MixedTrails allows for modeling uncertain classification of tourists which covers the underlying processes better than a deterministic group assignment (*det: tourists=near*).

often. By contrast, a deterministic split γ_{det} does not cover the uncertainty of classifying tourists and locals.

Overall, this study indicates that further investigating a combination of tourists and locals to explain the photowalking behavior of Flickr users may lead to hypotheses that better explain the overall transition behavior.[10] Also, this case study illustrates how MixedTrails can be used to incorporate heterogeneous aspects — such as the difference between tourists and locals — into navigation models specifically featuring probabilistic group assignments.

7.5. Discussion

In this case study, we have conducted extensive experiments to gain a better understanding of the underlying processes that are employed when people take photos while moving through cities. We dedicate this section to discuss characteristics specific to our approach and corresponding results, and highlight some potential limitations.

Data characteristics. Next, we shortly discuss four relevant aspects regarding our data: (i) splitting photo trails due to time constraints, (ii) observation sparsity, (iii) Flickr movement characteristics and (iv) state granularity:

[10] When comparing the plots from Figure 7.5 to the results in Section 7.4.1 please beware that the state space differs. Thus results are not directly comparable.

(i) We have considered the sequence of *all* photos of a user as a single photo trail regardless of the time span in between two photos. However, if such a time span is too long (e.g., a week or even a few hours), the corresponding two photos are most likely unrelated. Thus, in additional experiments, we have removed transitions with time intervals exceeding 6 or 24 hours respectively. The results are very similar to the ones reported in Section 7.4.1.

(ii) Since we are using grids with 200m by 200m cells over relatively large areas, the number of observations for corresponding transitions is limited. However, as HypTrails automatically focuses on observed states, the sparsity of the data does not randomly bias our results. Thus, no further testing of possibly derivable predictors is necessary since all available information is drawn directly from the data via Bayesian inference.

(iii) Due to our focus on studying Flickr, we are only able to make judgments about behavioral aspects that emerge when people move through a city and take photos as captured by Flickr. Studying other forms of mobility data might reveal different results. However, we assume that certain behavioral aspects are similar, regardless of the type of data we look at as suggested by Cho et al. [116]. This may be verified, for example by contrasting different cities or by considering different kinds of movement data, e.g., social check-ins, call details records, or business reviewing data.

(iv) We have focused on intra-city behavior using 200m by 200m grids. However, studies in the geo-spatial context have recognized the "modifiable areal unit problem" (MAUP) [257, 534] which refers to the problem of choosing appropriate areal units. Studying the stability of our results with regard to various unit settings may yield further insights into the studied processes. Similarly, we might observe different movement patterns if we extended our scope of interest. That is, by focusing on cities, we constrain our studies to a small geographic area which might favor proximity based hypotheses. If we extended the scope, for example to a country or continent level, the results will most likely largely differ. However, then, other types of hypotheses may be more plausible to study.

Choice of hypotheses. The observations in this work are limited by our choice of which hypotheses to study and how to express them; they have mostly been motivated by related work. Many other kinds of hypotheses are conceivable and can be investigated with HypTrails and our data. We suggest some potential candidates: (i) A hypothesis expressing the belief that a river is a natural barrier in a city. (ii) Also, district boundaries may be some kind of barrier. Additionally, (iii) demographic aspects (such as crime rates) might influence movement patterns in a city. And finally, (iv) hypotheses based on the notion of intervening opportunities [378] have not been tested yet.

Tourists and other heterogeneous aspects. We provide some preliminary case studies regarding the heterogeneous nature of human navigation behavior in Sections 7.4.2 and 7.4.3, mainly focusing on the user groups of tourists and locals. In particular, while

our hypotheses on tourists and locals in Section 7.4.3 were better than the uniform baseline, they did not explain the data well.

However, other studies also suggests that the photographing behavior on Flickr differs between tourists and residents of a city [135, 206]. For example, De Choudhury et al. [135] argue that residents are not under the direct pressure of visiting as many POIs within a certain time span as tourists are. This implies that hypotheses similar to the ones we formulated in Section 7.4.3 should be able to explain the observed photo trails better. Using more advanced spreading models (as opposed to a Gaussian spread) or employing a different state space may yield superior results. Similarly, a number of other user groups or sub-groups, such as visitors from different countries, or users from different generations, may be interesting to study. Finally, seasonal effects are also worth investigating. We leave these ideas to future work.

7.6. Related work

For a general overview on human navigation behavior in the context of human mobility and navigation on the web, we refer to Chapter 2. However, there are also more specific studies concerning human navigation behavior in the context of Flickr: For example, De Choudhury et al. [135] aimed at leveraging photo trails for automatically constructing travel itineraries through cities by utilizing the popularity of POIs. Travel routes have also been derived from other photo-sharing platforms like Panoramio [332]. Similarly, Tai et al. [476] used past landmarks photographed by users for recommending sequences of new landmarks derived from sequential information by other users on Flickr. Furthermore, Girardin et al. have conducted several studies on Flickr photo trails. In Girardin et al. [205], they studied digital footprints and in Girardin et al. [206] they focused on tourist dynamics based on concentrations and spatio-temporal flows revealing popular points of interests, density points, and common trails tourists follow. Also, Beiró et al. [47] used Flickr data to evaluate a method for predicting human mobility based on the gravity model (cf. Section 2.1.3). Apart from trails and mobility, Flickr has also been studied in other contexts like tagging [128, 342, 448] and social network properties [96, 360].

7.7. Conclusion

In this case study, we investigated and compared hypotheses about urban photo trails across different cities by analyzing sequences of geo-tagged photos uploaded to the Flickr platform using HypTrails [453], SubTrails, and MixedTrails (see Section 3.3.2 and Chapters 4 and 5, respectively). For this, we used discretization to transform the continuous geo-spatial observations into a discrete state space. Furthermore, for the informed specification of hypotheses, we utilized additional data sources such as DBpedia, YAGO, and view counts of Wikipedia articles which allowed us to find advanced explanations for human movement trajectories. Our results suggest that cities share interesting commonalities and differences. For example, while proximity was an overall good explanation across all cities, for the city of Los Angeles we observed movement

patterns on a different scale. Most prominently, our results suggest — at least on our data — that humans seem to prefer to consecutively take photos at proximate POIs that are popular on Wikipedia. Finally, we found exceptional sub-processes in the navigational behavior of Flickr users. In particular, we studied the difference between tourists and locals incorporating their characteristic behavior into a heterogeneous hypothesis which explained the observed navigation better than the homogeneous hypotheses we compared against. For the interactive exploration of location sequences and hypotheses, we also refer to our tool VizTrails (cf., Section 6.2).

In future work, we plan on extending our experiments by investigating additional cities. Furthermore, it would be interesting to expand the current city-level analysis to a larger scale, e.g., trails across different cities or countries.

8. Navigation processes during a participatory sensing campaign

The data collected via the EveryAware platform (Section 6.3.2) presents the opportunity to study human navigation behavior in a seldom covered context. That is, we study exploration processes for a participatory sensing campaign on air pollution. For this, we report results from our previously published article on the participatory sensing campaign "APIC" [457]. However, instead of also covering aspects like subjective perception and guesses of black carbon, we mainly focus on navigation and mobility related results. We furthermore conduct a preliminary study on corresponding navigation hypotheses using the HypTrails approach (cf. Section 3.3.2) for which we introduce a novel approach for formulating hypotheses tailored to continuous movement data.

8.1. Introduction

In Section 6.3, we have introduced the EveryAware platform for collecting mobile sensor data providing a framework for participatory sensing campaigns, quantified self projects, or Internet of Things applications. By explicitly supporting *mobile* measurements, EveryAware provides a unique source for studying human navigation behavior.

Problem setting. In this case study, we analyze human navigation behavior during the *AirProbe International Challenge (APIC)* which was held as part of the EU project EveryAware[1]. APIC was a participatory sensing campaign mapping air quality in the form of black carbon measurements across different cities.

Approach. In this context, we particularly focus on two aspects of human mobility relevant to participatory sensing campaigns, i.e., activity and coverage. The former is concerned with how much effort people put into the campaign and the latter quantifies the spatial and temporal coverage of the measurements. This allows us to extract meaningful information with regard to the behavior and goals of the users. To further study the behavioral processes of the participants, we also specifically analyze their navigation behavior by using the HypTrails approach [453] (cf. Section 3.3.2).

Contribution and findings. Overall, we employ the collected data for analyzing and studying user interests, activity patterns, as well as navigation behavior during the challenge. We find varying coverage characteristics for different locations and campaign settings. Furthermore, both, coverage and measured pollution levels, indicate that the participants had a tendency to monitor familiar areas when there was no restriction while measuring more polluted spots at the same time. And finally, similar to the coverage

[1] `http://everyaware.eu`, accessed: December 2017

characteristics, the navigation behavior shows differences with regard to the tasks given by the campaign.

Structure. In the following, we first cover background information on the APIC challenge (Section 8.2) and summarize the collected data (Section 8.3). Then, we go over several results derived from the collected data including activity and coverage statistics as well as navigation behavior in Section 8.4. We conclude this case study in Section 8.5.

8.2. The AirProbe International Challenge

The AirProbe International Challenge (APIC) challenge was aimed at studying the behavior and perceptions of citizens involved in monitoring air quality, during a large scale international test case. This was organized simultaneously in four cities: Antwerp (Belgium), Kassel (Germany), London (UK) and Turin (Italy). In this test case a web-based game, air quality sensing devices, and a competition-based incentive scheme were combined to collect both, objective air quality data and data on perceived air quality, in order to analyze participation patterns and (changes in) perception and behavior of the participants. The test case was organized as a competition between the cities to enhance participation.

During this test case, volunteer participants were asked to get involved in two activity types. The first one — which we focus on in this case study — consisted of using a sensing device (sensorbox) to measure air pollution (black carbon (BC) concentrations) in their daily life; generating what we call *objective* data. The second activity was playing a web game, where volunteers were asked to estimate the pollution level in their cities by placing flags (so called *AirPins*) on a map and tagging them with estimated black carbon (BC) concentrations on a scale from 0 to 10 $\mu g/m^3$; resulting in *subjective* data on air pollution (perceptions). Volunteers involved in the measuring activities were encouraged to play the game and bring other players as well (create a team).

The two data types allow for an analysis of user behavior and perception throughout the challenge. To enable this, the test case was composed of three phases. In phase 1, only the online game was available, so we could obtain an initial map of the perceived air pollution. In phase 2 the measurements started in a predefined area in each of the cities (corresponding also to the web game area), with the web game running in parallel. Phase 3 introduced a change in the game, so that players could acquire limited information about the real pollution in their cities in the form of sensor box measurements averaged over small areas (so called AirSquares). At the same time, measurements were continued, this time *without a restriction* of the area to be mapped. In phase 2 and 3, the volunteers received points depending on the spatial and temporal coverage they achieved with their measurements. Additionally, incentives in the form of prizes were given at the end of each phase to the best teams/players (please see the supplementary file S1 of Sirbu et al. [457] for more details).

For the first time, to our knowledge, an end-to-end scientific platform for participatory air pollution sensing was used, as developed as part of the EveryAware project. This includes a dedicated sensor box for measuring black carbon (BC), a corresponding Android

application for managing the data from sensor box, a framework for collecting, analyzing and visualizing the sensor data, and a web platform for the online game. For more information on the quality and representativeness of the collected air quality data as well as an in-depth analysis see Sirbu et al. [457]. And, for more information on the data collection platform EveryAware also see Section 6.3.

8.3. Data

We employ the mobile sensor data collected during the APIC challenge as introduced in Section 8.2. Since, in this case study, we focus on human navigation and behavior, we skip the analysis of perceptions collected via the online game.[2] This allows us to concentrate on phase 2 and 3 (of the three phases) of APIC where air quality was actively measured by volunteers carrying the EveryAware sensorboxes [162].

The APIC challenge has successfully involved 39 teams of volunteers in 4 European cities (Antwerp, London, Kassel, Turino), gathering 6,615,409 valid geo-localized data points during the second and third phase of the challenge (the measuring device collects one data point per second). Phase 2 was held from the 4th to the 17th of November 2013 and phase 3 took place from the 18th of November to the 1st of December. An additional 3,326,956 data points were uploaded to our servers in the same period but were not included in the analysis since they were missing complete GPS information. Some of these measurements contained labels (tags), with 742 geo-localized overall tags coming mostly from one location of the challenge (London).

Additional information on perception of pollution has been extracted from the online game. The platform had 288 users in total, over six weeks, 97 of which played the game at least ten times. Their activity resulted in 70,758 AirPins at the end of the test case, which were used to assess perceived pollution levels by Sirbu et al. [457].

8.4. Results

Volunteer involvement and activity levels are among the most important elements in participatory monitoring campaigns, since these can determine the success of a campaign. Large activity is required for acquiring meaningful data including both objective measurements, for analyzing of the environment itself, and subjective information, for analyzing of social behavior. In the following we discuss several navigational aspects including the activity levels of the participants, spatial and temporal coverage, as well as the goals and strategies of volunteers while measuring air quality. We also conduct a study on navigation behavior investigated by applying the previously covered HypTrails approach (see Section 3.3.2).

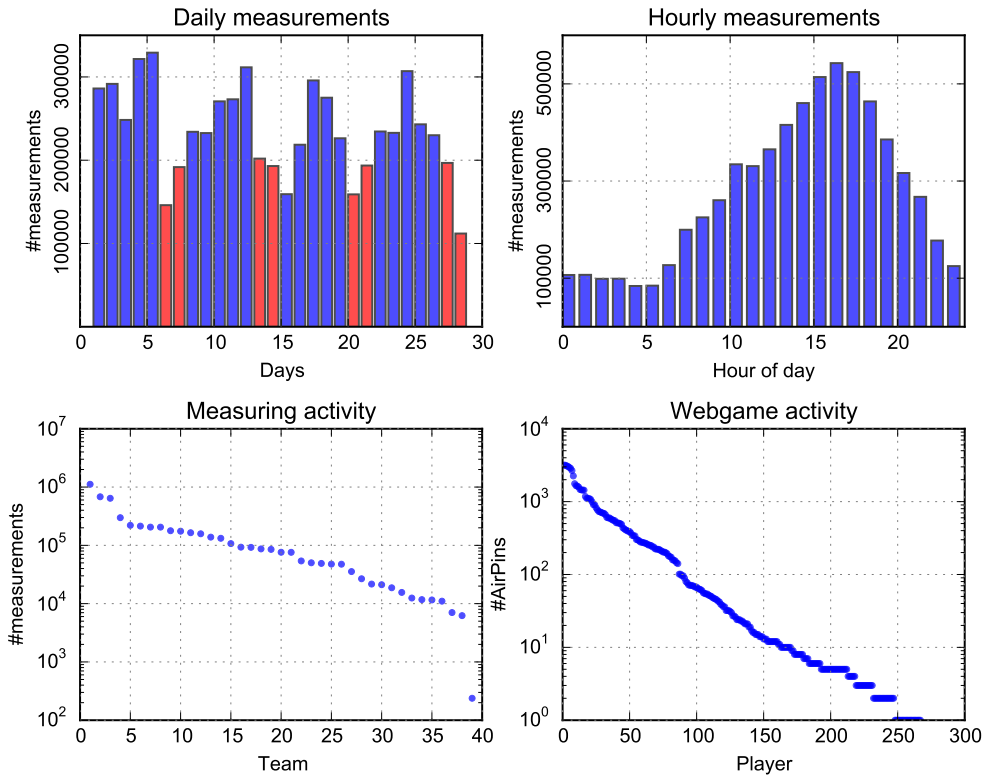


Figure 8.1.: Volunteer activity patterns during the APIC case study. The subplots in the top row show daily (weekends shown in red) and hourly measurements by volunteers. The distribution of the number of measurements performed per team is depicted at the bottom-left. We also show the distribution of the web game activity among players in the bottom-right subplot, for reference. The distributions are displayed by ranking the volunteers by activity and then displaying the number of measurements (measuring activity) and AirPins (webgame activity) in descending order, using a rank-frequency plot.

8.4.1. Activity

With regard to general user activity, Figure 8.1 shows general participation patterns. Further details about the data from the web game as well as the participation patterns for each of the four cities of APIC, can be found in [457] and its supplementary file S1. The daily number of measurements show larger activity during the week compared to weekends, with almost twice the activity in the peak days (Wednesday/Friday). This indicates that the volunteers were strongly interested in monitoring their exposure in relation to the routine activities of the week, which probably include commuting and access to highly polluted environments. It might also mean that it was easier for participants to monitor as part of their weekly routine whereby at the weekend monitoring would require more effort as (for example) it would not comprise part of their commute, or may have impacted on other leisure activities that they wanted to carry out. Daily patterns (hourly measurements) indicate a peak in activity in the afternoon, around 5 pm, again probably due to afternoon commuting. However, measurements are performed at all hours of the day, indicating the presence of very dedicated volunteers. In fact, the total number of measurements per team indicates several teams with very high activity levels, with the most active team reaching almost 1 million points (equivalent to over 270 hours of measurements). However, team activity was very heterogeneous, with some teams collecting much less data than the others. This heterogeneity was found within each city (e.g., the highly active teams are spread over three of the four cities), indicating that differences in activity were in general based on personal predisposition and not location. However, some of the heterogeneity between the cities can also be explained by the differences in instructions, emphasis, and incentives (also see the supplementary material file S1 of Sirbu et al. [457] for more details, e.g., on incentives).

8.4.2. Coverage

Besides activity in terms of number of measurements, another important aspect of participatory sensing domains is *coverage*, both in *space* and *time*. As we have seen before, measurements have been performed at all hours of the day and days of the week. However, usually not all areas and time frames are covered equally. Here we analyze aspects of coverage (for more details on individual locations please see the supplementary file S1 of Sirbu et al. [457]). In order to compute the coverage, the area of each of the four participating cities was divided into 10 by 10 meter squares (*tiles*).

8.4.2.1. Spatial coverage

In the following, one square was considered "spatially covered" if at least one measurement was performed within it. Figure 8.2 shows how the number of squares covered grows as users perform more measurements, both overall and for each phase individually. The volunteers had different tasks in the two measuring phases (phase 2 and 3 of the test

[2]For a analysis on the recorded perceptions we refer to the work of Sirbu et al. [457]

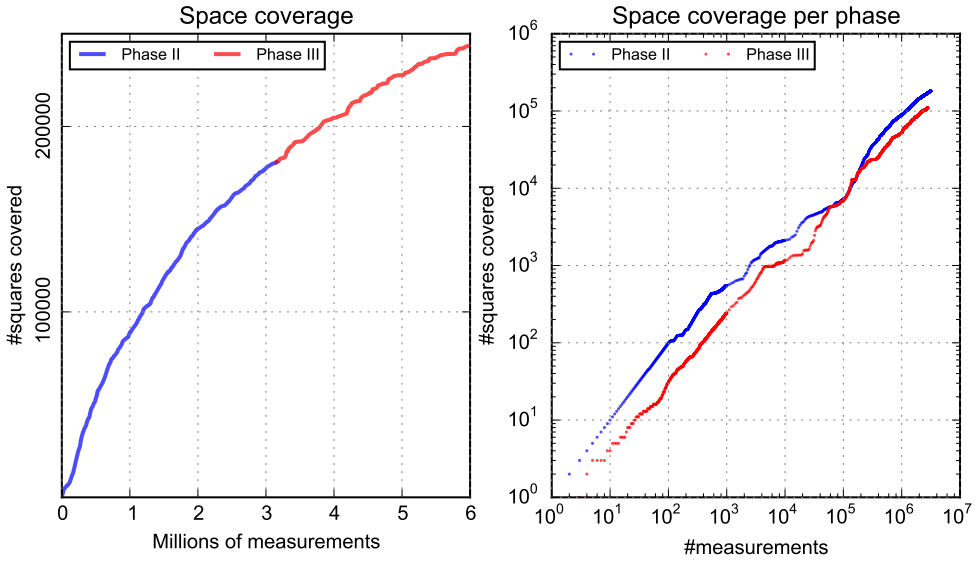


Figure 8.2.: General space coverage data. Left panel: growth of the number of squares covered for the entire challenge. Right panel: growth of the number of squares covered per phase, in a log-log plot.

case). In phase 2, they had to concentrate on covering as much as possible of a specific area, while in phase 3 they could explore any area they wanted.

In this context, Figure 8.2 indicates that space coverage grows steadily with the number of measurements, meaning that users continue to explore new areas over the course of the challenge. However, while at the beginning of the challenge the growth is fast, it decreases with time. This indicates less exploration as the challenge evolves, due to the fact that volunteers measure at the same location multiple times. Also the restricted measurement areas (especially in phase 2) may explain this effect. When looking at individual phases, it appears that during phase 2 space coverage was much better than in phase 3. This does indeed mean that volunteers displayed a better exploratory behavior at the beginning and when asked to cover a specific area of the city, compared to when they were asked to map any place they wished. In the latter case, they went for their daily routes that were not so extensive, and did not explore further. For both phases the growth of the space coverage follows a power-law, with exponent 0.73 in phase 2 and 0.79 in phase 3. This suggests that, although on the short term, the space coverage in phase 2 is larger, in the long run the strategy of phase 3 might actually produce better coverage. However, the restricted time frame of our challenge can not provide further proof for this hypothesis.

8.4.2.2. Temporal coverage

Since pollution levels vary both in space *and time*, it is important to have many measurements at the same location. So, for each tile, we also look at how measurements are spread in time, i.e., time coverage. First we separated the working days (Monday to

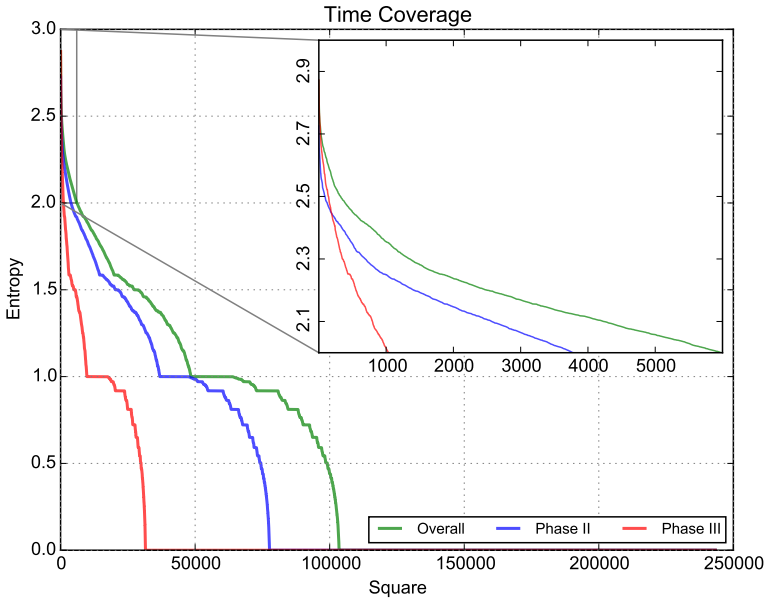


Figure 8.3.: General time coverage data. Time coverage per phase and overall. The inset shows an enlarged view of the leftmost part of the plot (top ranked squares).

Friday) from the weekends (Saturday and Sunday). Each of the two groups were divided into 4 further categories, by setting time thresholds at hours 08:00, 14:00, 18:00 and 23:00. The entropy of the resulting sets was computed. For each square, we obtained the fraction f_i of measurements in each category i as the ratio between measurements falling into that category and the overall number of measurements in that square. Then the entropy for that square is $S = -\sum_{i=1}^{8} f_i \log_2 f_i$. A higher entropy indicates a better spread of measurements in time. Figure 8.3 shows the distribution of the entropy for all squares covered in a rank-entropy plot (squares are sorted descending by entropy and the entropy values plotted for each square). A few squares had a very good time coverage. These correspond to hubs in the four cities such as leisure locations (e.g., Königsstrasse in Kassel), main squares (e.e., Piazza Castello in Turino), and transportation hubs (e.g., the Barbican and Bank subway exits in London). At the other extreme there are many squares (more than half) that have been covered only in one time slot (entropy is 0). Between the two extremes, time coverage is dropping fast.

The curves display jumps and it appears that squares can be divided into sets based on time coverage. One first set (rightmost) includes those squares that have measurements only at one time of the day (entropy 0), which is followed by those covered in 2 time slots, ending with those that are covered at all times of the day (leftmost). Within each set, coverage decays differently. While for the highly covered squares decay appears to be exponential (as plotted in the inset), this becomes slower as the coverage decreases, with curves resembling polynomial decay. This hints at different measurement processes, i.e., a *stationary* process were sensorboxes are left at the same location for extended periods of time, *routine* measuring where the sensorbox is used on routinely visited paths (e.g.,

Figure 8.4.: Heatmap representation of time and space coverage for phase 2. Red dots represent strongly covered tiles. Only the mapping area for each city is represented. The cities are, from left to right: Antwerp, Kassel, London and Turin.

Figure 8.5.: Heatmap representation of time and space coverage for phase 3. Red dots represent strongly covered tiles. The entire area of the four cities is represented. The cities are, from left to right: Antwerp, Kassel, London and Turin.

from home to work and back), and an *explorative* process probably in remote areas were a repetitive coverage is unlikely. Further investigation is necessary to confirm and explore the characteristics these processes.

When comparing the two phases, time coverage in phase 2 is much better overall than in phase 3. This indicates that volunteers not only explored more in space, but also in time, during phase 2, while in phase 3 they followed their daily schedule which resulted in poor time coverage. This underlines again the importance of giving volunteers a specific mapping area in order to obtain a better measurement spread.

8.4.2.3. Overall coverage

The overall coverage results are also displayed as spatial heat maps in Figure 8.4 (phase 2) and Figure 8.5 (phase 3). These show the areas of the four cities (mapping area for phase 2 and the entire city for phase 3) with the covered tiles. Bright colors correspond to higher time coverage, with bright red indicating the locations with most measurements. It is clear that the mapping area (phase 2) is much better covered than others (phase 3), with a few clear locations containing many measurements. These mainly correspond to landmarks and main roads in the four cities, as discussed earlier.

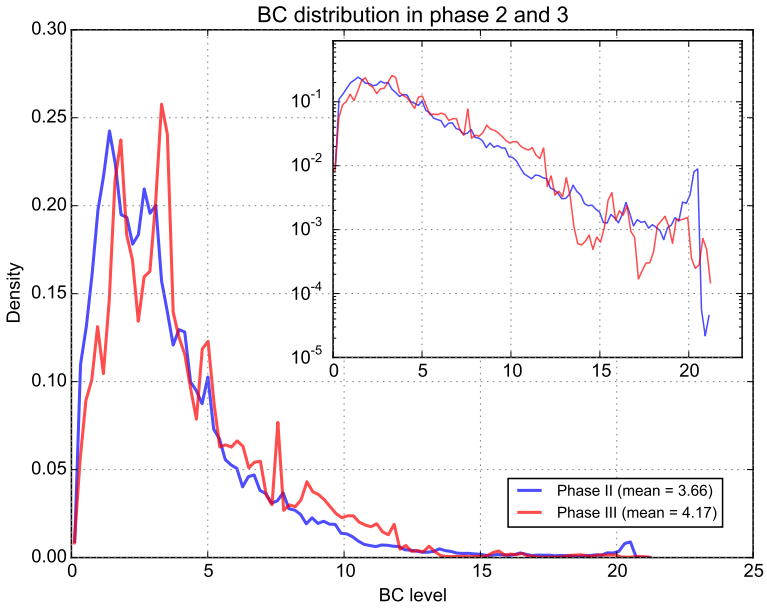


Figure 8.6.: Overall pollution levels compared between the two phases of APIC. The distribution of BC levels are shown for the two measuring phases of the challenge. The inset shows the same plot but with a logarithmic vertical axis, to emphasize the tail of the distribution.

8.4.3. Goals and strategies

The measured BC levels can also provide useful insight into the aims and strategies of the volunteers during the challenge. To this end, we can examine how these change from phase 2 to phase 3. Thus, Figure 8.6 shows graphs of BC levels measured in the two phases, and we can observe larger BC values in phase 3 (the distribution is shifted to the right). A Kolmogorov-Smirnov test was performed to test whether differences are significant and a p-value of 2.2e-16 was obtained, confirming the difference. When volunteers can freely choose where to take measurements, it appears that they primarily target more polluted areas. When the mapping area is restricted, they tend to have a more systematic approach and cover lower pollution levels as well. One may argue that pollution levels may change naturally from one day to another, so the shift we see could be due to a higher average pollution level from phase 2 to phase 3. However, a comparison with reference data seems to suggest that this is not the case (cf., supplementary material S1 of Sirbu et al. [457], where we also study additional comparisons for each location).

8.4.4. Navigation behavior

In the previous sections, we have studied activity, coverage, as well as goals and strategies of the APIC participants using different forms of aggregate statistics. While the corresponding observations already provide some information on the navigation strategies

Table 8.1.: Bounding boxes used to define grids for discretizing the APIC city areas.

	min lon.	min lat.	max lon.	max lat.
Turino	7.6017	45.0080	7.7336	45.1326
Kassel	9.3454	51.2533	9.5650	51.3617

in the context of APIC, this section specifically aims at studying hypotheses about the corresponding underlying navigational processes. We are particularly interested in the differences between phases 2 and 3, where the sensorbox users have been observed to behave differently in the previous sections.

To this end, similar to Chapter 7, we use the HypTrails approach [453] (cf. Section 3.3.2) to formulate and compare navigational hypotheses which we compare across phase 2 and 3 of APIC. In particular, we focus on hypotheses taking into account the street network of the cities and hypothesize that depending on the phase different types of streets are preferred.

Thus, in the following, we first formulate hypotheses with regard to the geo-spatial navigation behavior in the context of APIC (Section 8.4.4.2), and then evaluate and compare them on the mobile sensorbox data (Section 8.4.4.3).

8.4.4.1. Data

Trails. The underlying data for studying hypotheses on human navigation behavior in the context of APIC is the same as for the previous experiments (Section 8.3). In order to derive "clean" trails, we first apply several pre-processing steps: We first select all measurements where an accuracy is given and remove all measurements with a value above 10. Then we group the measurements by device id and sort them by their recording time to attain one trail for each sensor box. Then, in order to ensure correctly functioning sensorboxes which take one measurement per second, we split these trails whenever the time difference ($>$ 2sec), the distance ($>$ 100m), or the speed ($>$50km/h) between two consecutive measurements is greater than a given threshold.

Then, we generate a discrete state space — which is a requirement to apply the HypTrails approach — similar to our coverage studies in Section 8.4.2. In particular, we employ a 200m by 200m grid based on the bounding boxes[3] as listed in Table 8.1. We map the points of each trail to the corresponding grid cells and then (analogously to Section 7.4.1) remove all self-transitions. Afterwards we filter all trails which contain only a single entry.

Road network. For the hypotheses in Section 8.4.4.2, we use the road network of each city. We extracted these networks from OpenStreetMap[4] for each city separately from bbike.org[5]. Using this data we extract roads from the `osm_world_line` table and only

[3] These bounding boxes are based on the corresponding *woeid* ids on the *town* level. For example: `https://www.flickr.com/places/info/725003`

[4] `https://www.openstreetmap.org/`

[5] `http://download.bbbike.org/osm/bbbike/`, file date: 10.08.2017

retain entries where the field `name`, i.e., the name of the road, and the field `highway`, which defines the type of the road, are not `null` in order to concentrate our hypotheses in Section 8.4.4.2 on roads that can be tracked across cells.

8.4.4.2. Hypotheses

Using the HypTrails framework [453] (cf., Section 3.3.2), in this section, we formulate several hypotheses modeling different aspects of navigation processes in the context of APIC. To this end, analogously to Section 7.3, we define transition probability matrices $\boldsymbol{\phi}$ over grid cells as transition functions $\bar{P}$ to represent hypotheses.

The uniform hypothesis and adjacency. As in our previous case study in Section 7.3, the uniform hypothesis $\bar{P}_{\text{uniform}}(s_j|s_i) = 1$ can be considered as one of the most uninformed hypotheses. It states that all target/destination cells, not matter the distance, are equally likely to "jump" to next. This provides a baseline every other (informed) hypothesis should be able to outperform. However, since our measurements are continuous (one sample per second), and we are using a state space with 200m by 200m grids, it becomes apparent that, given the current grid cell, the next measurement can only be recorded in one of the adjacent cells.[6] Consider for example the gray cell in Figure 8.7a as the current cell. Only the cells in its immediate vicinity are candidates to navigate to. Thus, we also define the adjacency hypothesis modeling this aspect. In particular we define $adj_i(j)$ to return 1 when the cell s_j is one of the adjacent cells of cell s_i (see the eight white cells in Figure 8.7a), and 0 otherwise. Then the adjacency hypothesis is defined as

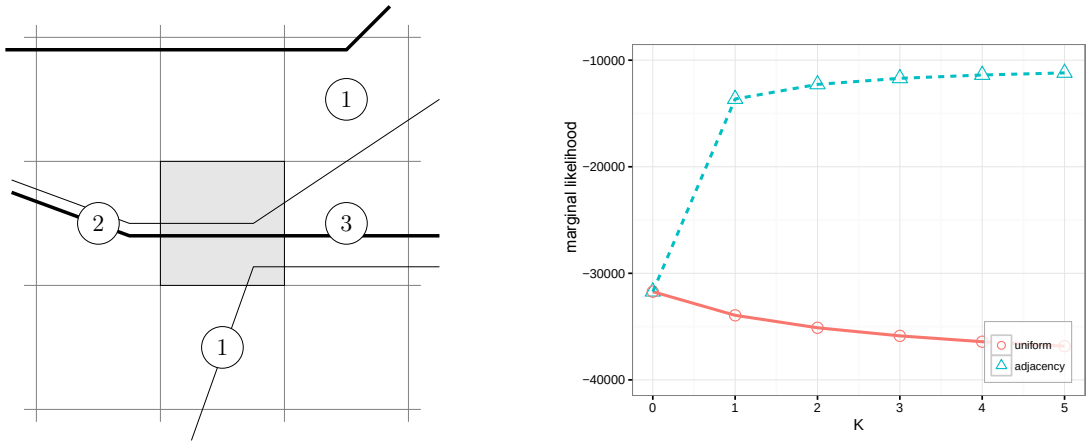
$$\bar{P}_{\text{adjacency}}(s_j|s_i) = adj_i(j) \tag{8.1}$$

Road counts. We further hypothesize that users move according to the road network. That is, given the current cell, we belief that the user will follow some road to one of the adjacent cells. We model this as follows: For each cell we extract the roads present in that cell. Then, given the names of the roads R_i of cell s_i, we count the number of roads $r_{i,j}$ of all the adjacent cells s_j which are also in cell s_i, i.e., $r_{i,j} = \sum_{x \in R_i \cap R_j} 1$. This represents the intuition that the more the roads between the source cell s_i and destination s_j overlap, the more likely a user will move to s_j. For an illustration, see Figure 8.7a: The cell below the top-right cell contains three roads also present in the grey source cell. Thus, a citizen will more likely go to that cell than to the top-right cell containing only one road also present in the (grey) source cell. We formally define the corresponding hypothesis as:

$$\bar{P}_{\text{roads}}(s_j|s_i) = r_{i,j} \cdot adj_i(j) \tag{8.2}$$

Footway and residential preference. With regard to phase 2 and phase 3 of APIC, we hypothesize that there is difference in navigation behavior. For example, we observed

[6] With regard to the notion of *global* and *local* hypotheses as introduced in Section 7.3, the limitation of transitions to adjacent cells results in only local hypotheses. This is because the transition probability distribution for each state is unique, i.e., local.



(a) Grid over street network

(b) Comparison of baseline hypotheses

Figure 8.7.: Comparison of baseline hypotheses on the APIC data. In (a) we illustrate how the road network is used to derive hypotheses preferring neighboring states with connecting roads. It shows a grid over a road network. The bold lines are residential roads, and the thin lines are footways. The number in each cell represents the count of the roads which connect to the current (grey) cell, i.e., zeros are left out. The higher the numnber of a cell, the higher the probability to move to that cell. Only adjacent cells are considered. In addition to simple counts, residential roads (thick lines) as well as footways (thin lines) can be weighted differently as done in Section 8.4.4.2. (b) shows the performance of baseline hypotheses illustrating that preferring adjacent cells is the more plausible baseline in a scenario of continuous trails.

changes in explorative behavior between phase 2 and phase 3 in Section 8.4.2. To address these characteristic properties, we investigate whether the type of the road users prefer to follow changes between the different phases. In this case study, we specifically focus on *residential* roads, as mostly found in cities, and *footways*, which are exclusively reserved for pedestrians and bicycle drivers.[7] Note however, that footways often are found alongside roads, including major roads.

Now, to model a preference for a specific road type, we weigh the different roads individually. Starting with the *residential* category, let $residential(x)$ be 2 if x is a road of the category *residential* and 1 otherwise. Then, we define the weighted sum $w_{i,j}^{\text{residential}} = \sum_{x \in R_i \cap R_j} residential(r)$ to represent the likelihood to move from cell s_i to s_j, where the residential roads are twice as important as all other road types. Consider, for example, Figure 8.7a where residential roads are bold and footways are thin. Using $w_{i,j}^{\text{residential}}$ instead of $r_{i,j}$, the weight of the cell below the top-right cell would be four instead of three. Formally, we define the hypothesis preferring residential roads as

$$\bar{P}_{\text{residential}}(s_j|s_i) = w_{i,j}^{\text{residential}} \cdot adj_i(j) \tag{8.3}$$

The hypothesis $\bar{P}_{\text{footway}}(s_j|s_i)$ is defined analogously.

[7] OpenStreetMap defines road categories *residential* and *footway* using the `highway` property. Also see: `http://wiki.openstreetmap.org/wiki/Key:highway` (accessed: 19.08.2017).

8.4.4.3. Results

In the following, we compare the hypotheses introduced in Section 8.4.4.2 on the data from phase 2 and phase 3 of APIC. As in most of the other case studies, we scale concentration factors κ with regard to the number of states in the state space. That is, we calculate the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_{i,j})$ elicited from the hypothesis matrix $\boldsymbol{\phi} = (\phi_{i,j})$ as $\alpha_{i,j} = \kappa \cdot m \cdot \phi_{i,j}$, where m is the number of states. See Section 3.3.2.3 for details. We only report results on Turino and Kassel. For the other cities (Antwerp and London), the general tendencies are the same but due to the smaller amount of data available for these cities, the results are not decisive with regard to the interpretation table of Kass and Raftery [273] (cf. Section 3.3.2.1). In contrast, the interpretations we report in the following are all backed by decisive differences.

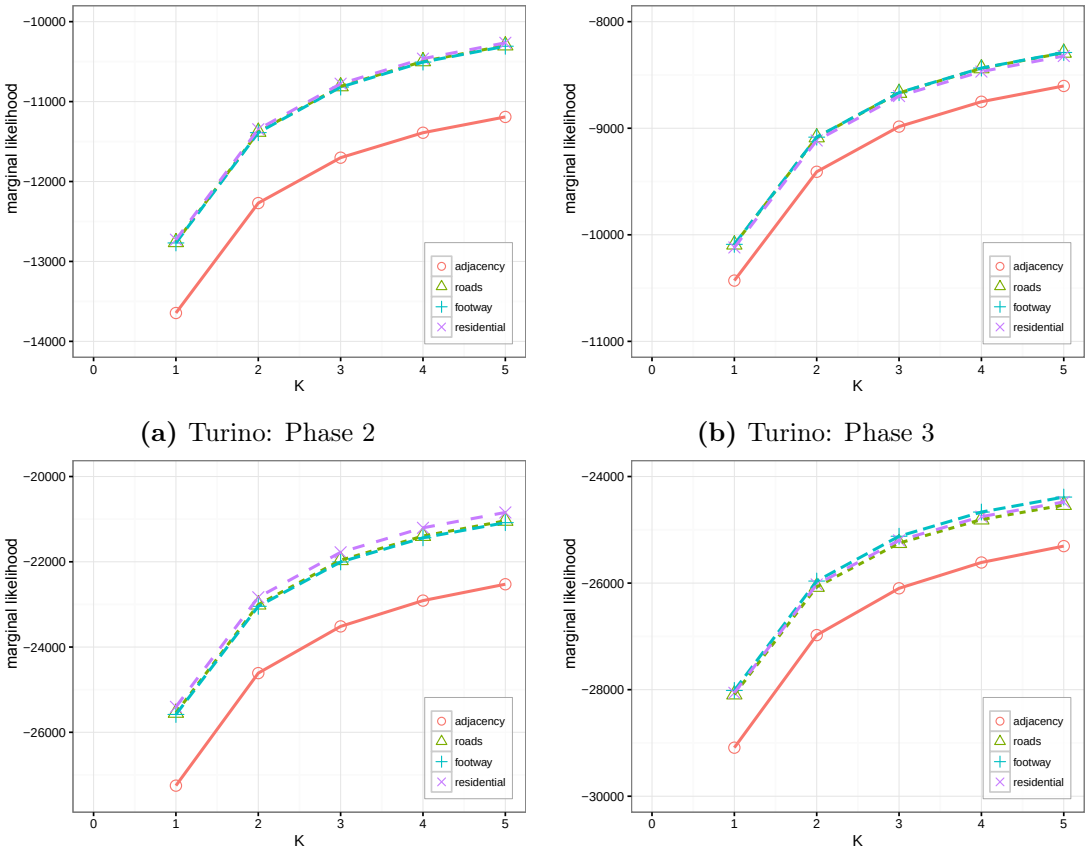
Baselines. We first compare the baselines — defined by $\bar{P}_{\text{uniform}}$ and $\bar{P}_{\text{adjacency}}$ — based on the data from phase 2 in Turin (see Figure 8.7b). As expected, we observe that — in a continuous setting — it is appropriate to restrict transitions to adjacent cells. That is, $\bar{P}_{\text{adjacency}}$ outperforms $\bar{P}_{\text{uniform}}$ by a large margin.

Roads. Next, we evaluate the performance of weighting the probability of a transitions to an adjacent cells by the number of common roads with the source cell. The results are shown in Figure 8.8. On both cities and both phases this hypothesis ($\bar{P}_{\text{roads}}$) shows large improvements on the baseline $\bar{P}_{\text{adjacency}}$. This indicates the general tendency of users to navigate according to the properties of the underlying street network so that more "links" between cells correspond to more people moving to that cell.

Residential roads and footways. Finally, we study the preference for residential roads and footways. We first concentrate on Turino (Figures 8.8a and 8.8b). In phase 2 we observe a clear improvement of the hypothesis preferring residential roads compared to weighting all roads equally ($\bar{P}_{\text{roads}}$) or preferring footways ($\bar{P}_{\text{footway}}$). This corresponds to the focused exploration of down-town Turino as visualized in Figure 8.4. What is hardly visible, is that $\bar{P}_{\text{roads}}$ slightly (but decisively) outperforms the hypothesis preferring footways ($\bar{P}_{\text{footway}}$). In contrast, in phase 2, the preference of residential roads cannot explain the navigation behavior of the users as well as the previously outperformed hypotheses ($\bar{P}_{\text{roads}}, \bar{P}_{\text{footway}}$). Here, however, the footway hypothesis ($\bar{P}_{\text{footway}}$) slightly (but not decisively) outperforms the unweighted roads hypothesis ($\bar{P}_{\text{roads}}$). This shows, that indeed, the navigation behavior between the two phases differs significantly with regard to the preference of residential roads.

Similar results can be observed for Kassel. That is, we observe the same tendencies for phase 2 where the general trend to follow residential roads is stronger than for Turino. In phase 3, things are slightly different. In particular, both, the residential and footway hypotheses outperform the unweighted roads hypothesis. Nevertheless, as for Turino, the footway hypothesis explains the data better than the residential hypothesis.

Interpretations. Comparing the different cities and the different phases two trends are apparent: i) A preference for residential roads or footways can improve on the unweighted roads hypothesis. ii) Residential roads are preferred in phase 2 while footways explain the navigation behavior better in phase 3. The former shows that road types generally carry



(a) Turino: Phase 2 **(b)** Turino: Phase 3

(c) Kassel: Phase 2 **(d)** Kassel: Phase 3

Figure 8.8.: Comparison of navigation hypotheses on the APIC data. We compare several hypotheses about human navigation during the two phases of APIC. Generally, we observe that the hypothesis assuming that participants follow *roads*, explains the data better than assuming random navigation (*adjacency*). We also find that refined information about street types (*residential* roads or *footways*) can improve on the unweighted roads hypothesis. Furthermore, we observe different preferences for one or the other road type dependent on the phase of the campaign and its objectives. This is in line with our findings in previous sections where we have observed different user behavior in the two phases.

information with regard to navigation preferences, and the latter indicates situational dependencies with regard to the overall goal and strategies of the users. This is in line with our findings in previous sections where we have observed different user behavior in the two phases. One explanation for this may lie in the focus of each phase (cf. Section 8.2): In the first phase users focused on the city centers trying to cover as much space as possible. Thus, they focused on the most common streets in these areas, namely the residential roads. When they were allowed to measure where they wanted to, the focus on the city center decreased, thus reducing the navigation on residential roads. See Figures 8.4 and 8.5 for a comparison of the respective coverage. The good performance of footways in Kassel (whereas in Turin there where hardly significant differences) may be due to the fact at the users mostly measured air quality while commuting and inherently using large roads which often have an attached footway in Kassel. Further studies will be necessary to clarify the corresponding details. Such work may explore for example the preferences for primary, secondary, and tertiary roads instead of footways in the third phase.

8.5. Conclusion

In this case study, we used the data collected in the participatory sensing campaign APIC for measuring air quality to study human behavior and mobility in terms of activity, coverage, and motivation. Our results indicated that better coverage is obtained when volunteers are assigned to a specific mapping area, compared to when they are asked to select the time and location of their measurements. Additionally, when allowed to measure freely, they (i) seemed to be attracted to places with higher pollution levels, and (ii) exhibited differing exploration behavior.

Furthermore, we applied HypTrails to study the underlying processes of the recorded measurement processes. In the corresponding experiments roads and road types played an important role for explaining the observed paths. In addition, the results confirmed a difference in navigational characteristics of the second and the third phase of APIC with regard to the road types being used. For our experiments, we also developed a way of formulating hypotheses in the context of temporally dense navigation paths (providing a location fix every second) in a continuous geo-spatial setting.

Overall, this study helps to understand the behavior of participants of participatory sensing domains. In particular, the corresponding information can be used to build user models, interpret the collected data better, or to develop new theories about the motivational processes of volunteers.

For future work, it may be interesting to extend the transition models as applied in Section 8.4.4, for example, by further investigating the influence of the road network in the context of the data provided by OpenStreetMap, by formulating heterogeneous hypotheses to explain the overall behavior of users during the APIC campaign using MixedTrails (Chapter 4), or by incorporating preferences for high BC recordings which may hint at certain measurement biases. Similarly, extending these studies to data from other participatory sensing campaigns may provide further insights into human navigation behavior as well as their incentives and goals in the context of environmental studies.

9. Browsing social tagging systems

As discussed in Chapter 1 and Section 2.2, human navigation behavior is not limited to geo-spatial navigation but can be observed in online environments as well. In this case study, we focus on such an online environment, namely the social bookmarking and publication management system BibSonomy[1]: Among other factors, we exploit the unique structural properties of folksonomies as well as the concept of semantic relatedness (cf. Section 2.2.4.3) to explain human navigation. Furthermore, we systematically study the navigational preferences of several specific user groups. Besides the possibility of employing MixedTrails (Chapter 4) and SubTrails (Chapter 5), this illustrates another way of analyzing the heterogeneous nature of human navigation. Furthermore, this study allows for novel insights in the navigational processes of online users on folksonomy structures. The content in the following sections is based on previously published work [373].

9.1. Introduction

In Chapters 7 and 8, we have studied human navigation behavior in a geo-spatial context. Contrasting, in this case study, we investigate a different form of navigational processes. Namely, we concentrate on navigation on the web. In particular, we focus on navigation on social tagging systems which have established themselves as a quick and easy way to organize and store information, such as bookmarks of websites[2] and publications[3]. In such systems, users can post resources and freely annotate them with keywords (called *tags*), for example, for later retrieval by themselves and by other users. The emerging structure over users, tags, and resources and their connections is called a *folksonomy* and serves as the main navigational concept in social tagging systems, providing links between co-occurring entities. Through those links, folksonomies possess an inherently semantic nature, i.e., as demonstrated by the emergence of a shared light-weight ontology from the assigned tags [56]. However, it is still largely unclear how users make use of the inherent structural properties of folksonomies as well as the navigation options they are given.

In particular, understanding user behavior as well as differences between various user groups is an important step towards assessing the effectiveness of a navigational concept and of improving it to better suit the users' needs. This task has attracted broad interest in the research community and previous work has focused on the navigation within one particular website [145, 519] or on the Web in general [108, 354]. Also, studying the navigation behavior of users in social tagging systems is of great interest, especially because

[1] `https://www.bibsonomy.org`, accessed: December 2017
[2] e.g., Delicious, `http://www.delicious.com`, accessed: December 2017
[3] e.g., BibSonomy, `http://www.bibsonomy.org`, accessed: December 2017

of these systems' inherent structural properties, i.e., represented by user-resource-tag triples. Consequently, different studies have addressed this issue: Heckner et al. [238] conducted a user survey on usage motivation in social tagging systems and Doerfel et al. [145] used log files to study actual navigation behavior of the overall user population through request counts. While these findings give first insights into the behavioral properties of user navigation, they focus on the overall population of users and their *global* request counts which do not capture the underlying *navigational* processes of the observed behavior.

Problem setting. Although the above mentioned work gives first insights into behavioral properties of user navigation, there exist several competing ideas and hypotheses about how users browse within a social tagging system based on *local* transition probabilities. Up to this point, such hypotheses were not objectively compared on actual navigation data and need further investigation.

Approach. To address this issue, we utilize log files of the social bookmarking system BibSonomy, which provide a unique opportunity to study the navigational trails of user groups in a folksonomy. In particular, we formulate several navigation hypotheses and compare them using HypTrails [453] (see Section 3.3.2), a method for comparing hypotheses about human movement on the Web. We also study how the performance of explaining navigation behavior differs on different data subsets, such as navigation grouped by gender, tagging behavior, or long-term experience. In the process, we revisit the aspects described by Doerfel et al. [145], and extend on their work, providing additional explanations for user intentions during navigation and their comparison. Furthermore, because tags and their inherent semantic information exert a great influence on social tagging systems, it is a logical assumption that navigating those systems is influenced by the semantic content. Thus, we explicitly search for a signal of the influence of tags on navigation, that is, a semantic component.

Contribution and findings. Our contributions and findings in this study can be summarized by three main points: i) We study different hypotheses about navigational user behavior in tagging systems. ii) We provide evidence for semantic influence on navigation. iii) We observe that users with different tagging behavior also exhibit differing navigational traits. Thus, overall, we contribute to a better understanding of navigation behavior in tagging systems and folksonomy structures. We consolidate the claims by Doerfel et al. [145] and shed light on general as well as subgroup specific behavior. Also, we expect our results to be relevant not only for researchers interested in understanding human behavior and social tagging, or operators of any system utilizing tags (e.g., Twitter), but also for the Semantic Web community.

Structure. This case study is structured as follows: Section 9.2 gives a formal definition of social tagging systems and Section 9.3 introduces the data we use in our experiments. Afterwards, we formulate hypotheses about navigation behavior on folksonomies in Section 9.4. The next section (Section 9.5) presents the results with regard to the behavior of the overall population as well as several subgroups. We finalize this case study by covering related work on navigation analysis on folksonomies in Section 9.6 and giving concluding remarks in Section 9.7.

9.2. Background

Social tagging systems have established themselves as popular means for organizing and managing digital resources on the Web. The basic idea of a tagging system is that each user can post resources and annotate these resources with freely chosen keywords (tags). By allowing users to assign arbitrary keywords to a resource, they form a powerful alternative to more traditional resource directories or catalogs with fixed taxonomies.

The structure emerging from tagging activities is called a *folksonomy*. Hotho et al. [246] models a folksonomy as a quadruple $\mathbb{F} := (U, T, R, Y)$, where U, T, and R are the finite sets of all users, tags, and resources, respectively. The set of *tag assignments* $Y \subseteq U \times T \times R$ is a ternary relation between these sets. Hereby, $(u, t, r) \in Y$ means that user u has annotated resource r with tag t. A *post* from a user u with a posted resource r and the annotated set of tags T_{ur} is defined as a set $P_{u,r} = \{(u, t, r)|t \in T_{ur}\} \subseteq Y$. This also implies that users cannot assign the same tag to a resource twice.

An example for a folksonomy is BibSonomy, a social tagging system for bookmarks and scientific publications (cf. Benz et al. [55]). Note that next to the navigation structure described above, real world implementations often introduce additional navigational features such as showing *related* tags on tag pages or a menu with links to a logged-in user's own pages. Because of this, navigation does not always strictly follow the folksonomy-induced link structure.

9.3. Data

The datasets used in this case study are based on web server logs and database contents of BibSonomy. Because in 2012 the login mechanism was modified, which introduced significant changes to the logging infrastructure, we restrict the datasets to data created between the start of BibSonomy in 2006 and the end of 2011. Anonymized datasets of logs and posts are made available to researchers by the BibSonomy team.[4] Because BibSonomy is a popular target for users who bookmark advertisements, the system uses a learning classifier as well as manual classification by the system's administrators to detect spam. In all experiments, we only use data of users that have been classified as non-spammers.

User and content dataset. We use the folksonomy data (all non-spammers with their respective resources and tags) from the BibSonomy database. In the considered time frame, 17,932 users were explicitly classified as non-spammers. They created 456,777 bookmark posts and 2,410,844 publication posts using 204,309 distinct tags. Since we need semantic similarity scores between pages for the semantic navigation hypothesis (where we assume users to navigate towards semantic similar resources/pages, cf. Section 9.4.1), we consider all tags which have been used at least twice in order to receive more meaningful results by avoiding typos or rarely used words. With this pruning step, we end up with 65,228 distinct tags.

[4] `http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/`, accessed: December 2017

Request log dataset. The BibSonomy log files include all HTTP requests to the system (caching is disabled), including common request attributes like IP address, date and referrer, as well as a session identifier and a cookie containing the name of the logged-in user. We only considered direct (i.e., no redirected) valid requests, which have been generated by logged-in non-spammers. Both the referrer and the target page of a request must be a retrieval page, that is a page that is used to retrieve information (e.g., a resource or a list of resources; we discuss each considered retrieval page type in the next subsection). For the semantic navigation hypothesis, we had to extract the tag cloud representation of each page. Because a successful request does not imply that the requested page contains any content (e.g., a user tried to filter her collection by a tag that she had not used), we only consider requests that yield a non-empty set of tags using the procedure described later in this section. The remaining dataset contains 103,415 distinct visited pages. We recorded 327,060 transitions between these pages. 123,452 transitions were self-transitions (i.e., transitions from a page to itself) and 261,300 were own-transitions (i.e., transitions, where the logged-in user owns both the source and the target page). One factor responsible for the large number of self-transitions are pagination effects.

Page types and categories. The pages we consider after filtering the request logs can be assigned to exactly one of six page types. These page types can be grouped into three categories, matching the three entity types of a folksonomy, i.e., *user*, *tag*, and *resource*. The six page types (with their corresponding categories) are:[5]

`/user/USER`
: lists all posts of the requested user `USER` (*user*).

`/user/USER/TAG`
: shows all posts which were tagged with tag `TAG` by user `USER` (*tag*).

`/tag/TAG`
: lists all resources with the tag `TAG` (*tag*).

`/url/RESOURCE_IDENTIFIER`[6]
: describes pages of bookmarked weblinks to the same web page (*resource*).

`/bibtex/RESOURCE_IDENTIFIER`
: describes pages that show all publication posts, with the same resource contributed by different users (*resource*). Similar to the previous page type.

`/bibtex/RESOURCE_IDENTIFIER/USER`
: shows all information that the user `USER` added for a specific publication (*resource*).

For bookmarks, no *details* pages exists. Instead, clicking a bookmark directly leads to a page outside of BibSonomy. Thus, these requests are not tracked by the logs.

Tag clouds as semantic page representations. Since each retrieval page in BibSonomy shows a set of posts, we can define a *tag cloud* for each page. Given a page s_k, its

[5] Both `/url/RESOURCE_IDENTIFIER` and `/bibtex/RESOURCE_IDENTIFIER` have been restructured and redesigned in mid 2016. They now show combined information about the web page or the publication instead of a list of the same resource.

[6] BibSonomy calculates an identifier for each resource (URL or publication). See `http://www.bibsonomy.org/help_en/InterIntraHash` (last accessed: December 2017) for more information.

tag cloud is defined as the set of tags with their respective frequencies, which are assigned to the posts of this page. For example, the tag cloud of a page showing two posts, one resource tagged with *social* and *web* and another resource tagged with *social*, *concept* and *web*, would be $tagcloud(s_k) := \{(social, 2), (concept, 1), (web, 2)\}$. The corresponding document-term vector v_k for the page s_k with the above mentioned tags as features would thus yield $v_k := (2, 1, 2)$, which can be used to represent the page s_k.

9.4. Hypotheses on navigation in social tagging systems

As in the previous case studies (Chapters 7 and 8) we formulated abstract ideas about geo-spatial human navigation as hypotheses to compare them by using HypTrails [453] (cf., Section 3.3.2). However, instead of discretized geo-spatial state spaces, here, users visit web pages provided by a social tagging system. Analogously, hypotheses are represented by transition probabilities $\boldsymbol{\phi} = (\phi_{i,j})$ between these pages which we formulate by defining transition functions $\bar{P}(s_j|s_i)$ which can be converted into the required probability distributions by normalizing the values for each source state s_i.

In this section, we first define basic hypotheses, motivated by general ideas about how users possibly navigate social tagging systems. Afterwards, we combine selected hypotheses in order to see how the different aspects influence each other.[7]

9.4.1. Basic hypotheses

First, we formulate basic hypotheses, each representing a basic aspect of navigation.

Uniform hypothesis. Similar to our example in Section 3.3.2.1 and previous case studies (Sections 7.3 and 8.4.4.2), the uniform hypothesis (also called teleportation hypothesis) serves as the *baseline* for all other hypotheses. It models the assumption that users randomly choose an arbitrary page to visit next, without regard for the underlying link structure. Formally, this is expressed as:

$$\bar{P}_{\text{uniform}}(s_j|s_i) = 1 \tag{9.1}$$

Since this hypothesis does not require any additional information, it can be considered the least informative one. Any informed hypothesis capturing a structurally interesting aspect of user behavior, should exhibits a higher evidence than this simple hypothesis.

Page consistent hypothesis. Results found by Doerfel et al. [145] motivate the idea that users often make a transition from a page to itself. This might be accounted for by various reasons, for example to follow pagination, that is, showing the next n elements in a truncated list. This hypothesis is formalized as:

$$\bar{P}_{\text{page}}(s_j|s_i) = \begin{cases} 1, & \text{if } s_i = s_j \\ 0, & \text{otherwise} \end{cases} \tag{9.2}$$

[7] With regard to the notion of *global* and *local* hypotheses as introduced in Section 7.3, all proposed hypotheses (except the uniform baseline) are local, since the transition probability distributions for each state depend on state specific attributes instead of global statistics that hold for all states equally.

Category consistent hypothesis. Doerfel et al. [145] found that transitions between two pages often occur between pages of the same category, i.e., after a user has visited a *tag* page, the next page is likely to be a *tag* page again. The same holds for *resource* and *user* pages. The classification of pages into one of these categories is described in Section 9.3. Thus, the hypothesis states that users stick to the same category. With $cat(s_k)$ denoting the category of page s_k, this is defined as:

$$\bar{P}_{\text{cat}}(s_j|s_i) = \begin{cases} 1, & \text{if } cat(s_i) = cat(s_j) \\ 0, & \text{otherwise} \end{cases} \tag{9.3}$$

User consistent hypothesis. Similarly to the category consistent hypothesis, this hypothesis assumes that a transition's target and source page belong to the same user. The motivating intuition for this hypothesis is that visitors, who are interested in the work of a specific user, will not only read one, but several of her articles and try to further explore her personomy (i.e., the subset of the folksonomy that only contains the user and the resources and tags posted by the user). With $user(s_k)$ denoting the user associated with page s_k, this is defined as:

$$\bar{P}_{\text{user}}(s_j|s_i) = \begin{cases} 1, & \text{if } user(s_i) = user(s_j) \\ 0, & \text{otherwise} \end{cases} \tag{9.4}$$

Folksonomy consistent hypothesis. Social tagging systems map links of the underlying folksonomy to actual hyperlinks of the system. For example, the page of a resource contains hyperlinks leading to the page of the resource's owner as well as to the pages of the assigned tags. For that reason, this hypothesis assumes that users navigate only to pages which are reachable (i.e., via a single action/link) using the folksonomy structure or by taking advantage of *related-tags* relations. To calculate reachability, we construct the page graph from the tag-assignments in the folksonomy dataset and (since they are an integral part of the BibSonomy user interface) we add tag-to-tag relations, when tags occur together at the same post. Formally, we define:

$$\bar{P}_{\text{folk}}(s_j|s_i) = \begin{cases} 1, & \text{if } s_j \text{ is directly reachable from } s_i \text{ in the folksonomy} \\ 0, & \text{otherwise} \end{cases} \tag{9.5}$$

Semantic navigation hypothesis. We aim to investigate the influence of a potential semantic component in navigation behavior. To compute the corresponding semantic similarities between two pages, a page s_k is treated as a document. The set of tags which appear on that page with respective frequencies (see Section 9.3) is treated as the document's "text", represented as the `TF-IDF` vector v_k. Thus, the similarity of two pages is calculated with the cosine measure $cossim(v_i, v_j) = \langle \hat{v}_i, \hat{v}_j \rangle$, where $\hat{v}_k$ denotes the normalized vector of v_k. The hypothesis is defined as:

$$\bar{P}_{\text{tfidf}}(s_j|s_i) = cossim(v_i, v_j) \tag{9.6}$$

9.4.2. Combining hypotheses

In order to investigate possible mutual influences between hypotheses, it is also possible to combine them. In the following, we motivate and describe certain combinations.[8]

Folksonomy consistent and semantic navigation hypothesis. As described earlier, it is a natural assumption that users utilize the folksonomy structure when navigating a social bookmarking system. If the folksonomy does indeed exhibit notable semantic properties, we should be able to see that adding a semantic component to folksonomic navigation improves the evidence of this hypothesis compared to the bare folksonomy navigation hypothesis. We define the hypothesis as:

$$\bar{P}_{\text{folk-tfidf}}(s_j|s_i) = \bar{P}_{\text{folk}}(s_j|s_i) \cdot \bar{P}_{\text{tfidf}}(s_j|s_i) \tag{9.7}$$

User consistent and semantic navigation hypothesis. A similar motivation as with folksonomic and semantic navigation arises when we combine user consistent and semantic navigation. Users are usually thematically restricted in their research interests and can thus serve as a good selector for a limited field of topics. Furthermore, navigation in the user's personomy is expected to show a strong semantic component. This hypothesis is defined as:

$$\bar{P}_{\text{user-tfidf}}(s_j|s_i) = \bar{P}_{\text{user}}(s_j|s_i) \cdot \bar{P}_{\text{tfidf}}(s_j|s_i) \tag{9.8}$$

User consistent and folksonomy navigation hypothesis. The intuition behind combing user consistent and folksonomic navigation is that, if navigation is mostly user consistent and partially follows folksonomy induced links, folksonomic navigation on pages from the same user should yield a good model of navigation. We define this hypothesis as:

$$\bar{P}_{\text{folk-user}}(s_j|s_i) = \bar{P}_{\text{user}}(s_j|s_i) \cdot \bar{P}_{\text{folk}}(s_j|s_i) \tag{9.9}$$

9.5. Results

In this section, we evaluate hypotheses about how users navigate on BibSonomy by employing HypTrails [453] as introduced in Section 3.3.2. We first compare our (homogeneous) hypotheses (cf. Section 9.4) on the overall request log dataset. Then, we manually (as opposed to using SubTrails from Chapter 5) analyze behavioral characteristics of different subsets of the data in order to investigate specific subgroups of the data in a target-oriented way.

Note that, in the following experiments, analogously to most of our case studies (e.g., Sections 7.4.1 and 8.4.4.3), we scale the concentration factors κ with regard to the number of state spaces. That is, we calculate the Dirichlet parameters $\boldsymbol{\alpha} = (\alpha_{i,j})$ elicited from the hypothesis matrix $\boldsymbol{\phi} = (\phi_{i,j})$ as $\alpha_{i,j} = \kappa \cdot m \cdot \phi_{i,j}$, where m is the number of states. For details please see Section 3.3.2.3.

[8] Note that we combine hypotheses by multiplication only. Due to the binary nature of most of our hypotheses, this results in a very restrictive process, i.e., the probabilities of many transitions that either of the combined hypotheses deems plausible are set to zero. However, we leave further combination schemes, e.g., using weights, to future work. See Chapter 10 for an example of using weights in the context of task choosing behavior on crowdsourcing platforms.

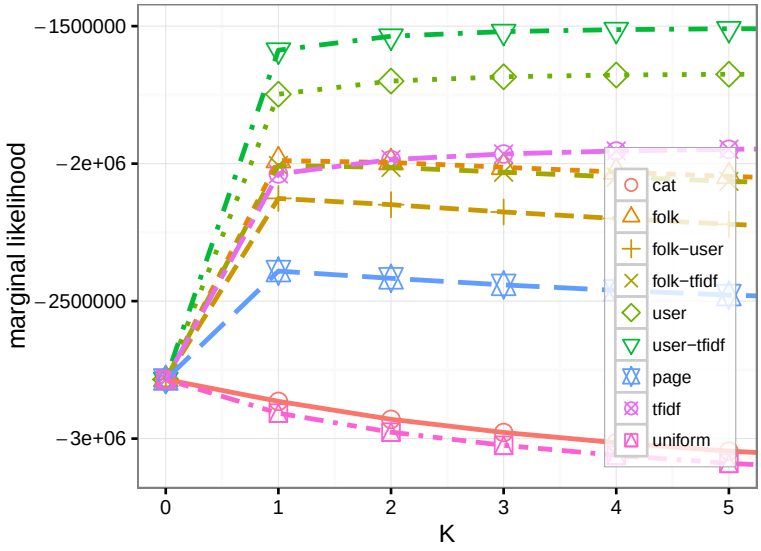


Figure 9.1.: Comparison of hypotheses for overall navigation behavior on BibSonomy. The chart shows the marginal likelihood curves for our hypotheses on the complete request log dataset. Of the basic hypotheses, the user consistent hypothesis explains the data best, followed by the semantic and the folksonomy hypotheses. When combining the user consistent hypothesis with a semantic bias, the evidence improves. This indicates that users are actually semantically biased while navigating through resources. In contrast, combining other hypotheses with the folksonomy hypothesis does not yield better explanations for the observed navigation.

9.5.1. Overall request log dataset

In this section, we focus on the overall dataset (cf. Section 9.3) on which we frist compare the basic hypotheses (cf. Section 9.4.1) followed by the combined hypotheses (cf. Section 9.4.2). Figure 9.1 shows the results.

Basic hypotheses. All of the basic hypotheses explain the observed transitions better than the baseline (the uniform hypothesis) to varying degrees. This indicates that they all contain at least some navigational characteristics explaining the observed transitions.

Besides this fact, there is a clear order of hypotheses: the user consistent hypothesis works best, the semantic and the folksonomy hypotheses are somewhat similarly plausible, followed by the page consistent and the category consistent hypotheses. Many of the observed effects are explainable by the large number of self-transitions in the dataset caused, for example, by pagination (cf. Section 9.3):

1. The page consistent hypothesis strongly improves on the uniform hypothesis.
2. The category consistent hypothesis is more plausible than the uniform hypothesis, even though it directly contradicts navigation as induced by a folksonomy structure.
3. The user consistent hypothesis as well as semantically induced hypotheses are strongly favored because of self-transitions.

Nevertheless, the user consistent as well as semantically induced hypotheses are also more plausible than the page consistent hypothesis, indicating that their structural properties cover further important factors. That is, the superiority of the user consistent hypothesis indicates that users indeed navigate mostly on their own resources (cf. Section 9.3). The good performance of the semantic hypotheses indicates that semantic similarity of pages (with regard to tags) is a strong explaining factor for navigation on our dataset from BibSonomy.

Finally, we consider the *folksonomy hypothesis* which models the navigation we expect in a folksonomy (see Section 9.4.1). It performs similarly well as the semantic hypotheses. We observe that the corresponding evidence curve crosses the semantic hypothesis (*tfidf*) for increasing concentration factors κ. This indicates that the folksonomy hypothesis covers an important factor of the navigation, but fails to model certain transitions, which are covered by the semantic hypothesis. The fact that the folksonomy hypothesis cannot cover certain transitions is due to navigation outside the folksonomy structure as elaborated in Section 9.3.

Combined hypotheses. Overall, the combination of the user consistent and the semantic hypotheses performs best, indicating that navigation on BibSonomy can mainly be explained by semantic navigation within the resources of a specific user.

In contrast, combining the folksonomy hypothesis with the semantic hypothesis decreases the observed evidence slightly. Also, combining the folksonomy hypothesis with the user consistent hypothesis decreases the observed evidence dramatically. Both observations indicate that users excessively take advantage of additional navigation features provided by BibSonomy (see Section 9.2) when navigating on their own resources. Interestingly, in Section 9.5.2, we see that this does not hold when users navigate outside their own scope, that is, on resources exclusively from other users.

9.5.2. Request log subsets

Due to the heterogeneous nature of human navigation (cf. Section 2.2.5), we expect that there are subsets of our data where some hypotheses perform differently than on the overall dataset. Instead of applying SubTrails to automatically find subgroups with exceptional transition behavior (cf. Chapter 5), here, we investigate prominent subgroups found in literature. In particular, we investigate different data subsets as listed in Table 9.1.

Inside and outside navigation. Motivated by the fact that users often navigate on their own pages (cf. Section 9.3), we investigate whether users behave differently when they are browsing the folksonomy *outside* of their own resources. In particular, we study the transitions where the source as well as the target state do not belong to the browsing user. This encompasses 42,192 transitions, cf. Table 9.1. The results can be seen in Figure 9.2 Note, that we only show the results for outside navigation because the results for inside navigation (the source and target state belong to the navigating user) hardly differs from the results on the overall dataset (see Figure 9.1).

The best explanation for the observed navigation is the *folk-tfidf* hypothesis which considers semantic behavior in combination with the structural properties of the folksonomy. This allows us to conclude that while users do not use the folksonomy structure

Table 9.1.: Details on the request-log subsets from BibSonomy.

	source states	links	counts
overall	55,129	149,542	327,060
inside	37,244	105,222	261,300
outside	14,757	28,760	42,193
male	23,090	61,616	130,988
female	5,598	14,413	29,705
neutral	28,726	73,575	161,830
lower_trr	30,368	83,268	176,755
upper_trr	7,084	15,474	32,517
lower_ten	3,459	6,959	15,451
upper_ten	51,542	140,844	307,072
short-term	10,285	21,912	48,221
long-term	45,535	126,453	274,302

when accessing their own resources (because they most likely explicitly access known publications), they fall back to the folksonomy structure when browsing resources outside of their scope. Furthermore, the evidence for the user consistent hypothesis drops strongly compared to the other hypotheses, because it is restricted to user consistent navigation *outside* of the browsing user's resources. This leads us to believe that browsing outside of the own resources is a process aimed at the discovery of new resources which in turn is not bound to the ownership of resources. Additionally, the fact that the (plain) user-consistent hypothesis performs similarly well as the self-transition hypothesis indicates that the observed user-consistent outside navigation is mostly due to pagination effects.

User gender. Since gender bias in online systems is an active research area [499], we also investigate the navigation for different genders. In BibSonomy, users can set their gender explicitly. If no gender was set, we assign the label *neutral*, otherwise, we can distinguish between *female* and *male*. These classes contain 161,830, 29,705, and 130,988 transitions respectively (cf. Table 9.1).

We hardly observed any difference between genders, thus we do not show dedicated plots. There is only a slight difference for the semantic hypotheses compared to the folksonomy hypothesis. It seems that navigation behavior of male users shows a tendency towards following the folksonomy structure whereas the navigation behavior of female and especially neutral users can be better explained by the semantic (*tfidf*) hypothesis.

Usage continuity. Since we expect users to adapt to systems they are using, we investigate if their navigation behavior changes over time. We divide users into *short-term* and *long-term users*, according to the temporal difference of their first and last request. If the difference is less than half a year, we classify a user as a *short-term user* and as a *long-term user* otherwise. This results in 48,221 transitions from short-term users and 274,302 transitions from long-term users, cf. Table 9.1. In Figure 9.3, we report the results for short-term users. The results for long-term users are very similar to the results of the overall dataset (cf. Section 9.4.1).

When comparing against the overall dataset (or, equivalently, the long-term user group),

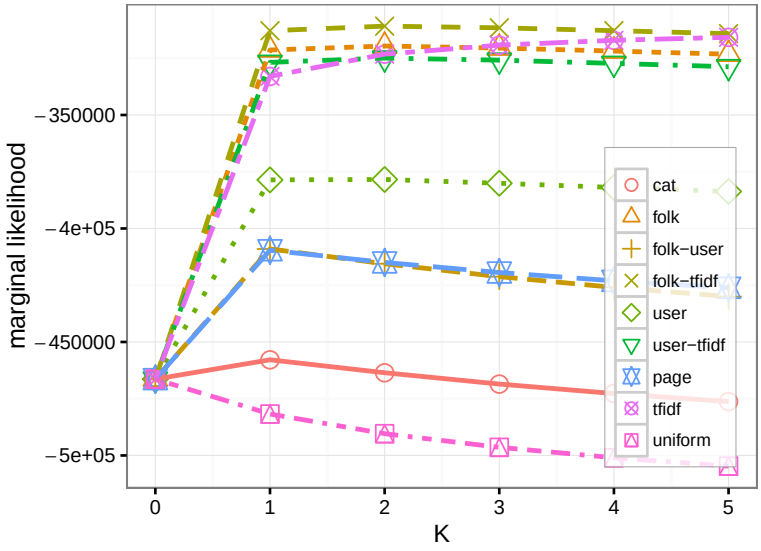


Figure 9.2.: Comparison of hypotheses for outside navigation on BibSonomy. This figure shows the marginal likelihood curves for hypotheses on the subset of navigation outside of the user's resources. In contrast to the overall dataset, we observe that outside navigation can be explained best by a hypothesis assuming semantic behavior on the folksonomy structure (cf. *folk-tfidf*).

we observe two conspicuous differences for short-term users. First, the semantic hypothesis performs significantly better when compared to the folksonomy hypothesis. Secondly, the page-consistent and the folk-user hypotheses are explaining navigation equally well. The former may be explained by the fact that new users are not as tuned to the folksonomy structure as long-term users. Thus, we may actually observe a learning process: the longer users work with the folksonomy, the more they exploit the folksonomy structure in order to navigate their own resources or to discover new ones. The similar evidence curves for the page-consistent and the folk-user hypotheses can be explained by increased pagination effects while exploring the system in combination with the lack of transitions on resources owned by the browsing user. In contrast to outside navigation, here, the lack of transitions on own resources can be explained by the fact that new users have no or a lot less own resources than long-term users.

Tagger classes. In Körner et al. [284] and Niebler et al. [375], different types of folksonomy users were characterized by their tagging behavior. Körner et al. [284] define *categorizers and describers* and Niebler et al. [375] identify *generalists and specialists*. Categorizers and describers are classified by their *tag-resource-ratio* (or short *trr*). That is, while categorizers use a small set of different tags for a large number of resources, indicating elaborate category systems, describers use many different tags, indicating a very descriptive approach when tagging. Generalists and specialists are classified using *tag entropy* (or short *ten*), where generalists have a high tag entropy, indicating a wide variety of tagged topics with regard to their resources, while specialists have a low entropy

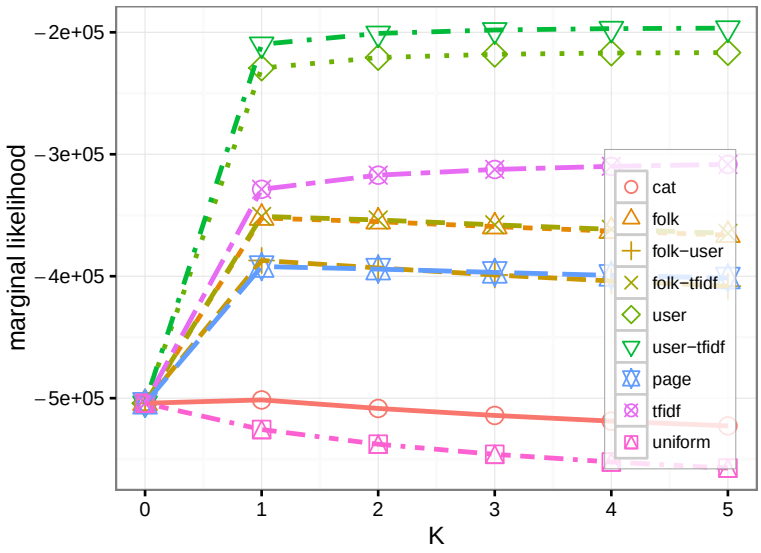


Figure 9.3.: Comparison of hypotheses for short-term users on BibSonomy. This figure shows the marginal likelihood curves for our hypotheses on the subset of short-term users ($\leq$ half a year according to their first and last request). We observe a stronger performance of the semantic hypothesis compared to the folksonomy hypothesis and see that the self and folk-user hypotheses perform equally well in explaining navigation. We attribute this to the increased browsing aspect of new users.

indicating a very specialized set of topics. For both classes we order users according to their *trr* and *ten* values separately and select the upper and lower 30%, respectively. For statistics on all tagger classes, such as the number of states, links, and transitions, please see Table 9.1.

We observe that categorizers and generalists show very similar evidence curves when compared to the overall navigation dataset. For the describers and specialists the curves are very similar to those of the previously studied short-term users, thus, we refrain from depicting them.

For both, describers and specialists, we see the same tendency as for short-term users (cf. Figure 9.3): the semantic hypothesis works better compared to the folk hypothesis and the user-consistent hypothesis has a tendency to perform equally well as the folk-user hypothesis.

The tendency towards semantic navigation over structural navigation on the folksonomy structure can most likely be explained by the nature of the tagging types: Specialists can be considered to be interested in rather few abstract topics, implying a more directed browsing behavior than generalists (whose interests are more varied). Consequently, their navigation is expected to also be more semantically influenced, because of their use of a small, but highly specialized tag subset. As for describers, resources are tagged with more keywords. Thus, for the semantic measure based on TF-IDF, calculating the similarity may simply be easier than on the very sparse tagging structures induced by a categorizer's

tagging habits.

In general, both types, specialists and describers, can be considered to be of a more explorative nature, as can be seen by the relative performance drop of the folk-user hypothesis and/or the increase of evidence for the self-transition hypothesis.

9.6. Related work

In the following, we cover work related to our analysis of human navigation in the context of folksonomies and social tagging systems. For a broader overview on navigation on the web see Section 2.2.

The term *folksonomy* was first mentioned by Vander Wal in 2004 on his personal blog.[9] He used this term to describe the underlying structure connecting users, tags, and annotated resources in social tagging systems. While Mathes [346] hypothesized that tag distributions follow a power law distribution, thus possibly causing semantic stabilization, Golder and Huberman [211] showed that after a certain time span, regularities in user activity, tag frequencies and relative frequency proportions could be observed. In turn, this motivated further investigations of tagging systems, especially about the effective extraction of semantically stable content [93] and motivation of tag usage [284]. Heckner et al. [238] conducted a user survey on the users' motivation for using social tagging systems, that is, whether users store resources for their own retrieval or social sharing purposes. While there exists a large amount of literature on tagging systems, to the best of our knowledge, there exists only a small amount of work utilizing and analyzing log data. Millen and Feinberg [357] investigated user logs of the social tagging system Dogear, which is internally used at IBM and thus not publicly available, as opposed to BibSonomy which we use in this case study. They found strong evidence for social navigation, that is, users are looking at posts from other people instead of mainly their own. In Doerfel et al. [145], a thorough study of user behavior on BibSonomy was presented. With our experiments in this case study, we extended these findings, focusing on the actual *navigation* behavior of users.

9.7. Conclusion

Understanding human navigation in web systems is an important step towards improving the design and usability of web pages. In this case study, we analyzed navigational behavior of users in a social tagging system. We presented several hypotheses on navigational patterns and evaluated them on a large web-log dataset of the social tagging system BibSonomy.

Beyond confirming the results by Doerfel et al. [145], that is, that users mainly navigate on their own resources, we were able to show that within these resources, navigation follows a semantic bias (cf. Section 9.5.1). Also, the semantic hypothesis performed well in general, confirming the semantic component in navigation behavior on BibSonomy.

[9] `http://vanderwal.net/random/category.php?cat=153`

Furthermore, we studied different navigation subsets which represents an alternative to applying MixedTrails and SubTrails (cf. Chapters 4 and 5, respectively) for studying heterogeneity in navigational processes. The results showed significant differences in behavioral characteristics. This includes that even though semantic, user consistent navigation represented a major aspect of the navigational characteristics of BibSonomy, users fell back to the folksonomy structure when browsing outside of their own pages (cf. Section 9.5.2). Also while different genders did not exhibit interesting behavioral deviations, short-term users, as well as different tagging types, followed certain behavioral patterns matching their individual characteristics. In particular, while it has been only hypothesized in prior work [284] that categorizers and describers (as well as generalists and specialists) differ in navigation behavior, we found specific components of their behavior which differed significantly, thus, indicating that navigation behavior and tagging pragmatics are indeed connected.

Overall, we were able to gain new insights into the underlying processes of navigation in tagging systems, which can be extended and leveraged in the future, for example, by considering new hypotheses, improving navigation experience, or extracting semantics.

10. Choosing campaigns on crowdsourcing platforms

In this case study — similar to Chapter 9 — we investigate navigation processes on the web. However, instead of directly studying browsing or searching behavior on web resources, we are investigating the more abstract notion of task choosing behavior on crowdsourcing platforms. As in the previous case studies, we formulate and compare hypotheses based on HypTrails (cf. Section 3.3.2). On a methodological level, this case study is noteworthy because it presents an approach to cope with temporal constraints when formulating hypotheses. That is, we account for the limited availability of individual campaigns to choose tasks from. Besides this theoretical contribution, to the best of the authors knowledge, this study presents the first study to explore hypotheses about task choosing behavior on crowdsourcing platforms based on large-scale log data instead of often small numbers of handcrafted surveys. We have previously published the work presented in this section [40].

10.1. Introduction

Crowdsourcing platforms are a relatively new type of large scale Internet services and represent a specific type of online labor markets. In contrast to traditional forms of organizing work, the crowdsourcing paradigm is characterized by the fact that tasks are not assigned to a specific person. Instead, employers define *campaigns* consisting of a set of *tasks* which are made available on crowdsourcing *platforms*. Then, the users of this platform — so called *workers* — freely choose from the pool of available tasks. The granularity of work on crowdsourcing platforms is smaller than in traditional forms of work organization [244]. This results in pools of hundreds to thousands of different tasks [254], which users have to navigate when choosing their workload.

Problem setting. The large number of tasks and campaigns on crowdsourcing platforms poses challenges: On the one hand, *workers* face the issue of efficiently finding tasks fitting their profile, e.g., according to their skills or their interest. On the other hand, *employers* need all their tasks to be completed. Both interests have to be addressed by the crowdsourcing companies. To solve these issues, mechanisms like task recommendation systems have been identified as a relevant research topic [281].

However, recommendation systems as well as similar mechanisms to support the workers in choosing their tasks need prior knowledge about task selection preferences. Unfortunately, there is only little information about how workers navigate the space of available tasks on crowdsourcing platforms: Current studies are based on surveys which only cover small subsets of workers and are also highly subjective.

Approach. In this case study, we address this lack of quantitative studies and objectively evaluate the influence of different factors involved in the selection process of tasks. To this end, we employ logs from the commercial crowdsourcing platform Microworkers.com[1] and — according to the navigation paradigm of this thesis — interpret the set of tasks a user has completed as ordered trails. Then, similar to the previous case studies, we apply the HypTrails approach [453] (cf. Section 3.3.2): Based on results from related papers, we formulate hypotheses on task choosing behavior and compare them directly on the observed trails which feature the work history of 39,100 workers over 6 years. Among others, the hypotheses considered in this work are based on campaign categories, monetary incentives, or semantic similarity of campaigns.

Contribution and findings. We objectively evaluate a considerable set of hypotheses and find that, in our scenario, those based on work categories and employers as well as campaign descriptions work best. Our approach enables crowdsourcing companies to better understand their users in order to optimize their platforms, e.g., by incorporating the knowledge gained about these factors into task recommendation systems.

Structure. The reminder of this case study is structured as follows. We first cover background on crowdsourcing platforms in Section 10.2. Then, the applied methodology and the underlying dataset are described in Section 10.3. The considered hypotheses are introduced in Section 10.4 and the results of our experiments are presented in Section 10.5. Section 10.6 discusses the results. Finally, in Section 10.7 we give a general overview of related work on influence factors on task selection in commercial crowdsourcing environments and conclude this case study in Section 10.8.

10.2. Background

We study how users choose tasks on crowdsourcing environments. Commercial crowdsourcing environments usually involve three actors: (i) platform users submitting work to the platforms, so called *employers*, (ii) users completing work submitted to the platform, so called *workers*, and (iii) the *platform operators*. As mentioned before, unlike in traditional forms of work organization, employers do not choose dedicated workers for completing the submitted work. Instead they define certain tasks and make them available through the crowdsourcing platform. The workers can then freely choose from the currently available work. Usually employers never communicate directly with workers. Instead, the platform and its operators are responsible for providing means to publish work on the platform, submit completed work, and transfer remuneration between worker and employer.

While a wide variety of crowdsourcing platforms exists, micro-tasking platforms, such as Microworkers, focus on highly repetitive tasks which can be completed in a short amount of time (a few minutes up to an hour). Micro-tasks include, e.g, tagging a series of images or categorizing the sentiment of a set of short text messages. Due to the repetitive nature of micro-tasks, employers publish a *campaign* describing a class of tasks and set a number of task instances for workers to complete for that particular campaign. A campaign ends

[1] `https://microworkers.com/` (accessed: Aug. 2015)

when all tasks have been completed. Each task can only be chosen once by a single worker. On the Microworkers platform, which we consider in this work, workers also cannot choose more than one task from the same campaign. Thus, in the following, the notion of *campaign* and *task* are used interchangeably.

Our main goal is to study how workers choose their tasks in crowdsourcing environments. Since we use HypTrails [453] (cf. Section 3.3.2) as our method of choice, we need to model this process as a set of transitions between campaigns. As each task is associated with a campaign, we can also derive a "trail" of campaigns for each user. That is, each trail consists of the campaigns associated with the tasks she has completed consecutively. At the same time, these trails define the transitions we need for the HypTrails approach.

10.3. Data

We use a dataset from the crowdsourcing platform Microworkers.com.[2] The data includes anonymized information about campaigns and users in a time period between the founding of the platform in May 2009 and January 2015. In the following, we first explain two specific characteristics of the dataset we need to consider in order to apply HypTrails [453] (cf. Section 3.3.2) and then introduce several campaign features we use for defining hypotheses (cf. Section 10.4).

Data restrictions. With respect to some special characteristics and features of Microworkers, we have to limit the data utilized for our HypTrails computation. Microworkers offers the possibility for employers to restrict their campaigns to workers from specific countries. To keep things as simple as possible, we choose to focus on US workers because they have access to most campaigns. In order to do so, we remove campaigns which place restrictions on US workers as well as campaign transitions from non-US workers.

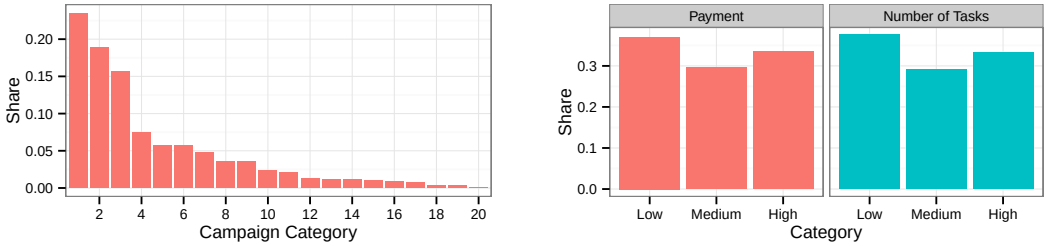
Additionally, instead of releasing all tasks of a campaign at once, employers can set its tasks to be released successively at a certain speed. However, since we define hypotheses based on transition probabilities between campaigns, we need to model which campaigns are available after finishing a task. The task release speed feature complicates this process. Thus, we only consider campaigns with a large enough speed in order to guarantee that this does no restrict the workers artificially.

In spite of these restrictions, our final dataset still contains $81,544$ campaigns and $3,415,119$ completed tasks. This includes 95% of the US workers and corresponds to 55% of the campaigns available to them as well as 60% of their completed tasks.

Campaign features. For defining hypotheses in Section 10.4, we use several features based on campaign properties. These include campaign categories, payment, the time required to finish a task, payment per hour, and the number of tasks offered by a campaign. These properties are introduced in the following.

On the Microworkers platform, each campaign is associated with one of 20 campaign *categories*, e.g., "promotion" or "writing". The distribution of the campaigns per category is shown in Figure 10.1a. The campaigns are not uniformly distributed, i.e., three categories

[2] `https://microworkers.com/` (accessed: Aug. 2015)



(a) Share of campaigns per predefined category on Microworkers.

(b) Share of campaigns per payment and campaign size classes.

Figure 10.1.: Selected statistics of the Microworkers dataset.

are very prominent. Each of these categories contains between 15% and 25% of the campaigns, whereas most of the other categories only contain 5% or less. On the one hand, the popular categories include simpler tasks with lower requirements to complete these tasks successfully. On the other hand, new categories were added over time, which is also responsible for this imbalance.

Besides the categories, campaigns differ concerning their *payment*. However, payments are strongly skewed towards small amounts of money. Since we later want to gauge whether workers generally tend to choose campaigns which are better payed, we divide the payment range into three classes, i.e., low, medium and high. We choose these intervals so that campaigns are as equally distributed as possible. Figure 10.1b shows the resulting distribution. The intervals are: \$0.0 to \$0.15 for low, \$0.15 to \$0.30 for medium, and amounts of more than \$0.30 for highly paid tasks.

The *time required to finish a task* is also a feature we will use to derive hypotheses. The time is set by the employer and is an estimation about how long a single task will take approximately. However, here it is not possible to derive equally sized intervals, since one time setting is far too prominent.

Using the payment and the time required to finish a task, we can derive the *payment per hour (pph)* for each campaign. Again, we define intervals so that campaigns are as equally distributed as possible. The intervals are: \$0.0 to \$2.4 for low, \$2.41 to \$6.0 for medium, and amounts of more than \$6.0 for high.

Additionally, each campaign defines a different number of tasks (also called *positions*) which ranges from 30 up to several hundred. Similar to payment, campaigns often only provide a rather low number of tasks. Consequently, we again choose to define equally sized intervals. The intervals are: only campaigns with 30 tasks for low, 31 to 90 tasks for medium, and 90 tasks and up for high volume campaigns. The resulting distribution is depicted in Figure 10.1b.

10.4. Hypotheses

In the crowdsourcing environment users choose tasks from different campaigns. These campaigns correspond to the states S of the Markov chain employed by HypTrails [453] (cf.

Section 3.3.2) to formulate hypotheses. Analogously to our other case studies, hypotheses are represented by transition probabilities $\boldsymbol{\phi} = (\phi_{i,j})$ between these pages formulated by transition functions $\bar{P}(s_j|s_i)$ which can be converted into the required probability distributions by normalizing the values for each source state s_i.

10.4.1. Uniform hypothesis and availability

With the uniform hypothesis we assume that, after finishing a task from a campaign, a user will randomly choose a task from any other campaign. As in the other case studies (e.g., Sections 7.3 and 9.4), this hypothesis will serve as our baseline due to its "uninformed" nature. Formally, the uniform hypothesis is defined as:

$$\bar{P}_{\text{uniform}}(s_j|s_i) = 1 \tag{10.1}$$

However, since campaigns are only available for a user to choose as long as there are some incomplete tasks, the point in time when a user chooses her next campaign defines the set of campaigns available to choose from. That is, a user can not choose a campaign whose tasks have all been completed. Now, assume that a user just finished working on a campaign s_i. Then, let $[t^{(i)}_{\text{start}}, t^{(i)}_{\text{end}}]$ denote the time interval in which campaign s_i was active, i.e., $t^{(i)}_{\text{start}}$ is the time campaign s_i was made available for users and $t^{(i)}_{\text{end}}$ is the time the last task has been completed. Thus, the user finished her task sometime between $t^{(i)}_{\text{start}}$ and $t^{(i)}_{\text{end}}$. In the best case, the user finished her task at $t^{(i)}_{\text{start}}$, i.e., right after campaign s_i started. If we now assume that a user may wait arbitrarily long before choosing her next campaign, all campaigns s_j ending after campaign s_i has started ($t^{(j)}_{\text{end}} \geq t^{(i)}_{\text{start}}$) are available to the user. However, all campaigns $s_{j'}$ which have ended before campaign s_i has started ($t^{(j')}_{\text{end}} < t^{(i)}_{\text{start}}$) are not available to the user. Now, for all users, given a campaign s_i they just finished, we define a set of available campaigns $S^{(i)}$:

$$S^{(i)}_{\text{after}} = \{s_j \in S \mid t^{(j)}_{\text{end}} > t^{(i)}_{\text{start}}\} \tag{10.2}$$

This availability setting assumes that a user may take an arbitrarily long time for choosing her next campaign. However, since tasks usually require a short amount of time to complete and campaigns only pay small amounts of money for theses tasks, a user whose goal is to earn a sensible amount of money will choose her campaigns in quick succession. Thus, it is realistic to assume that users only pick from campaigns available at the time they finish. This can be modeled by assuming that a campaign s_j is only available from another campaign s_i, if the active time of both campaigns overlap, i.e., $[t^{(i)}_{\text{start}}, t^{(i)}_{\text{end}}] \cap [t^{(j)}_{\text{start}}, t^{(j)}_{\text{end}}] \neq \emptyset$. Formally, the corresponding set of available campaigns, given a campaign s_i is defined as:

$$S^{(i)}_{\text{overlap}} = \{s_j \in S \mid [t^{(i)}_{\text{start}}, t^{(i)}_{\text{end}}] \cap [t^{(j)}_{\text{start}}, t^{(j)}_{\text{end}}] \neq \emptyset\} \tag{10.3}$$

Thus, given different definitions of availability $S^{(i)}$, the most natural extension of the uniform hypothesis is to set the probability of campaigns which are not available from a given campaign to zero:

$$\bar{P}_{\text{av}}(s_j|s_i) = \begin{cases} 1, & \text{if } s_j \in S^{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{10.4}$$

We formulate two corresponding hypotheses, namely $\bar{P}_{\text{after}}$ and $\bar{P}_{\text{overlap}}$. The uniform hypothesis $\bar{P}_{\text{uniform}}$ is equivalent to the availability hypothesis where all campaigns are available from every campaign: $\forall s_i \in S : S^{(i)} = S$.

10.4.2. Category and employer

An important factor when choosing campaigns can be the tendency of users to stick with categories of tasks which they are used to and employers which they know judge their work fairly. Next, we formulate hypotheses incorporating these two aspects.

Category. Crowdsourcing platforms often define a set of categories in order to group certain types of campaigns (see Section 10.3). Building on the idea that users like certain types of tasks better than others (for example "interesting" ones as indicated by Aris [17]), we propose a hypothesis which favors follow-up campaigns of the same category. Let the category of a campaign s_i be denoted as cat_i, and let α define the weight for campaigns with a category of the same type and β define the weight for campaigns with a category of a different type, then the corresponding hypothesis is:

$$\bar{P}^{\alpha,\beta}_{\text{cat}}(s_j|s_i) = \begin{cases} \alpha \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } cat_i = cat_j \\ \beta \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{otherwise} \end{cases} \tag{10.5}$$

For this, we only consider campaigns s_j *available from* the given campaign $s_i \in S$ as denoted by the factor $\bar{P}_{\text{av}}(s_j|s_i)$.

Employers. Furthermore, Schulze et al. [438] have shown that reputable employers are generally favored. Thus, we define a hypothesis assuming that users are consistent with regard to their employer when choosing their new tasks. Let the employer of a campaign be denoted as emp_i, then we define the employer hypothesis as:

$$\bar{P}^{\alpha,\beta}_{\text{emp}}(s_j|s_i) = \begin{cases} \alpha \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } emp_i = emp_j \\ \beta \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{otherwise} \end{cases} \tag{10.6}$$

Mixture. While both hypotheses can be a good explanation for how users choose their next task, we want to combine the two notions. That is, we assume that users choose from the same employer and at the same time they also like to work on the same type of tasks, thus also staying consistent regarding the category:

$$\bar{P}^{\alpha,\beta,\beta',\gamma}_{\text{cat\&emp}}(s_j|s_i) = \begin{cases} \alpha \cdot \bar{P}_{av}(s_j|s_i), & \text{if } cat_i = cat_j \wedge\ emp_i = emp_j \\ \beta \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } cat_i = cat_j \wedge\ emp_i \neq emp_j \\ \beta' \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } cat_i \neq cat_j \wedge\ emp_i = emp_j \\ \gamma \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{otherwise} \end{cases} \tag{10.7}$$

For the mixture, in this work, we set $\beta = \beta'$ and write $\bar{P}^{\alpha,\beta,\gamma}_{\text{cat\&emp}}(s_j|s_i)$.

Skewed probabilities. However, as mentioned in Section 10.3, the overall distribution of campaigns within categories and by employers is not equally distributed. As a result, there are significantly more campaigns in some categories. Consequently, more campaigns of this category will be chosen by workers. This possibly favors the category and employer hypotheses mentioned above. In order to investigate whether this is true, we also formulate a hypothesis based on overall category frequencies. Let f_i denote the frequency of a category in our corpus. Then we define:

$$\bar{P}_{\text{cat-freq}}(s_j|s_i) = f_j \tag{10.8}$$

Equivalently, we define $\bar{P}_{\text{emp-freq}}(s_j|s_i)$, for employer frequencies. These hypotheses model the overall probabilities to choose specific categories and employers.

10.4.3. Payment, positions, and time

As has been shown in several studies [e.g., 438, 549], the amount of money to be earned from a task can be considered a decisive factor in choosing new tasks. In this context, there are two aspects to cover, namely task payment [549] and hourly earnings [438]. The former is the amount of money to be payed for finishing a task. The latter also takes into account the estimated time required to finish a task. Another factor which has been shown to influence the users' preferences to choose tasks is the number of available positions of the campaign [113, 438]. That is, different campaigns provide different numbers of tasks and users seem to favor campaigns which provide more tasks.

Stratified classes for payment and positions. Both, the payment and the position factors, have in common that a higher value, i.e., higher payment or more positions, implies a higher probability to choose the corresponding campaign. Let $value_i$ denote the corresponding value. A straightforward formulation of a corresponding hypothesis would be

$$\bar{P}_{\text{value}}(s_j|s_i) = value_i \tag{10.9}$$

However, when considering the value distributions, we notice that payment as well as positions are strongly skewed towards low values. In order to model a tendency towards higher values, we divide the value range into stratified intervals consisting of an equal number of campaigns. In our case, we choose three different classes: low, middle, high. For further details, see Section 10.3. Now, let $class_i$ denote the class a campaign s_i is assigned to. Then the hypothesis is formulated as:

$$\bar{P}^{\alpha,\beta,\gamma}_{\text{lmh}}(s_j|s_i) = \begin{cases} \alpha \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } class_i = \text{low} \\ \beta \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } class_i = \text{middle} \\ \gamma \cdot \bar{P}_{\text{av}}(s_j|s_i), & \text{if } class_i = \text{high} \end{cases} \tag{10.10}$$

Based on this, we define three hypotheses for payment, payment per hour and positions, by specifying $\bar{P}_{\text{pay}}$, $\bar{P}_{\text{pph}}$, and $\bar{P}_{\text{pos}}$, respectively.

Time. In Section 10.3, we have also mentioned the time required to finish a task as a factor influencing the user when choosing a new task. Tasks estimated to be time intensive may deter users from choosing it [438]. Assuming normalized time values $value_i$ in a range from 0 to 1 we define the corresponding hypothesis as:

$$\bar{P}_{\text{time}}(s_j|s_i) = 1 - value_i \tag{10.11}$$

Since we were not able to derive stratified classes for the required time spans, we are not formulating this hypothesis based on intervals.

10.4.4. Title and description

The category hypothesis assumes that tasks of the same type are chosen consistently. However, this may not accurately represent the similarity of tasks required for example to capture the notion of always choosing "interesting" tasks [438]. This is especially true considering the skewed distribution of campaigns across categories (see Figure 10.1a). Thus, we further investigate this line of thought by comparing the title and the description of the campaigns instead of just their categories.

Both the title and the description can be represented as a bag-of-words. Thus, to compare titles and description, respectively, we employ the cosine distance based on TF-IDF vectors [30] (we use MLlib[3] to calculate the corresponding vectors). Note that we do not apply any other pre-processing steps like stop-word removal or stemming. Now, let $tfidf_i$ denote the TF-IDF vector of a document, i.e., either a title or the description, then we define the respective hypothesis as:

$$\bar{P}_{\cos}(s_j|s_i) = cos(tfidf_i, tfidf_j) \tag{10.12}$$

For the corresponding hypotheses for the titles and the descriptions, we write $\bar{P}_{\text{title}}(s_j|s_i)$ and $\bar{P}_{\text{desc}}(s_j|s_i)$, respectively.

10.5. Results

In order to compare the relative plausibility of the hypotheses introduced in Section 10.4, we apply the HypTrails approach as outlined in Section 3.3.2 based on the data described in Section 10.3. The results for the individual hypotheses are reported in Section 10.5.1 through 10.5.4 and visualized in Figures 10.2 and 10.3. We start by assessing the performance of the uniform hypothesis and different availability assumptions. Since we will find that the availability assumption based on overlap is the most realistic one, the following hypotheses are all grounded on availability (i.e., $\bar{P}_{\text{av}} := \bar{P}_{\text{overlap}}$). We give a summarizing comparison of our hypotheses in Section 10.5.5.

Note that, analogously to our case studies in the previous chapters (e.g., Chapters 7 and 9), we scale the concentration factor κ with regard to the number of state spaces (see Section 3.3.2.3 for details).

[3] `https://spark.apache.org/mllib/`, accessed: August 2015

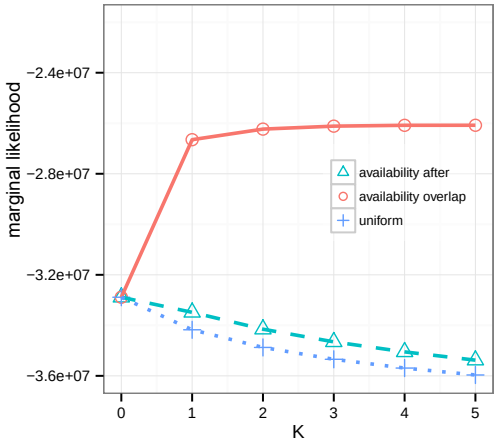


Figure 10.2.: Comparison of campaign availability models. We compare the different availability models we defined for reflecting time constraints which apply to compaigns on a crowdsourcing platform (see Section 10.4.1). We find that availability is best modeled by overlap, i.e., the user only chooses campaigns which are available upon finishing her last task.

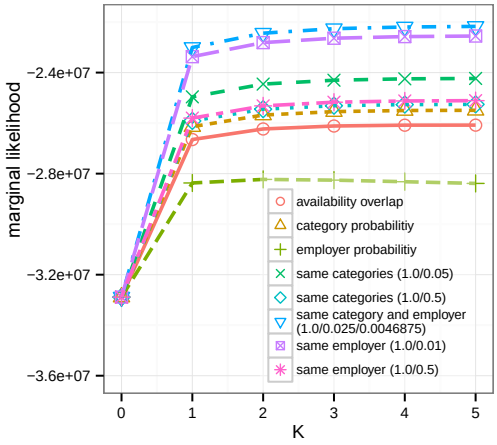
10.5.1. Uniform hypothesis and availability

Since campaigns are only available for a limited amount of time (as long as some tasks have not been completed), we have introduced several notions of availability in Section 10.4.1. In this section, we compare the corresponding availability hypotheses: $\bar{P}_{\text{uniform}}$, $\bar{P}_{\text{after}}$ and $\bar{P}_{\text{overlap}}$. Figure 10.2 shows the results. Generally, the uniform hypothesis is the most unrealistic one, since it assumes that — independent of time constraints — all campaigns are available from every other campaign. Thus, as expected, it performs worse than both the $\bar{P}_{\text{after}}$ and the $\bar{P}_{\text{overlap}}$ hypothesis. When comparing the latter two, the overlap hypothesis $\bar{P}_{\text{overlap}}$ is strongly superior. As outlined in Section 10.4.1, this is due to the fact that users choose their campaigns in quick succession. Since the overlap availability $\bar{P}_{\text{overlap}}$ is the one which is most plausible in our setting, we use it as the $\bar{P}_{\text{av}}$ component required by our other hypotheses as outlined in Section 10.4. Thus, we are looking for hypotheses which improve on $\bar{P}_{\text{overlap}}$. Consequently, $\bar{P}_{\text{overlap}}$ serves as our baseline.

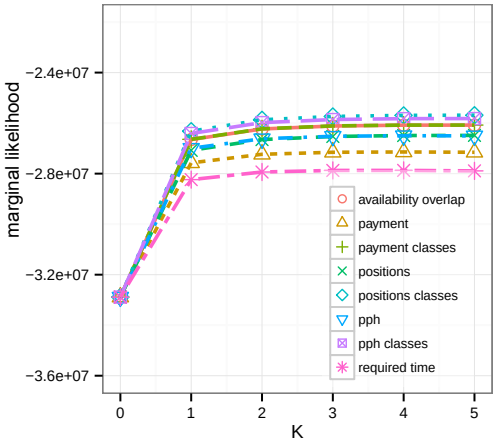
10.5.2. Category and employer

As introduced in Section 10.4.2, some straightforward hypotheses are those based on the fact that users may tend to choose campaigns from categories and/or employers which they already know. In this section, we compare different category and employer-based hypotheses. First we investigate whether users tend to pick campaigns from the same category or the same employer. Afterwards we combine both factors.

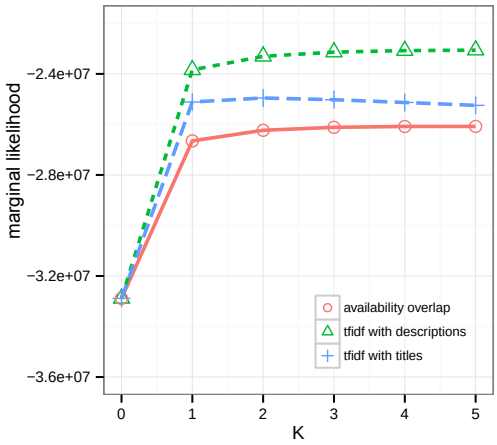
Individual results. Figure 10.3a shows the results for categories and employers separately. Independent of the respective parameters, both hypotheses are superior to the availability hypothesis indicating that users tend to stay within the same category and



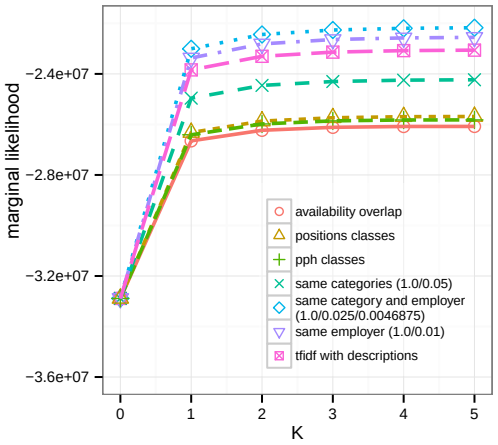
(a) Categories and employers



(b) Payment, positions and time



(c) Title and descriptions



(d) Comparison

Figure 10.3.: Comparison of hypotheses on task choosing behavior. In this figure, we show the results for different hypothesis types. (a) compares different campaign category and employer hypotheses, (b) focuses on hypotheses based on payment, available tasks per campaign (positions) as well as the time required to finish a task, and (c) depicts evidence values for description and title based hypotheses. In (d) we compare the best hypotheses of each category. We find that staying with the same employer is a strong influencing factor when choosing campaigns directly, followed by the description based hypothesis.

prefer campaigns from the same employer. For both hypotheses we initially set parameters so that moving to the same category or employer is two times as likely as moving to a campaign from another category or employer ($\alpha = 1, \beta = 0.5$). We experimented with different parameter values and found that further favoring the same categories/employers increases the plausibility of the hypothesis up to a certain point. In Figure 10.3a, we also show the best parameter settings we have found. The results indicate that users strongly favor the same categories and employers. We further note that staying with the same employer is a more plausible hypothesis, i.e., because we find greater evidence for it and also because a stronger focus on the same employer can be set ($\beta = 0.01$) before the evidence is decreasing again. This is in line with the findings of Schulze et al. [438], who report a tendency to choose campaigns from reputable employers.

Mixture. Since both (hypotheses based on categories and hypotheses based on employers) perform well, we investigate if a combination of both, as mentioned in Section 10.4.2, can further improve our results. And, indeed, we find that strongly favoring campaigns from the same category and the same employer ($\alpha = 1$), reducing the weights for choosing either the same campaign or the same employer ($\beta = 0.025$), and setting the weight for totally different campaigns rather low ($\gamma = 0.0046825$), results in the most plausible hypothesis so far. We have also tried setting different parameters for only favoring one of the two, category or employer (by using β, β'), which only resulted in marginal improvements.

Skewed probabilities. As mentioned in Section 10.3, the overall probabilities of campaigns within categories and from employers is not a uniform (cf. Figure 10.1a). Thus, we defined hypotheses based solely on employer and category frequencies in Section 10.4.2 ($\bar{P}_{\text{cat-freq}}, \bar{P}_{\text{emp-freq}}$) to contrast the category and employer hypotheses evaluated above. We find that, both hypotheses perform worse than the hypotheses which stick to the same category or employer. This shows that the corresponding, previously observed tendency does not stem from skewed category or employer frequencies.

Summary. Overall, we can conclude that the idea that users choose campaigns from the same category and the same employer can indeed explain parts of the campaign transitions we observe. Particularly, the employer is an important factor.

10.5.3. Payment, position, and time

As introduced in Section 10.4.3, we also consider several hypotheses based on payment, available positions per campaign, and the time required to complete a task. To this end, we directly use the values of payment, payment per hour (pph), positions and the inverse of the required time. The results are shown in Figure 10.3b. All of these hypotheses perform badly compared to the overlap hypothesis. This is due to their skewed value distributions which have a strong tendency towards small value ranges, cf. Section 10.3. Thus, we introduced three stratified classes for payment, pph and required time in order to model a tendency to pick low, average or high prices. We weight these classes as follows: $low = 1$, $average = 2$ and $high = 3$. We observe, that all the resulting hypotheses, even if marginal, have a greater evidence than the uniform hypothesis on overlapping availabilities (while the payment classes are hardly distinguishable in the graph, evidence

values actually do differ significantly from those of the overlapping availability hypothesis). Thus, confirming the findings of Chilton et al. [113], Schulze et al. [438], and Yuen et al. [549], we conclude that all of these factors play a role in choosing campaigns. However, we were not able to derive a hypothesis based on the "required time for a task" explaining the corresponding influence factor found by Schulze et al. [438].

10.5.4. Title and description

As introduced in Section 10.4.4, we also compare similarities between the title and the descriptions of the campaigns. The results are shown in Figure 10.3c. Both hypotheses clearly perform better than the baseline hypothesis assuming a uniform distribution over all overlapping campaigns ($\bar{P}_{\text{overlap}}$). This is a strong indicator that both, the title and the description, and thus the semantic content of the task to complete, are a decisive factor for choosing campaigns. Note though, that the similarity based on the title clearly performs worse than the similarity based on descriptions. This can be interpreted by the fact that while a certain type of campaign will have the same description, the title may differ. For example, consider a campaign whose goal is to annotate a certain corpus of documents. While the description might be the same, the title for one campaign might be "Annotating Literature", while the other might be "Annotate 10 Research Papers". The description based hypothesis captures the similarity of those two campaigns, while the title based hypothesis does not.

10.5.5. Summary

For an overall comparison, we show the best hypotheses of each category in Figure 10.3d. We observe that three hypotheses yield especially high marginal likelihood values, that is, the mixture of category and employer performed best, followed by the hypothesis solely based on employers, and the hypothesis based on description similarities. The hypotheses based on payment and positions hardly improve on the uniform overlap hypothesis. The hypothesis based solely on categories performs worse than the description based hypothesis but better than the hypotheses based on payment and positions.

Categories and employers. The fact that the category hypothesis performs significantly worse than the employer hypothesis, and that combining the employer hypothesis with categories only marginally improves the evidence, is an indicator that users primarily choose campaigns from the same employer instead of focusing on categories. Yet when they have chosen their employer they prefer to stay within the same campaign category. Overall, the large evidence values for employer-based hypotheses are in line with results found by Schulze et al. [438]. They imply that workers are loyal to reputable employers when choosing campaigns: since employers rate the completed tasks by their quality, and only good ratings result in money being payed, workers prefer employers who rate fairly.

Payment, position, and time. The result that the payment and positions hypotheses perform badly seems counter-intuitive. This might be due to the fact that users optimize for certain types of tasks and can earn more money when they stay at the same employer and within the same category accounting for the high plausibility of the corresponding

hypotheses. Also, Aris [17] finds that payment in general is not a consistent factor influencing how users choose their tasks. Schulze et al. [438] and Schnitzer et al. [437] even find that at least US workers (which we have focused on here) are not mainly interested in the amount of money they earn for completing a campaign. However, the bad performance may also be due to the way we model the corresponding hypotheses. For example, we have noticed that directly incorporating the payment value into the hypothesis does not perform well (see Section 10.5.3). Now, while the stratified classes reveal a certain tendency to choose well payed campaigns or campaigns with many positions, the resulting increase in evidence is not as large as could be expected. Thus, we might not capture the influence of payments or positions correctly. That is, different classes or different weighting strategies might result in better hypotheses. Also, the payment may strongly interact with other factors like the tendency to choose similar tasks as mentioned before. Regarding these issues, we will propose ideas for further research in the discussion section.

Title and description. Finally, the good performance of the description based hypothesis is an indicator that users not only choose from within the same campaign, but also try to choose similar campaigns with regard to the actual task they have to work on. One reason for this might be that familiar tasks are easier and more quickly to handle than unknown ones. This may also be a result of users choosing similar tasks based on their area of interest as suggested by Aris [17]. Note that the description hypothesis might strongly overlap with the employer hypothesis since the same employer may often use the same description for her campaigns of a similar type. This needs to be investigated further.

Overall. We studied a considerable amount of hypotheses partially motivated by related work using methods solely based on data from the crowdsourcing platform Microworkers and without resorting to error-prone and possibly biased user studies. In the process, we focused on US workers and were able to show from observed task transition data that most factors found in literature indeed influence the process of how workers choose campaigns. Employer and description based hypotheses worked best. Whereas hypotheses based on payment only showed a marginal influence on how users choose their tasks.

10.6. Discussion

We have tested several hypotheses about how users choose their campaigns on the crowdsourcing platform Microworkers. In this section, we give a short overview of particular limitations of our approach and propose possible future work.

Data. First of all, the dataset we are using is limited to workers from the US. This is because users from the US are free to choose nearly all campaigns. For non-US workers many campaigns are not available. Thus, incorporating non-US workers would introduce restrictions on transitions which can not be directly modeled using the original HypTrails framework [453]. Additionally, for example, Schulze et al. [438] and Schnitzer et al. [437] suggest that there may be strong differences between certain user groups, e.g., from different countries. Thus, in further studies it might be useful to actively incorporate different user groups, e.g., by applying MixedTrails (cf. Chapter 4).

Furthermore, we are evaluating our hypotheses on only one dataset. It would be interesting to check if the hypotheses behave similarly on different crowdsourcing platforms.

Hypotheses. While we have studied quite a few different hypotheses, there are more to consider. For example, it might be interesting to study if users tend to prefer recently created campaigns as suggested by Chilton et al. [113]. Also, Aris [17] implies that intrinsic factors are more important than extrinsic ones. In this work, we have mainly focused on extrinsic ones.

Furthermore, we have only combined the category with the employer hypothesis. Other combinations might yield better results. Also, we have not checked how the hypotheses are related to each other in a sense that the features used to build them are correlated. An example would be that the same employer will often use similar descriptions for her campaigns. Thus, as mentioned in Section 10.5.5, the description hypothesis might be correlated to the employer hypothesis. This needs further investigation.

Finally, the payment, positions and time related hypotheses did not yield good results when compared to category or employer based hypotheses. While there are explanations in literature [17, 438], this may also be due to a poor understanding of how these factors influence the choice of campaigns. We have approximated the influence using three stratified classes. Other approaches might be more appropriate.

Availability. One limitation of HypTrails is that it is not built to model states that are only available at certain time intervals. We solved this by introducing the notion of availability. In our scenario, we used local availability (from a specific campaign to other campaigns) based on time intervals. However, this approach is an approximation. Further research may find a better solution in corporate such time-dependent availability into HypTrails and/or in the process of formulating hypotheses.

10.7. Related work

The motivation of working on crowdsourcing platforms in general and the preferences of selecting tasks are the subject of several studies. Most of this research is based on user surveys leading to varying answers depending on the way questions are asked, and consequently limiting the understanding of the respective influence factors [438]. However, the influence factors derived form such studies can be used for formulating hypotheses about how users choose their tasks, which we objectively evaluated in this case study.

Aris [17] reviewed research results of motivational factors of participation in the area of mobile crowdsourcing. In contrast to the Microworkers platform investigated in this study, the platforms and services analyzed by Aris were from the field of creative tasks: for example, participating in innovation contests, generating news content, or even more specialized social tasks like assisting foreign visitors in Japan. The main influence factor in the reviewed studies was found to be "personal benefit", which can be categorized into intrinsic and extrinsic motivation. Intrinsic motivation is given if a task is fun, a new experience is gained, or because it is challenging, whereas extrinsic motivation describes participation based on awards like points or a monetary reward. Overall, Aris found that intrinsic aspects are more important than extrinsic aspects. Furthermore, Aris recognized

that the results about monetary rewards were not consistent. It can be assumed that — similar to the case of mobile crowdsourcing — users on micro tasking platforms (like the Microworkers platform we study in this case study) are also affected by intrinsic and extrinsic factors.

Indeed, a model for the workers' motivation by Kaufmann et al. [275] confirmed the importance of intrinsic aspects on Amazon Mechanical Turk (MTurk). At the same time, extrinsic factors have were found to be relevant. Such factors include task related factors as well as motivation based on learning and training skills. Regarding extrinsic factors, Chilton et al. [113] also found that task related properties and characteristics, like the creation date or the overall number of tasks, provided by a campaign, influence the selection of tasks. The results were based on the analysis of data scraped from MTurk and a survey about the workers' task searching behavior.

In contrast to Aris and Kaufmann et al., the user study of Yuen et al. [549] showed that a high monetary reward is the most important task selection criterion. In addition, the workers answered that they chose their work based on the nature and the difficulty of the different tasks.

Finally, Schulze et al. [438] observed that the preferences and influence factors differ with respect to the location of the workers: For workers from the United States, the most important aspect for selecting a task was their interest in it. This was followed by payment, the simplicity of tasks, and a high reputation of the employer. In contrast, Indian workers preferred well payed and simple tasks.

Schnitzer et al. [437] confirmed this observation by a user study about worker demands on task recommendation. Here, the similarity of tasks is the most important task property for workers from the United States, whereas Asian and European workers were mostly interested in tasks offering the highest monetary reward.

Overall, there are many factors influencing the selection of new tasks in crowdsourcing environments. Some results are even contradictory. However, in contrast to this case study, none of the cited papers conducted an objective comparison of the proposed factors.

10.8. Conclusion

We studied how users choose their next task on the crowdsourcing platform Microworkers. To this end, we formulated different hypotheses about the underlying processes based on properties like the similarity of campaign descriptions, categories, employers, or payment information. In the process, we developed an approach to cope with temporal aspects with regard to availability of states in the underlying Markov chain, i.e., campaigns in our crowdsourcing environment. Then, utilizing campaign transition data from Microworkers, we objectively compared the resulting hypotheses by means of the Bayesian approach HypTrails. While the results highly depended on how hypotheses are formulated, in our scenario, combinations of category and employer as well as the description-based hypothesis worked best. Overall, instead of using survey-based investigations — as similar studies do — we successfully applied the Bayesian method HypTrails to objectively compare hypotheses about how users choose their next campaign solely on data already

available from the crowdsourcing platform. This is a significant step forward in providing crowdsourcing companies with the means to gauge the preferences and the behavior of their users in order to optimize their platforms, e.g., by incorporating the knowledge gained about these factors into task recommendation systems.

11. Small scale case studies

In this section, we study three more domains where human navigation behavior can be observed, i.e., geo-spatial navigation when exploring urban noise pollution, navigation on Wikipedia pages, and listening behavior on the last.fm music service. These are small scale studies aimed at providing further insights into human navigation, as well as at illustrating the application of MixedTrails (Chapter 4) and SubTrails (Chapter 5).

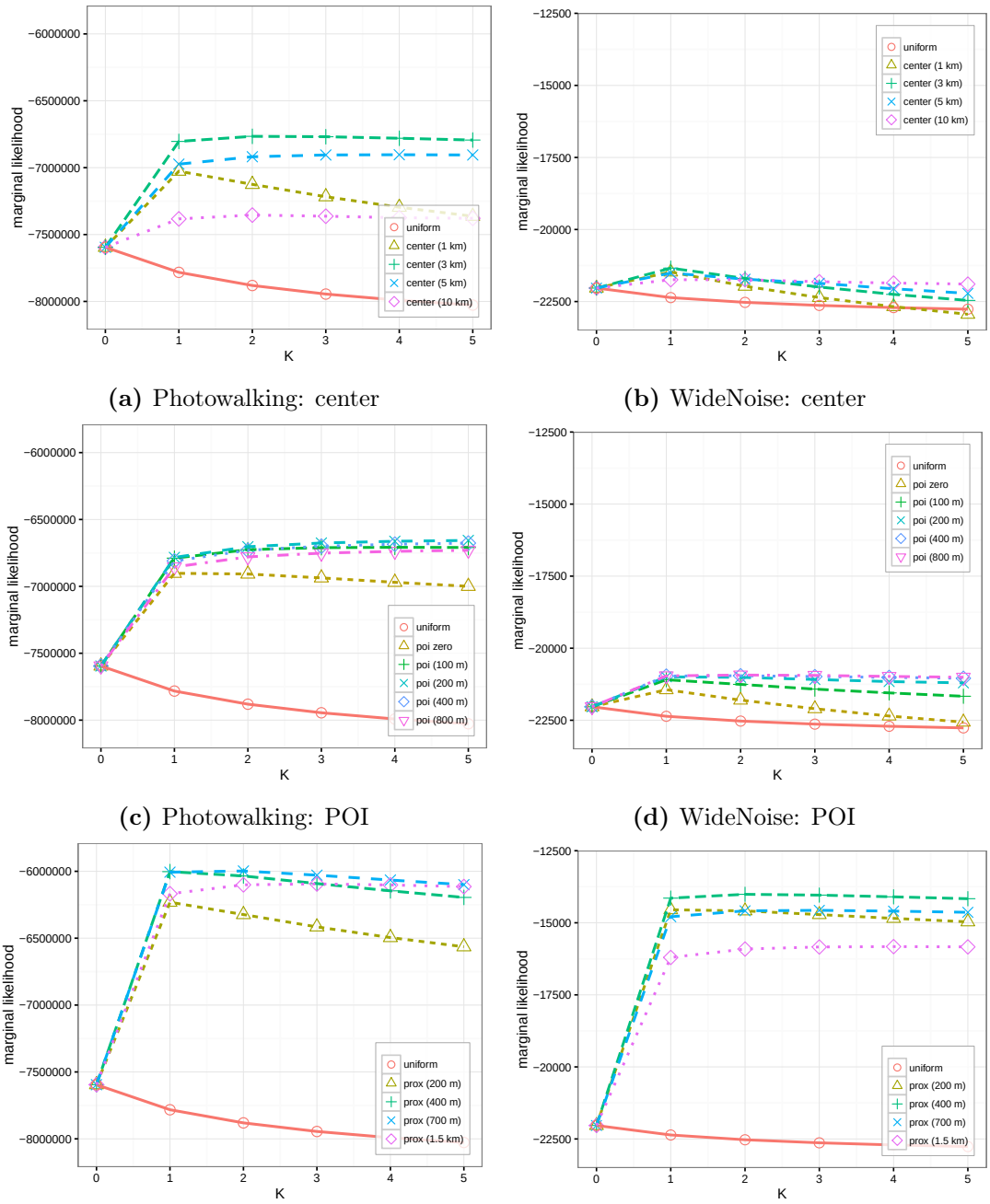
11.1. Noise pollution exploration

In this small scale case study, we explore the difference of navigation processes in different application scenarios. In particular, we compare the extensively studied photowalking behavior of Flickr users as covered in Chapter 7 with the navigational processes in the context of exploring noise pollution. To this end we focus on the city of London and find fundamental differences in behavioral characteristics.

Data. For the data on photowalking we use the same data as in Chapter 7. For the noise pollution measurements, we employ data from the WideNoise application of the EveryAware platform (cf. Section 6.3.3) which consists of overall 62,216 measurements collected from 08.11.2011 to 15.08.2017 by 18,350 devices. We limit these measurements to a bounding box around London as already mentioned in Table 7.1 leaving 2,798 measurements. Similar to the photos used for the photowalking case study in Chapter 7, we group these measurements by device id[1]) and sort them according to the time the samples have been recorded. This results in 352 noise trails with an average length of 6.95.

Experimental setup. Overall, the experimental setup is identical to the photowalking case study in Section 7.4.1, i.e., we employ a grid of 200m by 200m on a bounding around London (cf. Table 7.1). With regard to hypotheses, we use three major hypothesis classes introduced in Section 7.3, namely the center, the POI (points of interest), and the proximity hypothesis. The center hypothesis represents the belief that the next noise measurement will be taken close to the city center independent of the current location. The POI hypothesis represents the belief that the next noise measurement will be taken in an area with many points of interest, again independent of the current location. And the proximity hypothesis represents the belief that the next noise measurement will be taken somewhere close to the last measurement according to a Gaussian spreading function. For details, please see Section 7.3. We compare each hypothesis class using several different parameter settings (as in the Berlin case study in Section 7.4.1.1 but focusing on London).

[1]Device ids are always set, while user ids only exist when the user is logged in.



(a) Photowalking: center

(b) WideNoise: center

(c) Photowalking: POI

(d) WideNoise: POI

(e) Photowalking: proximity

(f) WideNoise: proximity

Figure 11.1.: Comparison of photowalking behavior and noise exploration in London. This figure visualizes the results for several navigational hypotheses in London based on photowalking behavior (left) and noise exploration (right). We visualize the center, the POI, and the proximity hypothesis analogously to Chapter 7. For the noise exploration behavior, the center and the POI hypothesis exhibit a stronger tendency towards larger spreading factors while the proximity hypothesis prefers smaller spreading factors. This indicates fundamental differences of navigational behavior in the different contexts.

Results. We compare photowalking behavior (PB) and noise exploration (NE) in London based on the center, POI, and proximity hypotheses introduced above.

For the center hypothesis the various parameter settings perform differently for PB and NE. In particular, the center hypothesis with a spread of 3km performs best for PB while a spread of 10km shows the best marginal likelihoods for NE with increasing concentration factors (κ). In combination with the fact that, for NE, the center hypotheses do not outperform the uniform hypothesis as clearly as for PB, this points to a less prominent tendency to stick to the city center for people who explore the noise environment of London. A similar phenomenon can be observed when comparing the marginal likelihoods of different parameterizations of the POI hypothesis: For NE, the POI hypotheses can not improve on the uniform hypothesis as strongly as in the case of PB, and, generally, larger spreads are favored for NE than for PB.

This tendency of NE for larger spreads for both hypotheses (center and POI), can most likely be explained by the inherent procedural differences of taking photos of a city and using WideNoise to explore noise. Photos are mostly taken by visitors and tourists, thus, concentrating on the city center and points of interest (POI). Similarly, locals also often focus their intereset on the same areas. This explains the comparably strong performance of both hypotheses in the case of PB. In contrast, when exploring noise pollution, people are more interested in their personal environment or hotspots relevant to them. Consequently, this is mostly focused on areas other than the city center or areas with POIs, which explains the tendency to favor larger spreads and the less prominent improvements of the informed hypotheses (center, POI, proximity) on the uniform hypothesis for NE.

The center and the POI hypotheses already clearly point towards significant differences of navigation behavior in the context of PB and NE. The proximity hypothesis further supports this observation. In particular, in comparison to PB, small spreads are strongly favored and specifically the hypothesis with a spreading factor of 400m performs exceptionally well for NE. Nevertheless, overall, the marginal likelihoods for the proximity hypotheses are the highest of the compared hypothesis classes. This indicates that the process of NE, generally, favors a tendency of users to stay close to a current location before taking the next measurement.

Thus, we observe that while PB and NE have inherently different navigational characteristics — with NE focusing on more narrow spreading factors — the overall tendency of humans to stick to close by locations is a stable concept.

Conclusion. In this case study, we showed that different navigational contexts exhibit fundamental differences with regard to several principles of human mobility. In particular, we observed that while a tendency to move towards the city center or POIs was an adequate explanation of photowalking behavior, those hypotheses were less viable for noise exploration processes. In contrast, a proximity component can be considered to be equally applicable to both scenarios even if with differing parameterizations (Gaussian spreading radius).

Note that this small scale case study is preliminary work due to several factors including the limited amount of data used to study noise exploration processes or the small number

of compared hypotheses and applied methods compared to Chapter 7. Even though, the observed results still give rise to the question if varying navigational contexts exhibit comparable differences and similarities. Thus, we aim to extend this study and encourage future work in this direction.

11.2. Exploration and homing-in phases on Wikipedia

Wikispeedia [519] is a game in which players aim to find the shortest path from a randomly given start article to a randomly given target article within Wikipedia by only navigating the available hyperlinks. In the context of this game, West and Leskovec [519] have hypothesized that "humans navigate more strongly according to degree in the early game phase, when finding a good hub is important [in order to be able to increase the amount of reachable concepts], and more strongly according to textual similarity later on, in the homing-in phase [when trying to find the actual target concept]". Here, we confirm this hypothesis using MixedTrails.

Data. Wikispeedia is based on a subset of 4,600 Wikipedia articles (from the 4,600-article CD version of "Wikipedia for Schools"[2]). A corresponding dataset [520] is freely available[3]. It consists of the plain text of each article, the link network, and a set of click sequences (including back clicks) created by participants playing the game. Like West and Leskovec [519], we remove back clicks (but keep the corresponding forward clicks which are undone by these back clicks) and then only keep click sequences of length 3 to 8 (number of clicks). The resulting dataset consists of over 25,000 click sequences with an average length of 5.6.

Hypotheses. To investigate the hypothesis by West and Leskovec [519], we consider two transition probability matrices: $\boldsymbol{\phi}_{\text{deg}}$ represents the hypothesis that people are trying to get to hubs in order to increase the number of concepts they can reach. Thus, if a link between a source article to a destination article exists, we set the belief in the corresponding transition proportional to the degree of the destination state (calculated as the sum of its in- and out-going links); and zero otherwise. Second, the transition probability matrix $\boldsymbol{\phi}_{\text{sim}}$ assumes a higher transition probability if there is a strong textual similarity between two articles. Again, we set the transition probability to 0 if there is no link between two articles. Otherwise, we set the belief in a transition proportional to the cosine similarity $cos(i, j)$ with respect to the corresponding *TF-IDF* vectors. For that, we removed words with a document frequency of over 80% and applied sub-linear scaling to the TF values.[4] For comparison, we additionally consider the link matrix $\boldsymbol{\phi}_{\text{link}}$ that expresses equal belief in all transitions for which a link exists.

Now, the first three hypotheses are homogeneous hypotheses assigning transitions to a single group: $H_{\text{link}} = (\boldsymbol{\gamma}_{\text{one}}, \boldsymbol{\phi}_{\text{link}})$, $H_{\text{deg}} = (\boldsymbol{\gamma}_{\text{one}}, \boldsymbol{\phi}_{\text{deg}})$, $H_{\text{sim}} = (\boldsymbol{\gamma}_{\text{one}}, \boldsymbol{\phi}_{\text{sim}})$. Furthermore,

[2] available at `schools-wikipedia.org` (version of 2007)

[3] `https://snap.stanford.edu/data/wikispeedia.html`

[4] Differing from our approach, West et al. [520] use the similarity between the clicked article and the target concept $cos(i, t)$, but report that along the game progress, the similarity of the current and the clicked/next article is qualitatively similar. Thus, we use the latter approach since we can only use information from already visited states, not future states.

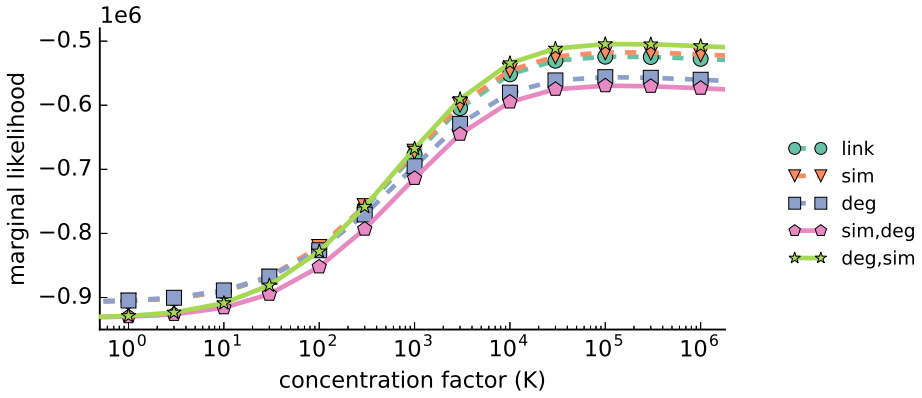


Figure 11.2.: Comparison of hypotheses on search trails on Wikispeedia. For the game Wikispeedia, players try to quickly navigate from one article to another using the underlying link structure of Wikipedia. One hypothesis (*deg,sim*) is, that to achieve this, players will first navigate to articles with a large degree, and then "home-in" on their target using similarity based navigation. The graph shows the results of modeling this heterogeneous hypothesis using the MixedTrails approach by splitting each click sequence at their second click. We also compare against several other homogeneous as well as heterogeneous hypotheses. Overall, of all the considered hypotheses, the heterogeneous *deg,sim*-hypothesis works best (for growing concentration factors), even though the initial split (low concentration factors) is not inherently advantageous. Note that, while the differences may seem marginal, they are decisive.

$H_{\text{deg,sim}}$ and $H_{\text{sim,deg}}$ are heterogeneous hypotheses that group transitions based on their position on the trail of articles left by users playing the game: the first two transitions are assigned to the "initial phase", and the rest of the transitions are assigned to the "homing-in phase". We name the corresponding group assignment probabilities $\boldsymbol{\gamma}_{\text{phases}}$. The heterogeneous hypotheses are then defined as: $H_{\text{deg,sim}} = (\boldsymbol{\gamma}_{\text{phases}}, (\boldsymbol{\phi}_{\text{deg}}, \boldsymbol{\phi}_{\text{sim}}))$ and $H_{\text{sim,deg}} = (\boldsymbol{\gamma}_{\text{phases}}, (\boldsymbol{\phi}_{\text{sim}}, \boldsymbol{\phi}_{\text{deg}}))$ assuming the degree and the similarity transition probability matrices to explain the "initial phase", respectively.

Results. Figure 11.2 shows that, as literature hypothesized, the heterogeneous hypothesis $H_{\text{deg,sim}}$ explains the navigational behavior of players better than all other considered hypotheses. While the additional variables introduced by the split (by means of Occam's Razor) result in lower marginal likelihoods compared to the homogeneous hypotheses for weak believes (low values of the parameter κ), it becomes apparent that the transition probability matrices of $H_{\text{deg,sim}}$ are modeling the corresponding movement behavior in each group better than the single transition probability matrix of the homogeneous hypotheses on the overall data. At the same time, the "opposite" hypothesis $H_{\text{sim,deg}}$ results in the lowest ML values even though it uses the same split as $H_{\text{deg,sim}}$. Among the homogeneous hypotheses, the similarity based hypothesis is the most plausible. By contrast, as it yields rather low marginal likelihood values, the degree hypothesis H_{deg} seems to be a very specialized hypothesis, which is applicable only for a specific subset of transitions—such as the first transitions in each sequence.

Conclusion. Overall, this example supports the claim by West and Leskovec [519] that there is an exploration and a homing-in phase when users search specific articles on Wikipedia; at least within the game environment of Wikispeedia. Furthermore, our results demonstrated the applicability of MixedTrails to a real world scenario. That is, we saw that a more fine-grained, heterogeneous hypothesis may explain the observed sequential data better than using a single, overly general, homogeneous hypothesis.

11.3. Exceptional listening behavior on the last.fm music service

In this case study, we analyze data from the last.fm music service. In particular we interpret playlists as trails over songs and use our method for finding subgroups with exceptional transition behavior (Chapter 5) to discovery subsets of the data with interesting music listening behavior.

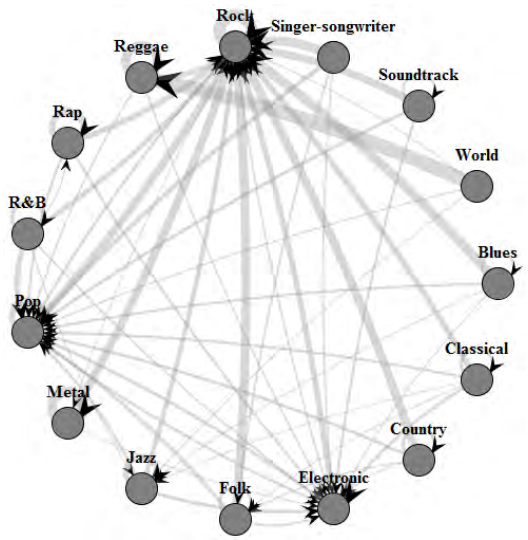
Data. We use the *1K listening data*[5] containing the full listening history of $1,000$ last.fm users featuring more than $19,000,000$ tracks (songs) by more than $170,000$ artists. With this data, we study sequences of music genres (such as *rock*, *pop*, *rap*, *classical*, etc.) of songs that users listened to, focusing on a list of 16 main genres. Since genre information is difficult to obtain on a track-level, we labeled each track with the best fitting genre for the respective artist as obtained by the EchoNest API[6]. In doing so, we could determine genres for more than 95% of the tracks. We then constructed genre transitions for each user based on the sequence of tracks she had listened to. We filtered subsequent tracks of the same artist to remove cases where the user listened to all songs of a single album. Additionally, we removed all transitions with unknown source or target state (genre). Thus, we obtained a dataset of $9,020,396$ transitions between tracks. Background knowledge includes user information about age, gender, origin and the year of sign-up to last.fm, and the point in time the source song of the transition was played, i.e., the hour of the day, the weekday, the month and the year.

Experimental setup. On the data described above, we applied our approach twice to find subgroups with exceptional transition behavior: once for subgroups described by a single selection condition only (e.g., Country=US), and once including combinations of two selection conditions (search with depth 2, e.g., # Tracks $>=$79277 $\wedge$ signup=2005).

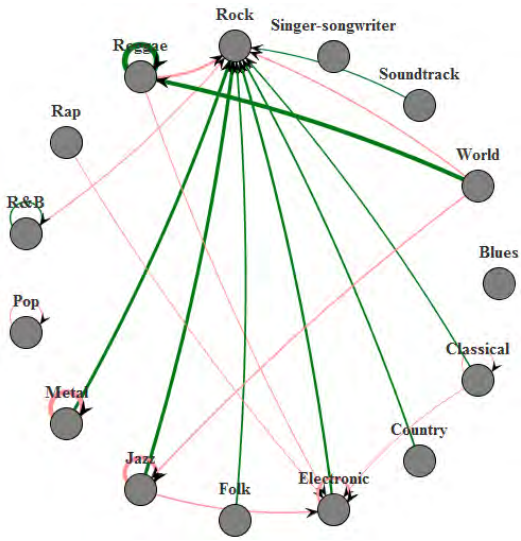
As selection expressions for subgroup descriptions, we used all attribute-value pairs for nominal attributes, and all intervals obtained by equal-frequency discretization into five groups for numeric attributes. This allowed to generate 86 selection conditions.

For computing the interestingness measure, we used $r = 1,000$ random samples. We confirmed our top results to be statistically significant on an $\alpha = 0.01$ level using the procedure presented in Section 5.2.2.

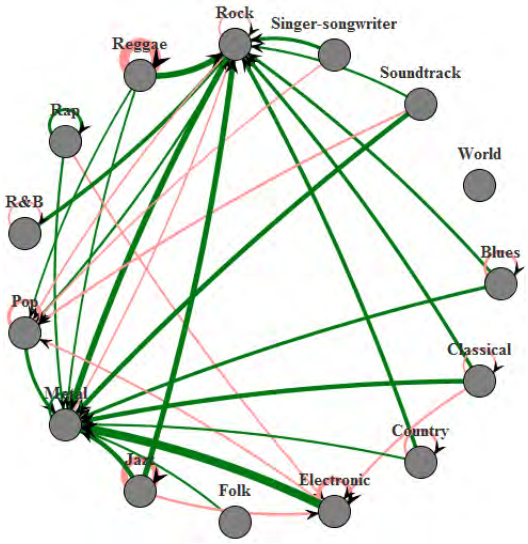
[5] http://ocelma.net/MusicRecommendationDataset/index.html
[6] http://developer.echonest.com/



(a) All transitions



(b) Users from the United States



(c) Users from Finland

Figure 11.3.: Exceptional transition models of last.fm users. The figures show transitions between music genres: stronger arrows represent higher transition probabilities. (a) shows all transitions in the data, (b) and (c) illustrate the differences of transition models in two exceptional subgroups. Green arrows indicate that transitions are more probable in the subgroup than in the overall dataset, red arrows the contrary. For instance, it can be observed in (b) that users from the US are more likely to listen to *Reggae* after *World* music; (c) shows that Finnish users have higher transition probabilities to *Metal*. Insignificant differences are removed for visibility.

Table 11.1.: Top subgroups for the last.fm dataset. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv}, and the unweighted total variation Δ_{tv}.

(a) Single selection conditions only.

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Country = US	2,576,652	420.37 ± 9.67	326,435	1.44
Country = Finland	384,214	408.37 ± 8.89	132,378	3.05
Country = Argentina	174,140	360.22 ± 8.62	84,285	4.54
# Tracks > 79277	1,803,363	355.27 ± 8.39	249,634	1.61
Country = Poland	378,003	346.09 ± 7.37	122,155	3.35

(b) Including combinations of selection conditions.

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Tracks >=79277 ∧ signup=2005	617,245	425.59 ± 9.30	186,048	3.38
# Tracks >=79277 ∧ age = [23–24]	155,998	421.67 ± 9.78	89,769	4.67
Country = US	2,576,652	420.37 ± 9.67	326,435	1.44
Country = Finland	384,214	408.37 ± 8.89	132,378	3.05
# Tracks >=79277 ∧ signup=2006	658,135	398.40 ± 8.74	182,690	3.38

Results. Results for single selection conditions are displayed in Table 11.2a. We can see that the country of origin of users is an important factor: the majority of top subgroups is described by this attribute. Specifically, users from the United States, from Finland, and from Argentina exhibit transitions between music genres that are unusual compared to the entire data. By contrast, date and time influence the transitions between genres only little and do not describe any of the top subgroups. For subgroups described by combinations of conditions, see Table 11.2b, we can see that users with a high number of tracks show unusual transition behavior, especially if they signed up to the system early, or if they are in a certain age group.

Figure 11.3 visualizes differences in transition behavior in comparison to the overall dataset for two top-subgroups. Here, each node represents a state (genre). The first graph gives an impression on the transition probabilities in the entire dataset. Stronger arrows represent higher probabilities. We omit probabilities below 0.1. The next two graphs show *deviations* from these probabilities in the subgroups *Country=US* and *Country=Finland*. Green arrows indicate transitions that are more likely in the subgroup than in the overall data, red arrows imply less likely transitions. Stronger arrows represent higher deviations; small deviations (< 0.05) are omitted.

We observe that users from the US and Finland deviate from the overall behavior in characteristic ways. For example, users from the US tend to skip to *Rock* more often, while the same is true for users from Finland with regard to *Metal*. Also, we observe interesting dynamics between genres that go beyond the different target state distributions: for example, users from the US are more likely to listen to *Reggae* after a song from the *World* genre, while the preference for *Rock* decreases in this case. We can also see, that although *Rock* is overall more popular in the US, it follows a track of *Reggae* less likely than in the entire data.

Conclusion. In this case study, we showed the heterogeneous nature of music listening behavior. In particular, using SubTrails (cf. Chapter 5), we were able to identify typical stereotypes like the preference of people from Finland for the Metal genre in their playlists (cf. Purhonen et al. [407]). Besides the results from Section 7.4.2, this demonstrates the applicability of SubTrails on real-world data.

12. Conclusion

In this thesis, we contributed to understanding human navigation by proposing novel methods for explaining its underlying processes as well as by conducting a wide variety of case studies. In particular, we addressed the challenges of hypothesis *comparison* and *conception* with an explicit focus on the inherently heterogeneous nature of human behavior (cf. Section 1.2). For this, we introduced novel methodology and supplemented current literature by studying human navigation behavior in geo-spatial contexts as well as on the web. In the following, we briefly summarize our contributions and conclude with an outlook on future work.

12.1. Summary

This thesis has three main parts. In Part I, we covered basic information including related studies in the field of human navigation analysis as well as the methodological foundations we use and extend throughout this work. In Part II, we introduced our main contributions, i.e., novel methods for analyzing human navigation behavior explicitly focusing on the heterogeneity of navigational processes. And finally, in Part III, we applied existing and novel methodology to a diverse field of different application scenarios. In the following, we briefly sum up each part.

Part I: Background. The background consists of two chapters, i.e., Chapters 2 and 3:

In Chapter 2, we reviewed a broad spectrum of studies concerning human navigation behavior in the geo-spatial context as well as on the web. We covered early work, modeling aspects, as well as regularities and patterns. Furthermore, we specifically targeted heterogeneity inherent to human navigation behavior since our methodological contributions are centered around this aspect.

Chapter 3 introduced the methodological foundations which we use throughout this work. This included a short introduction of the framework of discrete navigation behavior in combination with background information which we use as the foundation of this thesis. Afterwards, we covered several other methods and approaches we applied and extended including Markov chains, exceptional model mining, as well as hypothesis comparison using HypTrails. The latter was especially emphasized, since we extended it in Chapter 4 and applied it heavily throughout our case studies in Part III.

Section 1.3.1: Novel methods. On the methodological level, we introduced several novel methods to study and understand human navigation behavior especially focusing on its inherent heterogeneity. In particular, we described MixedTrails for *formulating and comparing heterogeneous hypotheses* (cf. Section 1.2.1) as well as SubTrails for discovering

subgroups with exceptional transition behavior supporting the process of *hypothesis conception* (cf. Section 1.2.2).

Chapter 4 covers MixedTrails which we developed for understanding the heterogeneous processes of human navigation. It extends existing methodology by allowing the formulation and comparison of *heterogeneous* hypotheses about transition processes in sequential data based on a Bayesian approach. MixedTrails provides a straightforward way to formulate such hypotheses and, thus, enables the comparison of a very flexible set of ideas and intuitions by incorporating a large variety of background information.

Our second contribution to understand heterogeneity of human navigation, called SubTrails (Chapter 5), supports the processes of conceiving new hypotheses for understanding navigation: It provides the methodology for automatically discovering subgroups with exceptional transition behavior. In contrast to MixedTrails, where holistic hypotheses about heterogeneous sequence data are formulated and compared, SubTrails aims at finding subgroups which differ from the overall dataset with regard to how users transition between different states. By using background data to build and describe such subgroups, the found patterns are easily interpretable. While not leading to overall explanations of the observed data, SubTrails gives various insights into the underlying heterogeneous processes and can stimulate new ideas for more general hypotheses.

In addition to MixedTrails and SubTrails, Part II also introduced several analysis tools useful for the comparison as well as the conception of hypotheses about human navigation (see Chapter 6). This includes SparkTrails (Section 6.1), a distributed implementation of the HypTrails approach, the VizTrails toolbox (Section 6.2), for visualizing transition behavior in geo-spatial contexts, as well as the EveryAware platform (Section 6.3), a system to actively support participatory sensing campaigns by providing functionality to collect, explore, and analyze sensory as well as subjective data.

Part III: Case studies. In Part III, we extensively applied HypTrails [453] (cf. Section 3.3.2) as well as our own novel methodology to a variety of application domains. This includes studies on geo-spatial data, i.e., photo trails from Flickr, as well as a participatory sensing campaign, navigation on social tagging systems, the process of choosing campaigns on a crowdsourcing platform, and more. In the process, in line with the challenges formulated in Section 1.2, we developed approaches for coping with application specific subtleties like continuous observations or temporal constraints and study the inherent heterogeneous nature of human behavior.

On the geo-spatial photo trails studied on Flickr (Chapter 7), we found that proximity is a good explanation for human behavior across several cities. In particular, humans seem to prefer to consecutively take photos at proximate POIs that are popular on Wikipedia. We furthermore analyzed the same data using SubTrails and MixedTrails in an integrated experiment finding and modeling subgroups with exceptional behavior, e.g., tourists and locals. In the processes of this study, we had to cope with the continuous nature of geo-spatial navigation.

Similar to the photowalking experiments, we also investigated geo-spatial human behavior in Chapter 8, where we focused on data from a participatory sensing campaign collected using the EveryAware platform from Section 6.3. We found characteristic

explorative navigation processes, based on the different phases of the campaign. Here, in addition to continuous observations, the temporal density of measurements had to be taken into consideration in order to formulate sensible hypotheses.

In Chapter 9, we switched from the geo-spatial domain to studying human navigation on the web. In particular, we analyzed navigation behavior of users in a social tagging system and found semantic components in the underlying processes of the studied subsets of the user population. While different genders did not exhibit significant behavioral deviations, short-term users, as well as different tagging types, followed certain behavioral patterns matching their individual characteristics. Our results even indicated that navigation behavior and tagging pragmatics are connected.

Additionally, we analyzed the process of choosing tasks in a crowdsourcing environment (Chapter 10). Previous studies are mostly survey-based while we formulated explicit hypotheses which we compared based on sequential data produced by users of the Microworkers platform. Our results show that users mostly stick to categories and employers or choose semantically similar tasks. Also, since crowdsourcing campaigns are only available until all their tasks have been completed, we developed an approach to cope with such temporal constraints.

Finally, in several smaller case studies (Chapter 11), we analyzed human navigation behavior in the context of exploring urban noise pollution, and used MixedTrails and SubTrails to study navigation of Wikipedia users when searching for specific articles, and music sequences from the music listening portal last.fm.

12.2. Outlook

In this thesis, we covered several novel methods for exploring, analyzing, and understanding human navigation behavior especially focusing on its inherently heterogeneous nature (Part II) and applied them in a variety of application scenarios (Part III). Our work is extensive in both areas and, thus, inspires several lines of future work. We

Methodology. HypTrails and MixedTrails provide a powerful framework for comparing a wide range of hypotheses about human navigation behavior. However, the foremost limitation of these approaches is that ideas and intuitions already need to exist. While explorative approaches — like the proposed SubTrails method — somewhat help to discover certain sub-processes, thus possibly inspiring new theories, they still do not result in overall explanations of the observed data. In this context, approaches with clustering characteristics may be helpful [e.g., 171]. However, usually such methods do not yield interpretable results out of the box or are limited in their use of available background information. Future work may address this issue and develop methods which can take into account a wide variety of background data and *automatically* extract explanations for human navigation behavior with regard to heterogeneous aspects as well as various navigation characteristics.

Also, for HypTrails as well as MixedTrails, even if the idea for several hypotheses exists, formulating them is a challenging task. While they provide a very flexible framework to incorporate many different aspects of human navigation behavior including a wide array

of background information, they require very exact specification and possibly tedious parameter tuning. In this context, it may be possible to formulate hypothesis on a more abstract level, that is, leaving the parameters unspecified or at least setting a prior to bias them, instead of defining and comparing the parameter settings as separate hypotheses. It may even be possible to allow formulating hypotheses in a more general, possibly non-parametric framework, automatically adapting the underlying model beyond the current Markov chain approach. However, while this would better represent probabilistic dependencies, it most likely will increase the complexity of interpreting the results. Sensible trade-offs will have to be investigated.

Case studies. With regard to case studies, we only scratched the surface of the experiments enabled by our methodological contributions. For example, we are sure that there are more interesting features to find in the photowalking data covered in Chapter 7, especially when considering structures across different cities. Also, our studies on choosing campaigns in crowdsourcing environments as well as social tagging systems can potentially be extended by applying MixedTrails and SubTrails.

Overall. Our proposed methods enable a wide range of novel studies on human navigation behavior in the light of its inherently heterogeneous nature. With this, we provide a solid building block for further, more advanced analysis methods. Especially, our methodological contributions — MixedTrails and SubTrails — that support both, hypothesis comparison as well as conception, open up new ways of studying navigational data. We firmly believe that this will trigger various novel insights in many application scenarios. Accordingly, our case studies illustrated the variety of domains and facets in which navigational processes can be studied and, thus, showcased the applicability and flexibility of our approaches. In the process, we presented new aspects of human navigation which we hope will inspire a multitude of future studies in order to ultimately help to better understand the multi-faceted, and inherently heterogeneous, nature of human navigation behavior.

Part IV.

Appendix

A. General notation table

S	set of all states $S = \{s_1, \ldots, s_n\}$
D	set of observed transitions $D = \{t_1, ..., t_m\}$
src_k, dst_k	source state src_k and the destination state dst_k of transtion t_k
i_k, j_k	index of the source state i_k and the destination state j_k of transtion t_k
$t_{i,j}$	a transition from state s_i to state s_j
$n_{i,j}$	the count of transitions from state s_i to state s_j in a specific dataset D
$\boldsymbol{T}$	the transition count matrix $\boldsymbol{T} = (n_{i,j})$ holds the transition counts $n_{i,j}$ between all states s_i, s_j in S with regard to a dataset D
$\theta_{i,j}$	probability of a transition from state s_i to state s_j
$\boldsymbol{\theta}_{s_i}$	transition probabilities from state s_i to all other states, i.e., $\boldsymbol{\theta}_{s_i} = (\theta_{i,1}, \ldots, \theta_{i,n})$
$\boldsymbol{\theta}$	transition probabilities between all states, i.e., $\boldsymbol{\theta} = (\theta_{i,j}) = \{\boldsymbol{\theta}_{s_i} \mid s_i \in S\}$
$\boldsymbol{\theta}_D$	the transition probability matrix derived from the transition count matrix $\boldsymbol{T}_D$
$\phi_{i,j}$	belief (as defined by a hypothesis) in the probability of a transition from state s_i to state s_j
$\boldsymbol{\phi}_{s_i}$	belief in transition probabilities from state s_i to all states, i.e., $\boldsymbol{\phi}_{s_i} = (\phi_{i,1}, \ldots, \phi_{i,n})$
$\boldsymbol{\phi}$	belief in transition probabilities between all states, i.e., $\boldsymbol{\phi} = (\phi_{i,j}) = \{\boldsymbol{\phi}_{s_i} \mid s_i \in S\}$
$\alpha_{i,j}$	Dirichlet parameter ($\in \mathbb{R}^+$) for the transition from state s_i to state s_j
$\boldsymbol{\alpha}_{s_i}$	Dirichlet parameters for state s_i, i.e., $\boldsymbol{\alpha}_{s_i} = (\alpha_{i,1}, \ldots, \alpha_{i,n})$
$\boldsymbol{\alpha}$	Dirichlet parameters for all transitions, i.e., $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_{s_i} \mid s_i \in S\}$

Table A.1.: General notations. Overview of the most important notations with regard to Markov chains (cf. Section 3.2) and HypTrails (cf. Section 3.3.2)

B. MixedTrails notation table

G	set of all groups $G = \{g_1, \ldots, g_o\}$
$\gamma_{g\|t}$	probability for transition t to belong to group g
$\boldsymbol{\gamma}_t$	group assignment probabilities for a single transitions $\boldsymbol{\gamma}_t = \{\gamma_{g\|t} \| g \in G\}$
$\boldsymbol{\gamma}$	group assignment probabilities for all transitions $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_t \| t \in D\}$
$\theta_{i,j\|g}$	probability of a transition from state s_i to state s_j in group g
$\boldsymbol{\theta}_{s_i\|g}$	transition probabilities from state s_i to all other states in group g, i.e., $\boldsymbol{\theta}_{s_i\|g} = (\theta_{i,1\|g}, \ldots, \theta_{i,n\|g})$
$\boldsymbol{\theta}_g$	transition probabilities between states in group g, i.e., $\boldsymbol{\theta}_g = \{\boldsymbol{\theta}_{s_i\|g} \mid s_i \in S\}$
$\boldsymbol{\theta}$	transition probabilities between all states for all groups $\boldsymbol{\theta} = \{\boldsymbol{\theta}_g \mid g \in G\}$
$\phi_{i,j\|g}$	belief (as defined by a hypothesis) in the probability of a transition from state s_i to state s_j in group g
$\boldsymbol{\phi}_{s_i\|g}$	belief in transition probabilities from state s_i to all states in group g, i.e., $\boldsymbol{\phi}_{s_i\|g} = (\phi_{i,1\|g}, \ldots, \phi_{i,n\|g})$
$\boldsymbol{\phi}_g$	belief in transition probabilities between states in group g, i.e., $\boldsymbol{\phi}_g = \{\boldsymbol{\phi}_{s_i\|g} \mid s_i \in S\}$
$\boldsymbol{\phi}$	belief in transition probabilities between all states
$\alpha_{i,j\|g}$	Dirichlet parameter ($\in \mathbb{R}^+$) for the transition from state s_i to state s_j in group g
$\boldsymbol{\alpha}_{s_i\|g}$	Dirichlet parameters for state s_i in group g, i.e., $\boldsymbol{\alpha}_{s_i\|g} = (\alpha_{i,1\|g}, \ldots, \alpha_{i,n\|g})$
$\boldsymbol{\alpha}_g$	Dirichlet parameters for the transitions in group g, i.e., $\boldsymbol{\alpha}_g = \{\boldsymbol{\alpha}_{s_i\|g} \mid s_i \in S\}$
$\boldsymbol{\alpha}$	Dirichlet parameters for all groups $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_g \| g \in G\}$
Ω	the set of all group assignments, i.e., $\Omega = \{\{(t_1, g_1), ..., (t_m, g_m)\} \| (g_1, ..., g_m) \in G^{\|D\|}\}$
ω	a fixed group assignment $\omega \in \Omega$, which assigns a fixed group for each transition in a transition dataset D
p_ω	the probability for group assignment $\omega \in \Omega$
$n_{i,j\|g,\omega}$	the number of transitions in dataset D from state s_i to state s_j given group $g \in G$ and group assignment $\omega \in \Omega$
$\boldsymbol{T}_{g,\omega}$	the matrix $\boldsymbol{T}_{g,\omega} = (n_{i,j\|g,\omega})$ holds the number of transitions in dataset D between all states given group $g \in G$ and group assignment $\omega \in \Omega$

Table B.1.: MixedTrails notations. This is an overview of the most important notations used in the context of the MixedTrails approach (Chapter 4). Also see Table A.1, for more general notations in the context of Markov chains and HypTrails.

C. Derivation of the marginal likelihood of MTMC for MixedTrails

This chapter elaborates on the derivation of the marginal likelihood of the MTMC model used by the MixedTrails approach introduced in Chapter 4.

Given the generative process from Section 4.2.2 and by exploiting the fact that the transition probabilities $\boldsymbol{\theta}_g$ for each group g as well as the group assignment probabilities $\gamma_{g|t_k}$ for each transition t_k are independent, we can write the marginal likelihood of MTMC as follows:

$$\Pr(D|H) = \int \underbrace{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})}_{\text{likelihood}} \underbrace{\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})}_{\text{prior}} \, d\boldsymbol{\theta} \tag{C.1}$$

$$= \int \underbrace{\prod_{t_k \in D} \sum_{g \in G} \gamma_{g|t_k} \theta_{i_k,j_k|g}}_{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})} \underbrace{\prod_{g \in G} \Pr(\boldsymbol{\theta}_g|\boldsymbol{\alpha}_g)}_{\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})} \prod_{g \in G} d\boldsymbol{\theta}_g \tag{C.2}$$

To solve this integral we take a similar path as in the homogeneous case (cf. [453]). Thus, we need to get the grouping out of the integral. First, we focus on the likelihood $\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})$ where we extend the multiplication over all transitions resulting in an outer sum over all possible group assignments:

$$\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma}) = \prod_{t_k \in D} \sum_{g \in G} \gamma_{g|t} \theta_{i_k,j_k|g} \tag{C.3}$$

$$= \sum_{\substack{\omega \in \Omega \\ \Omega = \{\{(t_1,g_1),\ldots,(t_m,g_m)\}|(g_1,\ldots,g_m) \in G^{|D|}\}}} \prod_{(t_k,g_k) \in \omega} \gamma_{g_k|t_k} \theta_{i_k,j_k|g_k} \tag{C.4}$$

$$= \sum_{\omega \in \Omega} \underbrace{\prod_{(t_k,g_k) \in \omega} \gamma_{g_k|t_k}}_{p_\omega} \prod_{(t_k,g_k) \in \omega} \theta_{i_k,j_k|g_k} \tag{C.5}$$

$$= \sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{s_i,s_j \in S} \theta_{i,j|g}^{n_{i,j|g,\omega}} \tag{C.6}$$

Here, each ω represents a single, fixed group assignment of the set of transitions in D, i.e., each transition has been assigned to a fixed group. The set of all possible group assignments ω is defined as $\Omega = \{\{(t_1,g_1),...,(t_m,g_m)\}|(g_1,...,g_m) \in G^{|D|}\}$. Furthermore, p_ω represents the probability of the respective group assignment $\omega \in \Omega$. Finally, $n_{i,j|g,\omega}$ denotes the number of transitions from state s_i to state s_j given the group g and the

group assignment ω. What we observe is that, given a specific group assignment ω, the likelihood is the same as the likelihood derived by Singer et al. [453].

We now substitute the likelihood $\Pr(D|\boldsymbol{\theta}, \boldsymbol{\gamma})$ in Equation (C.2) with this reformulated likelihood (Equation (C.6)) and write the priors for the group dependent transition probabilities $\Pr(\boldsymbol{\theta}_g|\boldsymbol{\alpha}_g)$ based on the multivariate beta function. Then, we can calculate the marginal likelihood $\Pr(D|H)$ by taking advantage of the independence of the transition probabilities $\boldsymbol{\theta}_g$ between groups $g \in G$ and source states $s \in S$ as well as the independence of group assignment probabilities $\gamma_{g_k|t_k}$ between transitions $t_k \in D$:

$$\Pr(D|H) = \int \underbrace{\sum_{\omega\in\Omega} p_\omega \prod_{g\in G} \prod_{s_i,s_j\in S} \theta_{i,j|g}^{n_{i,j|g,\omega}}}_{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})} \prod_{g\in G} \underbrace{\prod_{s_i\in S} \frac{1}{B(\boldsymbol{\alpha}_{s_i|g})} \prod_{s_j\in S} \theta_{i,j|g}^{\alpha_{i,j|g}-1}}_{\Pr(\boldsymbol{\theta}_g|\boldsymbol{\alpha}_g)} \prod_{g\in G} d\boldsymbol{\theta}_g \tag{C.7}$$

$$= \sum_{\omega\in\Omega} p_\omega \prod_{g\in G} \prod_{s_i\in S} \frac{1}{B(\boldsymbol{\alpha}_{s_i|g})} \int \prod_{s_j\in S} \theta_{i,j|g}^{n_{i,j|g,\omega}+\alpha_{i,j|g}-1} d\boldsymbol{\theta}_g \tag{C.8}$$

$$= \sum_{\omega\in\Omega} \theta_\omega \prod_{g\in G} \underbrace{\prod_{s_i\in S} \frac{B(\boldsymbol{n}_{s_i|g,\omega} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})}}_{\Pr(D_{g|\omega}|\boldsymbol{\alpha}_g)} \tag{C.9}$$

This concludes the derivation of the marginal likelihood formula in Equation (4.4).

D. Total weighted variation as a special case of Bayesian belief update

As we have hinted on in Section 5.2.2, the weighted total variation ω_{tv} — which we use in our interestingness measure q_{tv} for discovering subgroups with exceptional transition behavior — can be interpreted as a special case of *Bayesian belief update*. We elaborate on this relation in this section.

Specifically, in Bayesian statistics, one's current beliefs H are expressed by probability distributions over some parameters $\boldsymbol{\mu}$. Given new information I, the *prior* belief $P(\boldsymbol{\mu}|H)$ is updated to a *posterior* belief $P(\boldsymbol{\mu}|H, I)$. In this context, *Bayesian belief update* is defined as the difference between the prior belief and the posterior belief. The amount of belief update was proposed by Silberschatz and Tuzhilin [449] as an interestingness measure for pattern mining in traditional settings. There, the belief update implied by a subgroup g is defined by the difference between the prior distribution of $\boldsymbol{\mu}$ (derived, e.g., from the overall dataset) and the posterior distribution of $\boldsymbol{\mu}$ after observing the instances covered by the subgroup g.

In our setting, the parameters $\boldsymbol{\mu}$ represent the transition probabilities of a Markov chain. In this context, we show that the *amount of Bayesian belief update* is order equivalent to the total variation measure $\omega_{tv}(g, D)$ from Section 5.2.2 if the transition probabilities derived from the reference matrix $\boldsymbol{T}_D$ are used as a very strong prior. That means that both measures ultimately imply the same ranking of subgroups.

Recall the total variation measure $\omega_{tv}(g, D)$, with $\boldsymbol{T}_g = (g_{i,j})$ being the transition counts from a subgroup g and $\boldsymbol{T}_D = (d_{i,j})$ being the entries of the reference matrix $\boldsymbol{T}_D$:

$$\omega_{tv}(g, D) = \sum_i \left(\sum_j g_{ij} \cdot \sum_j \left| \frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{d_{i,j}}{\sum_j d_{i,j}} \right| \right) \tag{D.1}$$

As Singer et al. [453] suggest, we can elicit the matrix of a Dirichlet prior $\boldsymbol{\alpha} = (\alpha_{i,j})$ using transition probabilities $\boldsymbol{\theta}_D = (\theta_{i,j}) = (d_{i,j}/\sum_j d_{i,j})$ derived from $\boldsymbol{T}_D$ by applying the formula $\alpha_{i,j} = (\kappa \cdot \theta_{i,j}) + 1$.[1] Here, κ specifies the strength of the belief expressed by the prior. The prior is updated to a posterior according to the transitions observed in a subgroup g which are given as a transition count matrix $\boldsymbol{T}_g = (g_{i,j})$. In this context, according to Singer et al. [454], the expected probabilities for a state transition from state

[1] Note that in Section 3.3.2, we use $\boldsymbol{\phi} = (\phi_{i,j})$ to represent the transition probability matrix representing the hypothesis. Here, we stick to the simplified notation of Chapter 5 which does not differentiate between hypothesis probability matrices and observed transition probability matrices.

s_i to state s_j in the prior are

$$\frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \tag{D.2}$$

and the expected probabilities in the posterior are

$$c_i \cdot \frac{g_{i,j}}{\sum_j g_{i,j}} + (1 - c_i) \cdot \frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \text{ , with } c_i = \frac{\sum_j g_{i,j}}{\sum_j (g_{i,j} + \alpha_{i,j})} \tag{D.3}$$

To determine the overall belief update BU for all state transitions, we compute the absolute difference between the posterior and the prior for each possible transition (from state s_i to state s_j) and aggregate these values by summing them up:

$$BU(H, D) = \sum_i \sum_j \left| \left(c_i \frac{g_{i,j}}{\sum_j g_{i,j}} + (1 - c_i) \frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \right) - \frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \right| \tag{D.4}$$

$$= \sum_i \sum_j \left| c_i \cdot \left(\frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \right) \right| \tag{D.5}$$

$$= \sum_i c_i \cdot \sum_j \left| \left(\frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \right) \right| \tag{D.6}$$

$$= \sum_i \frac{1}{\sum_j (g_{i,j} + \alpha_{i,j})} \sum_j g_{i,j} \sum_j \left(\left| \frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{\alpha_{i,j}}{\sum_j \alpha_{i,j}} \right| \right) \tag{D.7}$$

Now, assume that we have a very strong belief in the prior, i.e., $\kappa \to \infty$ and thus $\alpha_{i,j} \gg g_{i,j}$. Then, the right hand sum converges to the total variation δ_{tv} between the observed transition count matrix $\boldsymbol{T}_g$ and the reference matrix $\boldsymbol{T}_D$. The factor $\sum_j g_{i,j}$ corresponds to the weights w_i. The additional factor $\frac{1}{\sum_j (\alpha_{i,j,} + g_{i,j})}$ is approximately constant across all subgroups if $\alpha_{i,j} \gg g_{i,j}$ since $\alpha_{i,j}$ is independent from the evaluated subgroup. Overall, the weighted total variation ω_{tv} describes the amount of belief update a subgroup induces to a prior that reflects a very strong belief in the transition probabilities given by the reference matrix $\boldsymbol{T}_D$.

Bibliography

[1] Myriam Abramson. "Learning temporal user profiles of web browsing behavior". In: *International Conference on Social Computing (SocialCom)*. 2014, pp. 1–9.

[2] Tarek Abudawood and Peter Flach. "Evaluation Measures for Multi-class Subgroup Discovery". In: *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*. 2009, pp. 35–50.

[3] Eytan Adar, Jaime Teevan, and Susan T Dumais. "Large scale analysis of web revisitation patterns". In: *Conference on Human Factors in Computing Systems (CHI)*. 2008, pp. 1197–1206.

[4] Deepak Agarwal, Bee-Chung Chen, and Pradheep Elango. "Spatio-temporal models for estimating click-through rate". In: *International Conference on World Wide Web*. 2009, pp. 21–30.

[5] Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. "Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction". In: *Data Mining and Knowledge Discovery* 24.3 (2012), pp. 663–696.

[6] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. "Mining association rules between sets of items in large databases". In: *International Conference on Management of Data (SIGMOD)*. 1993, pp. 207–216.

[7] Rakesh Agrawal and Ramakrishnan Srikant. "Fast algorithms for mining association rules". In: *International Conference on Very Large Databases (VLDB)*. 1994, pp. 487–499.

[8] Rakesh Agrawal and Ramakrishnan Srikant. "Mining Sequential Patterns". In: *International Conference on Data Engineering (ICDE)*. 1995, pp. 3–14.

[9] Hirotogu Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. 1973, pp. 199–213.

[10] Hirotugu Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723.

[11] Siaw Akwawua and James A Pooler. "The development of an intervening opportunities model with spatial dominance effects". In: *Journal of Geographical Systems* 3.1 (2001), pp. 69–86.

[12] Laura Alessandretti, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli. "Multi-scale spatio-temporal analysis of human mobility". In: *PLOS ONE* 12.2 (2017), pp. 1–17.

[13] Corin R Anderson, Pedro Domingos, and Daniel S Weld. "Relational Markov models and their application to adaptive web navigation". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2002, pp. 143–152.

[14] William J Anderson. *Continuous-time Markov chains: An applications-oriented approach*. 2012.

[15] Natalia Andrienko, Gennady Andrienko, and Peter Gatalsky. "Exploratory spatio-temporal visualization: an analytical review". In: *Journal of Visual Languages & Computing* 14.6 (2003), pp. 503–541.

[16] Theo Arentze, Frank Hofman, Henk van Mourik, and Harry Timmermans. "ALBATROSS: multiagent, rule-based model of activity pattern decisions". In: *Transportation Research Record* 1706 (2000), pp. 136–144.

[17] Hazleen Aris. "Influencing Factors in Mobile Crowdsourcing Participation: A Review of Empirical Studies". In: *Conference on Computer Science and Computational Modelling*. 2014.

[18] Akinori Asahara, Kishiko Maruyama, Akiko Sato, and Kouichi Seto. "Pedestrian-movement Prediction Based on Mixed Markov-chain Model". In: *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*. 2011, pp. 25–33.

[19] Fereshteh Asgari, Vincent Gauthier, and Monique Becker. "A survey on Human Mobility and its applications". In: *CoRR* abs/1307.0814 (2013).

[20] Martin Atzmueller, Martin Becker, Stephan Doerfel, Mark Kibanov, Andreas Hotho, Björn-Elmar Macek, Folke Mitzlaff, Juergen Mueller, Christoph Scholz, and Gerd Stumme. "Ubicon: Observing Social and Physical Activities". In: *International Conference on Green Computing and Communications*. 2012, pp. 317–324.

[21] Martin Atzmueller, Stephan Doerfel, and Folke Mitzlaff. "Description-oriented community detection using exhaustive subgroup discovery". In: *Information Sciences* 329 (2016), pp. 965–984.

[22] Martin Atzmueller and Florian Lemmerich. "Exploratory pattern mining on social media using geo-references and social tagging information". In: *International Journal of Web Science* 2.1–2 (2013), pp. 80–112.

[23] Martin Atzmueller and Florian Lemmerich. "Fast subgroup discovery for continuous target concepts". In: *International Symposium on Methodologies for Intelligent Systems (ISMIS)*. 2009, pp. 35–44.

[24] Martin Atzmueller and Florian Lemmerich. "VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics". In: *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*. 2012, pp. 842–845.

[25] Martin Atzmueller and Folke Mitzlaff. "Efficient Descriptive Community Mining". In: *International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. 2011, pp. 459–464.

[26] Martin Atzmueller and Frank Puppe. "SD-Map–A fast algorithm for exhaustive subgroup discovery". In: *European Conference on Principles of Data Mining and Knowledge Discovery*. 2006, pp. 6–17.

[27] Daniel Austin, Robin M Cross, Tamara Hayes, and Jeffrey Kaye. "Regularity and predictability of human mobility in personal space". In: *PloS one* 9.2 (2014), e90256.

[28] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. "Sequential pattern mining using a bitmap representation". In: *International Conference on Knowledge Discovery and Data Mmining (SIGKDD)*. 2002, pp. 429–435.

[29] Lars Backstrom, Eric Sun, and Cameron Marlow. "Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity". In: *International Conference on World Wide Web*. 2010, pp. 61–70.

[30] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*. Vol. 463. 1999.

[31] Ricardo Baeza-Yates and Alessandro Tiberi. "Extracting semantic relations from query logs". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2007, pp. 76–85.

[32] Duygu Balcan, Vittoria Colizza, Bruno Goncalves, Hao Hu, Jose J. Ramasco, and Alessandro Vespignani. "Multiscale mobility networks and the spatial spreading of infectious diseases". In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21484–21489.

[33] Shumeet Baluja, Rohan Seth, D Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. "Video suggestion and discovery for youtube: taking random walks through the view graph". In: *International Conference on World Wide Web*. 2008, pp. 895–904.

[34] Sasha A Barab, Bruce E Bowdish, and Kimberly A Lawless. "Hypermedia navigation: Profiles of hypermedia users". In: *Educational Technology Research and Development* 45.3 (1997), pp. 23–41.

[35] Albert-László Barabási and Réka Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (1999), pp. 509–512.

[36] Gabriele Barbieri, Francois Pachet, Pierre Roy, and Mirko Degli Esposti. "Markov constraints for generating lyrics with style". In: *European Conference on Artificial Intelligence (ECAI)*. 2012, pp. 115–120.

[37] Daniele Barchiesi, Helen Susannah Moat, Christian Alis, Steven Bishop, and Tobias Preis. "Quantifying international travel flows using Flickr". In: *PloS one* 10.7 (2015), e0128470.

[38] Marcia J. Bates. "The design of browsing and berrypicking techniques for the online search interface". In: *Online Review* 13.5 (1989), pp. 407–424.

[39] Roberto J Bayardo Jr. "Efficiently mining long patterns from databases". In: *International Conference on Management of Data (SIGMOD)* (1998), pp. 85–93.

[40] Martin Becker, Kathrin Borchert, Matthias Hirth, Hauke Mewes, Andreas Hotho, and Phuoc Tran-Gia. "MicroTrails: Comparing Hypotheses About Task Selection on a Crowdsourcing Platform". In: *International Conference on Knowledge Technologies and Data-driven Business*. 2015, 10:1–10:8.

[41] Martin Becker, Florian Lemmerich, Philipp Singer, Markus Strohmaier, and Andreas Hotho. "MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data". In: *Data Mining and Knowledge Discovery* 31.5 (2017), pp. 1359–1390.

[42] Martin Becker, Hauke Mewes, Andreas Hotho, Dimitar Dimitrov, Florian Lemmerich, and Markus Strohmaier. "SparkTrails: A MapReduce Implementation of HypTrails for Comparing Hypotheses About Human Trails". In: *International Conference Companion on World Wide Web*. 2016, pp. 17–18.

[43] Martin Becker, Juergen Mueller, Andreas Hotho, and Gerd Stumme. "A Generic Platform for Ubiquitous and Subjective Data". In: *Conference on Pervasive and Ubiquitous Computing Adjunct Publication*. 2013, pp. 1175–1182.

[44] Martin Becker, Philipp Singer, Florian Lemmerich, Andreas Hotho, Denis Helic, and Markus Strohmaier. "Photowalking the City: Comparing Hypotheses About Urban Photo Trails on Flickr". In: *International Conference on Social Informatics*. 2015, pp. 227–244.

[45] Martin Becker, Philipp Singer, Florian Lemmerich, Andreas Hotho, Denis Helic, and Markus Strohmaier. "VizTrails: An Information Visualization Tool for Exploring Geographic Movement Trajectories". In: *Conference on Hypertext & Social Media*. 2015, pp. 319–320.

[46] Richard Becker, Ramon Caceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. "Human Mobility Characterization from Cellular Network Data". In: *Communications of the ACM* 56.1 (2013), pp. 74–82.

[47] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. "Predicting human mobility through the assimilation of social media traces into mobility models". In: *EPJ Data Science* 5.1 (2016), p. 30.

[48] Vitaly Belik, Theo Geisel, and Dirk Brockmann. "Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases". In: *Phys. Rev. X* 1.1 (2011), p. 011001.

[49] Richard Bellman. "A Markovian decision process". In: *Journal of Mathematics and Mechanics* 6.5 (1957), pp. 679–684.

[50] Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. "A Bayesian Wilcoxon signed-rank test based on the Dirichlet process". In: *International Conference on Machine Learning (ICML)*. 2014, pp. 1026–1034.

[51] Fabricio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgilio A. F. Almeida. "Characterizing User Behavior in Online Social Networks". In: *Conference on Internet Measurement Conference (SIGCOMM)*. 2009, pp. 49–62.

[52] Yoshua Bengio. "Markovian Models for Sequential Data". In: *Neural computing surveys* 2.199 (1999), pp. 129–162.

[53] Simon Benhamou. "How many animals really do the Levy walk?" In: *Ecology* 88.8 (2007), pp. 1962–1969.

[54] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the royal statistical society. Series B (Methodological)* 57.1 (1995), pp. 289–300.

[55] D Benz, A Hotho, R Jäschke, B Krause, F Mitzlaff, C Schmitz, and G Stumme. "The social bookmark and publication management system BibSonomy". In: *VLDB Journal* 19.6 (2010), pp. 849–875.

[56] Dominik Benz, Andreas Hotho, Stefan Stützer, and Gerd Stumme. "Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge". In: *Web Science Conference (WebSci)*. 2010.

[57] Dominik Benz, Beate Krause, G Praveen Kumar, Andreas Hotho, and Gerd Stumme. "Characterizing Semantic Relatedness of Search Query Terms". In: *Workshop on Explorative Analytics of Information Networks at ECML PKDD*. 2009, pp. 119–135.

[58] Pauline van den Berg, Theo Arentze, and Harry Timmermans. "A path analysis of social networks, telecommunication and social activity–travel patterns". In: *Transportation Research Part C: Emerging Technologies* 26 (2013), pp. 256–268.

[59] Tim Berners-Lee, Robert Cailliau, Jean-François Groff, and Bernd Pollermann. "World-Wide Web: The Information Universe". In: *Internet Research* 2.1 (1992), pp. 52–58.

[60] Tim Berners-Lee, James Hendler, Ora Lassila, et al. "The semantic web". In: *Scientific american* 284.5 (2001), pp. 28–37.

[61] Chandra R. Bhat and Frank S. Koppelman. "Activity-Based Modeling of Travel Demand". In: *Handbook of Transportation Science*. 1999, pp. 35–61.

[62] Mikhail Bilenko and Ryen W White. "Mining the search trails of surfing crowds: identifying relevant websites from user activity". In: *International Conference on World Wide Web*. 2008, pp. 51–60.

[63] Amy Blackstone. *Sociological Inquiry Principles: Qualitative and Quantitative Methods*. 2012.

[64] David M Blei and Pedro J Moreno. "Topic segmentation with an aspect hidden Markov model". In: *Conference on Research and Development in Information Retrieval (SIGIR)*. 2001, pp. 343–348.

[65] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. "A survey of results on mobile phone datasets analysis". In: *EPJ Data Science* 4.1 (2015), p. 10.

[66] Chiara Boldrini and Andrea Passarella. "HCMM: Modelling spatial and temporal properties of human mobility driven by users' social relationships". In: *Computer Communications* 33.9 (2010), pp. 1056–1074.

[67] J Bollen, H Van de Sompel, A Hagberg, L Bettencourt, R Chute, M A Rodriguez, and L Balakireva. "Clickstream data yields high-resolution maps of science". In: *PLoS ONE* 4.3 (2009), e4803.

[68] Geoffray Bonnin and Dietmar Jannach. "Automated Generation of Music Playlists: Survey and Experiments". In: *ACM Computer Surveys (CSUR)* 47.2 (2014), 26:1–26:35.

[69] Jose Borges and Mark Levene. "Data Mining of User Navigation Patterns". In: *International Workshop on Web Usage Analysis and User Profiling*. 2000, pp. 92–112.

[70] Jose Borges and Mark Levene. "Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions". In: *IEEE Transactions on Knowledge and Data Engineering* 19.4 (2007), pp. 441–452.

[71] Jose Borges and Mark Levene. "Generating Dynamic Higher-Order Markov Models in Web Usage Mining". In: *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. 2005, pp. 34–45.

[72] José Borges and Mark Levene. "Testing the Predictive Power of Variable History Web Usage". In: *Soft Computing* 11.8 (2007), pp. 717–727.

[73] Vincent Borrel, Franck Legendre, Marcelo Dias De Amorim, and Serge Fdida. "Simps: Using sociology for personal mobility". In: *IEEE/ACM Transactions on Networking* 17.3 (2009), pp. 831–842.

[74] Patricia L Brantingham and Paul J Brantingham. "Burglar mobility and crime prevention planning". In: *Coping with burglary*. 1984, pp. 77–95.

[75] D. Brockmann, L. Hufnagel, and T. Geisel. "The scaling laws of human travel". In: *Nature* 439.7075 (2006), pp. 462–465.

[76] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. 2011.

[77] Duncan P Brumby and Andrew Howes. "Good Enough But I'll Just Check: Web-page Search As Attentional Refocusing". In: *Conference on Cognitive Modeling (ICCM)*. 2004, pp. 46–51.

[78] Randolph E Bucklin and Catarina Sismeiro. "A model of web site browsing behavior estimated on clickstream data". In: *Journal of marketing research* 40.3 (2003), pp. 249–267.

[79] Peter Buhlmann and Abraham J. Wyner. "Variable Length Markov Chains". In: *Annals of Statistics* 27 (1999), pp. 480–513.

[80] Kenneth P Burnham and David R Anderson. "Multimodel inference understanding AIC and BIC in model selection". In: *Sociological methods & research* 33.2 (2004), pp. 261–304.

[81] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. "Model-Based Clustering and Visualization of Navigation Patterns on a Web Site". In: *Data Minining and Knowledge Discovery* 7.4 (2003), pp. 399–424.

[82] Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. "Human mobility prediction based on individual and collective geographical preferences". In: *International Conference on Intelligent Transportation Systems (ITSC)*. 2010, pp. 312–317.

[83] Tracy Camp, Jeff Boleng, and Vanessa Davies. "A survey of mobility models for ad hoc network research". In: *Wireless communications and mobile computing* 2.5 (2002), pp. 483–502.

[84] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. "Towards Context-aware Search by Learning a Very Large Variable Length Hidden Markov Model from Search Logs". In: *International Conference on World Wide Web*. 2009, pp. 191–200.

[85] Donald L Capone and Woodrow W Nichols Jr. "Urban structure and criminal mobility". In: *American Behavioral Scientist* 20.2 (1976), pp. 199–213.

[86] Malachy Carey, Chris Hendrickson, and Krishnaswami Siddharthan. "A method for direct estimation of origin/destination trip matrices". In: *Transportation Science* 15.1 (1981), pp. 32–49.

[87] Erran Carmel, Stephen F. Crawford, and Hsinchun Chen. "Browsing in hypertext: A cognitive study". In: *IEEE Transactions on Systems, Man, and Cybernetics* 22.5 (1992), pp. 865–884.

[88] Gerald AP Carrothers. "An historical bedew of the gravity and potential concepts of human interaction". In: *Journal of the American Institute of Planners* 22.2 (1956), pp. 94–102.

[89] Ronald P Carver. "The Case Against Statistical Significance Testing". In: *Harvard Educational Review* 48.3 (1978), pp. 378–399.

[90] George Casella and Roger L Berger. *Statistical inference*. Vol. 2. 2002.

[91] Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. "From Taxi GPS Traces to Social and Community Dynamics: A Survey". In: *ACM Computing Surveys (CSUR)* 46.2 (2013), 17:1–17:34.

[92] Lara D. Catledge and James E. Pitkow. "Characterizing browsing strategies in the World-Wide Web". In: *Computer Networks and ISDN Systems* 27.6 (1995), pp. 1065–1073.

[93] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. "Semantic Grounding of Tag Relatedness in Social Bookmarking Systems". In: *International Conference on The Semantic Web*. 2008, pp. 615–631.

[94] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. "Dynamics of person-to-person interactions from distributed RFID sensor networks". In: *PloS one* 5.7 (2010), e11596.

[95] Robert Cervero. *Land-use mixing and suburban mobility*. Tech. rep. University of California Transportation Center, 1989.

[96] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. "A measurement-driven analysis of information propagation in the flickr social network". In: *International Conference on World Wide Web*. 2009, pp. 721–730.

[97] Ann Chadwick-Dias, Michelle McNulty, and Tom Tullis. "Web usability and age: how design changes can improve performance". In: *ACM SIGCAPH Computers and the Physically Handicapped*. 73-74. 2003, pp. 30–37.

[98] Matthew Chalmers, Kerry Rodden, and Dominique Brodbeck. "The order of things: activity-centred information access". In: *Computer Networks and ISDN Systems* 30.1 (1998), pp. 359–367.

[99] C Chen, J Ma, Y Susilo, Y Liu, and M Wang. "The promises of big data and small data for travel behavior (aka human mobility) analysis". In: *Transportation Research Part C: Emerging Technologies* 68 (2016), pp. 285–299.

[100] Ming-Syan Chen, Jong Soo Park, and Philip S Yu. "Data mining for path traversal patterns in a web environment". In: *International Conference on Distributed Computing Systems*. 1996, pp. 385–392.

[101] Ming-Syan Chen, Jong Soo Park, and Philip S. Yu. "Efficient data mining for path traversal patterns". In: *IEEE Transactions on knowledge and data engineering* 10.2 (1998), pp. 209–221.

[102] Wang Chen, Qiang Gao, and Hua-Gang Xiong. "Uncovering urban mobility patterns and impact of spatial distribution of places on movements". In: *International Journal of Modern Physics C* 28.01 (2017), p. 1750004.

[103] Wei Chen, Fangzhou Guo, and Fei-Yue Wang. "A survey of traffic data visualization". In: *IEEE Transactions on Intelligent Transportation Systems* 16.6 (2015), pp. 2970–2984.

[104] Xin Chen and Xiaodong Zhang. "A Popularity-Based Prediction Model for Web Prefetching". In: *Computer* 36.3 (2003), pp. 63–70.

[105] Yanguang Chen. "The distance-decay function of geographical gravity model: Power law or exponential law?" In: *Chaos, Solitons & Fractals* 77 (2015), pp. 174–189.

[106] Chen Cheng, Haiqin Yang, Michael R Lyu, and Irwin King. "Where You Like to Go Next: Successive Point-of-Interest Recommendation". In: *International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 13. 2013, pp. 2605–2611.

[107] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. "Exploring millions of footprints in location sharing services". In: *International Conference on Web and Social Media (ICWSM)*. 2011, pp. 81–88.

[108] Ed H. Chi, Peter L. T. Pirolli, Kim Chen, and James E. Pitkow. "Using Information Scent to Model User Information Needs and Actions and the Web". In: *Conference on Human Factors in Computing Systems (SIGCHI)*. 2001, pp. 490–497.

[109] Siddhartha Chib. "Marginal likelihood from the Gibbs output". In: *Journal of the American Statistical Association* 90.432 (1995), pp. 1313–1321.

[110] Siddhartha Chib and Ivan Jeliazkov. "Marginal likelihood from the Metropolis–Hastings output". In: *Journal of the American Statistical Association* 96.453 (2001), pp. 270–281.

[111] Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. "Are web users really Markovian?" In: *International Conference on World Wide Web*. 2012, pp. 609–618.

[112] Terry L Childers, Christopher L Carr, Joann Peck, and Stephen Carson. "Hedonic and utilitarian motivations for online retail shopping behavior". In: *Journal of retailing* 77.4 (2001), pp. 511–535.

[113] Lydia B Chilton, John J Horton, Robert C Miller, and Shiri Azenkot. "Task search in a human computation market". In: *Workshop on Human Computation*. 2010, pp. 1–9.

[114] Wai Ki Ching, Eric S Fung, and Michael K Ng. "Higher-order Markov chain models for categorical data sequences". In: *Naval Research Logistics (NRL)* 51.4 (2004), pp. 557–574.

[115] Wai-Ki Ching, Ximin Huang, Michael K Ng, and Tak-Kuen Siu. "Higher-Order Markov Chains". In: *Markov Chains*. 2013, pp. 141–176.

[116] Eunjoon Cho, Seth A Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2011, pp. 1082–1090.

[117] Chun Wei Choo, Brian Detlor, and Dan Turnbull. "Information seeking on the Web: An integrated model of browsing and searching". In: *First Monday* 5.2 (2000).

[118] Alvin Chua, Ernesto Marcheggiani, Loris Servillo, and Andrew Vande Moere. "FlowSampler: Visual analysis of urban flows in geolocated social media data". In: *International Conference on Social Informatics (Socinfo)*. 2014, pp. 5–17.

[119] Gerda Claeskens and Nils Lid Hjort. "The focused information criterion". In: *Journal of the American Statistical Association* 98.464 (2003), pp. 900–916.

[120] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data". In: *SIAM review* 51.4 (2009), pp. 661–703.

[121] Jacob Cohen. "The earth is round ($p < .05$)". In: *American Psychologist* 49.12 (1994), pp. 997–1003.

[122] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. "Data preparation for mining world wide web browsing patterns". In: *Knowledge and information systems* 1.1 (1999), pp. 5–32.

[123] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web". In: *International Conference on Tools with Artificial Intelligence (ICTAI)*. 1997, pp. 558–567.

[124] Caitlin Cottrill, Francisco Pereira, Fang Zhao, Iněs Dias, Hock Lim, Moshe Ben-Akiva, and P Zegras. "Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore". In: *Transportation Research Record: Journal of the Transportation Research Board* 2354 (2013), pp. 59–67.

[125] J.F. Cove and B.C. Walsh. "Online text retrieval via browsing". In: *Information Processing & Management* 24.1 (1988), pp. 31–37.

[126] David R Cox. "Further results on tests of separate families of hypotheses". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 24.2 (1962), pp. 406–424.

[127] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. "Inferring social ties from geographic coincidences". In: *Proceedings of the National Academy of Sciences* 107.52 (2010), pp. 22436–22441.

[128] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. "Mapping the world's photos". In: *International Conference on World Wide Web*. 2009, pp. 761–770.

[129] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. "Bridging the Gap Between Physical Location and Online Social Networks". In: *International Conference on Ubiquitous Computing*. 2010, pp. 119–128.

[130] Imre Csiszár, Paul C Shields, et al. "The consistency of the BIC Markov order estimator". In: *The Annals of Statistics* 28.6 (2000), pp. 1601–1619.

[131] Dana Cuff, Mark Hansen, and Jerry Kang. "Urban sensing: out of the woods". In: *Communications of the ACM* 51.3 (2008), pp. 24–33.

[132] Alexander Dallmann, Thomas Niebler, Florian Lemmerich, and Andreas Hotho. "Extracting Semantics from Random Walks on Wikipedia: Comparing Learning and Counting Methods". In: *The Workshops of the Tenth International AAAI Conference on Web and Social Media Wiki*. 2016, pp. 33–40.

[133] Cam Davidson-Pilon. *Probabilistic Programming & Bayesian Methods for Hackers*. 2015.

[134] Brian D Davison. "Learning Web Request Patterns". In: *Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. 2004, pp. 435–459.

[135] Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. "Automatic Construction of Travel Itineraries Using Social Breadcrumbs". In: *Conference on Hypertext and Hypermedia*. 2010, pp. 35–44.

[136] Yves-Alexandre De Montjoye, Cesar A Hidalgo, Michel Verleysen, and Vincent D Blondel. "Unique in the crowd: The privacy bounds of human mobility". In: *Scientific reports* 3 (2013), p. 1376.

[137] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters". In: *Communications of the ACM* 51.1 (2008), pp. 107–113.

[138] Ayhan Demiriz. "webspade: A parallel sequence mining algorithm to analyze web log data". In: *International Conference on Data Mining (ICDM)*. 2002, pp. 755–758.

[139] Merkebe Getachew Demissie, Francisco Antunes, Carlos Bento, Santi Phithakkitnukoon, and Titipat Sukhvibul. "Inferring origin-destination flows using mobile phone data: A case study of Senegal". In: *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 2016, pp. 1–6.

[140] Urška Demšar and Kirsi Virrantaus. "Space–time density of trajectories: exploring spatio-temporal patterns in movement data". In: *International Journal of Geographical Information Science* 24.10 (2010), pp. 1527–1542.

[141] Mukund Deshpande and George Karypis. "Selective Markov models for predicting Web page accesses". In: *ACM Transactions on Internet Technology* 4.2 (2004), pp. 163–184.

[142] James Diebel and Sebastian Thrun. "An application of markov random fields to range sensing". In: *International Conference on Neural Information Processing Systems (NIPS)*. 2005, pp. 291–298.

[143] Frans M Dieleman, Martin Dijst, and Guillaume Burghouwt. "Urban form and travel behaviour: micro-level household attributes and residential context". In: *Urban studies* 39.3 (2002), pp. 507–527.

[144] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. "What Makes a Link Successful on Wikipedia?" In: *International Conference on World Wide Web*. 2017, pp. 917–926.

[145] Stephan Doerfel, Daniel Zoller, Philipp Singer, Thomas Niebler, Andreas Hotho, and Markus Strohmaier. "What Users Actually do in a Social Tagging System: A Study of User Behavior in BibSonomy". In: *ACM Transactions on the Web* 10.2 (2016), 14:1–14:32.

[146] Xing Dongshan and Shen Junyi. "A New Markov Model For Web Access Prediction". In: *Computing in Science and Engineering* 4.6 (2002), pp. 34–39.

[147] Lennart Downar and Wouter Duivesteijn. "Exceptionally Monotone Models: The Rank Correlation Model Class for Exceptional Model Mining". In: *International Conference on Data Mining (ICDM)*. 2015, pp. 111–120.

[148] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. "Understanding the relationship between searchers' queries and information goals". In: *Conference on Information and Knowledge Management (CIKM)*. 2008, pp. 449–458.

[149] Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. "Exceptional model mining". In: *Data Mining and Knowledge Discovery* 30.1 (2016), pp. 47–98.

[150] Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. "Different Slopes for Different Folks: Mining for Exceptional Regression Models with Cook's Distance". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2012, pp. 868–876.

[151] Wouter Duivesteijn and Arno Knobbe. "Exploiting False Discoveries–Statistical Validation of Patterns and Quality Measures in Subgroup Discovery". In: *International Conference on Data Mining (ICDM)*. 2011, pp. 151–160.

[152] Wouter Duivesteijn, Arno Knobbe, Ad Feelders, and Matthijs van Leeuwen. "Subgroup Discovery Meets Bayesian Networks–An Exceptional Model Mining Approach". In: *International Conference on Data Mining (ICDM)*. 2010, pp. 158–167.

[153] Christopher W. Dunn, Minaxi Gupta, Alexandre Gerber, and Oliver Spatscheck. "Navigation Characteristics of Online Social Networks and Search Engines Users". In: *Workshop on Online Social Networks*. 2012, pp. 43–48.

[154] Olive J. Dunn. "Multiple Comparisons Among Means". In: *Journal of the American Statistical Association* 56.293 (1961), pp. 52–64.

[155] Nathan Eagle and Alex (Sandy) Pentland. "Reality mining: sensing complex social systems". In: *Personal and Ubiquitous Computing* 10.4 (2006), pp. 255–268.

[156] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. "Inferring friendship network structure by using mobile phone data". In: *Proceedings of the National Academy of Sciences* 106.36 (2009), pp. 15274–15278.

[157] Nathan Eagle and Alex Sandy Pentland. "Eigenbehaviors: identifying structure in routine". In: *Behavioral Ecology and Sociobiology* 63.7 (2009), pp. 1057–1066.

[158] Andrew M. Edwards, Richard A. Phillips, Nicholas W. Watkins, Mervyn P. Freeman, Eugene J. Murphy, Vsevolod Afanasyev, Sergey V. Buldyrev, M. G. E. da Luz, E. P. Raposo, H. Eugene Stanley, and Gandhimohan M. Viswanathan. "Revisiting Levy flight search patterns of wandering albatrosses, bumblebees and deer". In: *Nature* 449.7165 (2007), pp. 1044–1048.

[159] Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26.

[160] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. 1994.

[161] Magdalini Eirinaki, Michalis Vazirgiannis, and Dimitris Kapogiannis. "Web Path Recommendations Based on Page Ranking and Markov Models". In: *International Workshop on Web Information and Data Management*. 2005, pp. 2–9.

[162] Bart Elen, Jan Theunis, Stefano Ingarra, Andrea Molino, Joris Van, den Bossche, Matteo Reggente, and Vittorio Loreto. "The EveryAware SensorBox: a tool for community-based air quality monitoring". In: *Sensing a Changing World Workshop 2*. 2012, pp. 1–7.

[163] Nicole B Ellison et al. "Social network sites: Definition, history, and scholarship". In: *Journal of Computer-Mediated Communication* 13.1 (2007), pp. 210–230.

[164] Peter Erdi and Gabor Lente. "Continuous Time Discrete State Stochastic Models". In: *Stochastic Chemical Kinetics*. 2014, pp. 25–70.

[165] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*. Vol. 3. 1990.

[166] Lisette Espin Noboa, Florian Lemmerich, Philipp Singer, and Markus Strohmaier. "Discovering and characterizing mobility patterns in urban spaces: A study of Manhattan taxi data". In: *International Conference Companion on World Wide Web*. 2016, pp. 537–542.

[167] Federico Michele Facca and Pier Luca Lanzi. "Mining interesting knowledge from weblogs: a survey". In: *Data & Knowledge Engineering* 53.3 (2005), pp. 225–241.

[168] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. *Advances in knowledge discovery and data mining*. Vol. 21. 1996.

[169] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases". In: *AI magazine* 17.3 (1996), p. 37.

[170] Zhenni Feng and Yanmin Zhu. "A Survey on Trajectory Data Mining: Techniques and Applications". In: *IEEE Access* 4 (2016), pp. 2056–2067.

[171] Flavio Figueiredo, Bruno Ribeiro, Jussara M. Almeida, and Christos Faloutsos. "TribeFlow: Mining & Predicting User Trajectories". In: *International Conference on World Wide Web*. 2016, pp. 695–706.

[172] Flavio Figueiredo, Bruno Ribeiro, Jussara M Almeida, Nazareno Andrade, and Christos Faloutsos. "Mining Online Music Listening Trajectories". In: *International Society for Music Information Retrieval Conference (ISMIR)*. 2016, pp. 688–694.

[173] Shai Fine, Yoram Singer, and Naftali Tishby. "The hierarchical hidden Markov model: Analysis and applications". In: *Machine learning* 32.1 (1998), pp. 41–62.

[174] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu., and V. S. Tseng. "SPMF: a Java Open-Source Pattern Mining Library". In: *Journal of Machine Learning Research (JMLR)* 15 (2014), pp. 3389–3393.

[175] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, and Yun Sing Koh. "A Survey of Sequential Pattern Mining". In: *Data Science and Pattern Recognition* 1.1 (2017), pp. 54–77.

[176] Philippe Fournier-Viger, Roger Nkambou, and Engelbert Nguifo. "A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems". In: *Mexican International Conference on Artificial Intelligence (MICAI)* (2008), pp. 765–778.

[177] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. "Bayesian nonparametric methods for learning Markov switching processes". In: *IEEE Signal Processing Magazine* 27.6 (2010), pp. 43–54.

[178] Enrique Frias-Martinez and Vijay Karamcheti. "A prediction model for user access sequences". In: *Workshop on Web Mining for Usage Patterns and User Profiles*. 2002.

[179] Nial Friel and Jason Wyse. "Estimating the evidence–a review". In: *Statistica Neerlandica* 66.3 (2012), pp. 288–308.

[180] Jon Froehlich, Joachim Neumann, and Nuria Oliver. "Sensing and Predicting the Pulse of the City through Shared Bicycling". In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2009, pp. 1420–1426.

[181] Sylvia Frühwirth-Schnatter and Sylvia Kaufmann. "Model-based clustering of multiple time series". In: *Journal of Business & Economic Statistics* 26.1 (2008), pp. 78–89.

[182] Wai-Tat Fu and Peter L. T. Pirolli. "SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web". In: *Human-Computer Interaction* 22.4 (2007), pp. 355–412.

[183] Yongjian Fu, Kanwalpreet Sandhu, and Ming-Yi Shih. "Clustering of web users based on access patterns". In: *Workshop on Web Mining*. 1999.

[184] Roland J. Fuchs and George J. Demko. "The Postwar Mobility Transition in Eastern Europe". In: *Geographical Review* 68.2 (1978), pp. 171–182.

[185] Barbara Furletti, Paolo Cintia, Chiara Renso, and Laura Spinsanti. "Inferring Human Activities from GPS Tracks". In: *International Workshop on Urban Computing (SIGKDD)*. 2013, 5:1–5:8.

[186] K. R. Gabriel and J. Neumann. "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv". In: *Quarterly Journal of the Royal Meteorological Society* 88.375 (1962), pp. 90–95.

[187] Mark Gahegan, Monica Wachowicz, Mark Harrower, and Theresa-Marie Rhyne. "The integration of geographic visualization with knowledge discovery in databases and geocomputation". In: *Cartography and Geographic Information Science* 28.1 (2001), pp. 29–44.

[188] Riccardo Gallotti, Armando Bazzani, Sandro Rambaldi, and Marc Barthelemy. "A stochastic model of randomly accelerated walkers for human mobility". In: *Nature communications* 7 (2016), 12600:1–12600:7.

[189] Sebastien Gambs, Marc-Olivier Killijian, and Miguel Nunez del Prado Cortez. "Next Place Prediction Using Mobility Markov Chains". In: *Workshop on Measurement, Privacy, and Mobility*. 2012, 3:1–3:6.

[190] Sebastien Gambs, Marc-Olivier Killijian, and Miguel Nunez del Prado Cortez. "Show Me How You Move and I Will Tell You Who You Are". In: *Workshop on Security and Privacy in GIS and LBS*. 2010, pp. 34–41.

[191] Crispin W Gardiner. *Stochastic methods*. 1985.

[192] Floriana Gargiulo, Maxime Lenormand, Sylvie Huet, and Omar Baqueiro Espinosa. "Commuting Network Models: Getting the Essentials". In: *Journal of Artificial Societies and Social Simulation* 15.2 (2012), 6:1–6:10.

[193] Tini Garske, Hongjie Yu, Zhibin Peng, Min Ye, Hang Zhou, Xiaowen Cheng, Jiabing Wu, and Neil Ferguson. "Travel patterns in China". In: *PloS ONE* 6.2 (2011), e16364.

[194] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. "Statistical methods for eliciting probability distributions". In: *Journal of the American Statistical Association* 100.470 (2005), pp. 680–701.

[195] Wolfgang Gaul and Lars Schmidt-Thieme. "Mining web navigation path fragments". In: *Measurement and Multivariate Analysis* (2000), pp. 249–260.

[196] Andrew Gelman. "Commentary: P values and statistical practice". In: *Epidemiology* 24.1 (2013), pp. 69–72.

[197] Andrew Gelman, Jennifer Hill, and Masanao Yajima. "Why we (usually) don't have to worry about multiple comparisons". In: *Journal of Research on Educational Effectiveness* 5.2 (2012), pp. 189–211.

[198] Mathias Gery and Hatem Haddad. "Evaluation of web usage mining approaches for user's next request prediction". In: *International Workshop on Web Information and Data Management*. 2003, pp. 74–81.

[199] Zoubin Ghahramani, Michael I Jordan, and Padhraic Smyth. "Factorial hidden Markov models". In: *Machine learning* 29.2-3 (1997), pp. 245–273.

[200] Fosca Giannotti, Mirco Nanni, and Dino Pedreschi. "Efficient mining of temporally annotated sequences". In: *SIAM International Conference on Data Mining*. 2006, pp. 348–359.

[201] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. "Unveiling the complexity of human mobility by querying and mining massive trajectory data". In: *The VLDB Journal* 20.5 (2011), pp. 695–719.

[202] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. "Trajectory Pattern Mining". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2007, pp. 330–339.

[203] Walter R Gilks. "Markov chain monte carlo". In: *Encyclopedia of Biostatistics* (2005).

[204] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. "Assessing Data Mining Results via Swap Randomization". In: *Transactions on Knowledge Discovery from Data* 1.3 (2007), pp. 167–176.

[205] Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. "Digital footprinting: Uncovering tourists with user-generated content". In: *Pervasive Computing* 7.4 (2008), pp. 36–43.

[206] Fabien Girardin, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. "Leveraging explicitly disclosed location information to understand tourist dynamics: a case study". In: *Journal of Location Based Services* 2.1 (2008), pp. 41–56.

[207] Edward L Glaeser and Matthew E Kahn. "Sprawl and urban growth". In: *Handbook of regional and urban economics* 4 (2004), pp. 2481–2527.

[208] Sharad Goel, Jake M Hofman, and M Irmak Sirer. "Who Does What on the Web: A Large-Scale Study of Browsing Behavior". In: *International Conference on Web and Social Media (ICWSM)*. 2012, pp. 130–137.

[209] David E Goldberg and Philip Segrest. "Finite Markov chain analysis of genetic algorithms". In: *International Conference on Genetic Algorithms*. 1987, pp. 1–8.

[210] Jacob Goldenberg and Moshe Levy. "Distance Is Not Dead: Social Interaction and Geographical Distance in the Internet Era". In: *arXiv preprint arXiv:0906.3202* (2009).

[211] Scott A. Golder and Bernardo A. Huberman. "Usage patterns of collaborative tagging systems". In: *Journal of Information Science* 32.2 (2006), pp. 198–208.

[212] Sharon Goldwater and Thomas L. Griffiths. "A fully Bayesian approach to unsupervised part-of-speech tagging". In: *Annual Meeting of the Association of Computational Linguistics*. Vol. 45. 1. 2007, pp. 744–751.

[213] Li Gong, Xi Liu, Lun Wu, and Yu Liu. "Inferring trip purposes and uncovering travel patterns from taxi trajectory data". In: *Cartography and Geographic Information Science* 43.2 (2016), pp. 103–114.

[214] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns". In: *Nature* 453.7196 (2008), pp. 779–782.

[215] Steven Goodman. "A dirty dozen: twelve p-value misconceptions". In: *Seminars in hematology*. Vol. 45. 3. 2008, pp. 135–140.

[216] Steven N Goodman. "Multiple comparisons, explained". In: *American journal of epidemiology* 147.9 (1998), pp. 807–812.

[217] Sheila M Gore. "Biostatistics and the medical research council". In: *Medical Research Council News* 35 (1987), pp. 19–20.

[218] Daniel A Griffith. "Modeling spatial autocorrelation in spatial interaction data: empirical evidence from 2002 Germany journey-to-work flows". In: *Journal of Geographical Systems* 11.2 (2009), pp. 117–140.

[219] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. "Tight optimistic estimates for fast subgroup discovery". In: *European Conference on Machine Learning and Knowledge Discovery in Databases*. 2008, pp. 440–456.

[220] Peter D Grünwald. *The minimum description length principle*. 2007.

[221] Sule Gündüz and M Tamer Özsu. "A web page prediction model based on click-stream tree representation of user behavior". In: *International Conference on Knowledge Discovery and Data Mining (KDD)*. 2003, pp. 535–540.

[222] Diansheng Guo, Jin Chen, Alan M MacEachren, and Ke Liao. "A visualization system for space-time and multivariate patterns (vis-stamp)". In: *IEEE Transactions on Visualization and Computer Graphics* 12.6 (2006), pp. 1461–1474.

[223] Rishi Gupta, Ravi Kumar, and Sergei Vassilvitskii. "On Mixtures of Markov Chains". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3441–3449.

[224] Muki Haklay. "Citizen science and volunteered geographic information: Overview and typology of participation". In: *Crowdsourcing Geographic Knowledge*. 2013, pp. 105–122.

[225] Martin Halvey, Mark T Keane, and Barry Smyth. "Mobile web surfing is the same as web surfing". In: *Communications of the ACM* 49.3 (2006), pp. 76–81.

[226] Martin Halvey, Mark T Keane, and Barry Smyth. "Time based segmentation of log data for user navigation prediction in personalization". In: *International Conference on Web Intelligence*. 2005, pp. 636–640.

[227] James D Hamilton. "Analysis of time series subject to changes in regime". In: *Journal of Econometrics* 45.1-2 (1990), pp. 39–70.

[228] Cong Han and Bradley P Carlin. "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review". In: *Journal of the American Statistical Association* 96.455 (2001), pp. 1122–1132.

[229] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. "Frequent pattern mining: current status and future directions". In: *Data Mining and Knowledge Discovery* 15.1 (2007), pp. 55–86.

[230] Jiawei Han, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". In: *ACM Sigmod Record*. Vol. 29. 2. 2000, pp. 1–12.

[231] Edward J Hannan and Barry G Quinn. "The determination of the order of an autoregression". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1979), pp. 190–195.

[232] Samiul Hasan, Christian M Schneider, Satish V Ukkusuri, and Marta C González. "Spatiotemporal patterns of urban human mobility". In: *Journal of Statistical Physics* 151.1-2 (2013), pp. 304–318.

[233] Samiul Hasan, Xianyuan Zhan, and Satish V Ukkusuri. "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media". In: *International Workshop on Urban Computing*. 2013, pp. 1–8.

[234] Timothy J Hatton and Jeffrey G Williamson. *Global migration and the world economy*. 2005.

[235] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. "Geo-located twitter as proxy for global mobility patterns". In: *Cartography and Geographic Information Science* 41.3 (2014), pp. 260–271.

[236] Brian Hayes. "First links in the markov chain". In: *American Scientist* 101.2 (2013), pp. 92–97.

[237] Kingsley E Haynes, Dudley L Poston Jr, and Paul Schnirring. "Intermetropolitan migration in high and low opportunity areas: Indirect tests of the distance and intervening opportunities hypotheses". In: *Economic Geography* 49.1 (1973), pp. 68–73.

[238] M. Heckner, M. Heilemann, and C. Wolff. "Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems". In: *International Conference on Web and Social Media*. 2009, pp. 42–49.

[239] Karsten Held, E Rota Kops, Bernd J Krause, William M Wells, Ron Kikinis, and H-W Muller-Gartner. "Markov random field segmentation of brain MR images". In: *Transactions on Medical Imaging* 16.6 (1997), pp. 878–886.

[240] Norman Herr. *The Sourcebook for Teaching Science, Grades 6-12: Strategies, Activities, and Instructional Resources*. 2008.

[241] Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. "An overview on subgroup discovery: Foundations and applications". In: *Knowledge and Information Systems* 29.3 (2011), pp. 495–525.

[242] Klaus Herrmann. "Modeling the sociological aspects of mobility in ad hoc networks". In: *Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems*. 2003, pp. 128–129.

[243] Sture Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70.

[244] Tobias Hoßfeld, Matthias Hirth, and Phuoc Tran-Gia. "Modeling of Crowdsourcing Platforms and Granularity of Work Organization in Future Internet". In: *International Teletraffic Congress*. 2011, pp. 142–149.

[245] Sahar Hoteit, Guangshuo Chen, Aline Viana, and Marco Fiore. "Filling the gaps: On the completion of sparse call detail records for mobility analysis". In: *ACM Chants* (2016), pp. 45–50.

[246] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. "Information Retrieval in Folksonomies: Search and Ranking". In: *The Semantic Web: Research and Applications*. 2006, pp. 411–426.

[247] Andreas Hotho, Rasmus Ulslev Pedersen, and Michael Wurst. "Ubiquitous data". In: *Ubiquitous Knowledge Discovery*. 2010, pp. 61–74.

[248] Chen-Ming Hsu, Chien-Yu Chen, Baw-Jhiune Liu, Chih-Chang Huang, Min-Hung Laio, Chien-Chieh Lin, and Tzung-Lin Wu. "Identification of hot regions in protein-protein interactions by sequential pattern mining". In: *BMC Bioinformatics* 8.5 (2007), S8.

[249] Wei-jen Hsu, Kashyap Merchant, Haw-wei Shu, Chih-hsin Hsu, and Ahmed Helmy. "Weighted waypoint mobility model and its impact on ad hoc networks". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 9.1 (2005), pp. 59–63.

[250] Yanqing Hu, Jiang Zhang, Di Huan, and Zengru Di. "Toward a Universal Understanding of the Scaling Laws in Human and Animal Mobility". In: *Europhysics Letters* 96.3 (2011), p. 38006.

[251] Lian Huang, Qingquan Li, and Yang Yue. "Activity identification from GPS trajectories using spatial temporal POIs' attractiveness". In: *International Workshop on Location Based Social Networks*. 2010, pp. 27–30.

[252] Wei Huang, Songnian Li, Xintao Liu, and Yifang Ban. "Predicting human mobility with activity changes". In: *International Journal of Geographical Information Science* 29.9 (2015), pp. 1569–1587.

[253] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. "Strong Regularities in World Wide Web Surfing". In: *Science* 280.5360 (1998), pp. 95–97.

[254] Panagiotis G Ipeirotis. "Analyzing the amazon mechanical turk marketplace". In: *XRDS: Crossroads, The ACM Magazine for Students* 17.2 (2010), pp. 16–21.

[255] Robert A Jarrow, David Lando, and Stuart M Turnbull. "A Markov model for the term structure of credit risk spreads". In: *Review of Financial Studies* 10.2 (1997), pp. 481–523.

[256] Harold Jeffreys. *The Theory of Probability*. Third. 1961.

[257] Dennis E Jelinski and Jianguo Wu. "The modifiable areal unit problem and implications for landscape ecology". In: *Landscape Ecology* 11.3 (1996), pp. 129–140.

[258] Hoyoung Jeung, Heng Tao Shen, and Xiaofang Zhou. "Mining trajectory patterns using hidden markov models". In: *Data Warehousing and Knowledge Discovery*. 2007, pp. 470–480.

[259] Bin Jiang, Junjun Yin, and Sijian Zhao. "Characterizing the human mobility pattern in a large street network". In: *Physical Review E* 80.2 (2009), p. 021136.

[260] Shan Jiang, Joseph Ferreira, and Marta C Gonzales. "Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore". In: *IEEE Transactions on Big Data* (2016).

[261] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. "A review of urban computing for mobile phone traces: current methods, challenges and opportunities". In: *International Workshop on Urban Computing*. 2013, pp. 1–9.

[262] George G Judge and Earl Raymond Swanson. "Markov chains: Basic concepts and suggested uses in agricultural economics". In: *Australian Journal of Agricultural Economics* 6.2 (1962), pp. 49–61.

[263] Woo-Sung Jung, Fengzhong Wang, and H Eugene Stanley. "Gravity model in the korean highway". In: *Europhysics Letters* 81.4 (2008), p. 48005.

[264] Ion Juvina and Herre van Oostendorp. "Individual differences and behavioral metrics involved in modeling web navigation". In: *Universal Access in the Information Society* 4.3 (2006), pp. 258–269.

[265] Ion Juvina and Herre van Oostendorp. "Modeling semantic and structural knowledge in Web navigation". In: *Discourse Processes* 45.4-5 (2008), pp. 346–364.

[266] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. "Bicycle cycles and mobility patterns-Exploring and characterizing data from a community bicycle program". In: *arXiv preprint arXiv:0810.4187* (2008).

[267] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system". In: *Pervasive and Mobile Computing* 6.4 (2010), pp. 455–466.

[268] Chaogui Kang, Song Gao, Xing Lin, Yu Xiao, Yihong Yuan, Yu Liu, and Xiujun Ma. "Analyzing and geo-visualizing individual human mobility patterns using mobile call records". In: *International Conference on Geoinformatics*. 2010, pp. 1–7.

[269] Chaogui Kang, Yu Liu, Diansheng Guo, and Kun Qin. "A generalized radiation model for human mobility: Spatial scale, searching direction and trip constraint". In: *PloS ONE* 10.11 (2015), e0143500.

[270] Chaogui Kang, Xiujun Ma, Daoqin Tong, and Yu Liu. "Intra-urban human mobility patterns: An urban morphology perspective". In: *Physica A: Statistical Mechanics and its Applications* 391.4 (2012), pp. 1702–1717.

[271] Ju-Young Kang and Hwan-Seung Yong. "Mining spatio-temporal patterns in trajectory data". In: *Journal of Information Processing Systems* 6.4 (2010), pp. 521–536.

[272] Kevin Karplus, Christian Barrett, and Richard Hughey. "Hidden Markov models for detecting remote protein homologies." In: *Bioinformatics* 14.10 (1998), pp. 846–856.

[273] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795.

[274] Richard W Katz. "On some criteria for estimating the order of a Markov chain". In: *Technometrics* 23.3 (1981), pp. 243–249.

[275] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. "More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk." In: *Americas Conference on Information Systems*. 2011.

[276] Minkyong Kim, David Kotz, and Songkuk Kim. "Extracting a Mobility Model from Real User Traces". In: *IEEE International Conference on Computer Communications*. 2006.

[277] Motoo Kimura. "Some problems of stochastic processes in genetics". In: *The Annals of Mathematical Statistics* (1957), pp. 882–901.

[278] Ross Kindermann and Laurie Snell. *Markov random fields and their applications*. Vol. 1. 1980.

[279] Andrew B King. *Speed up your site: Web site optimization*. 2003.

[280] Muneo Kitajima, Marilyn H. Blackmon, and Peter G. Polson. "A Comprehension-based Model of Web Navigation and Its Application to Web Usability Analysis". In: *People and Computers XIV — Usability or Else!* 2000, pp. 357–373.

[281] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. "The future of crowd work". In: *Conference on Computer Supported Cooperative Work*. 2013, pp. 1301–1318.

[282] Willi Klösgen. "Explora: A Multipattern and Multistrategy Discovery Assistant". In: *Advances in Knowledge Discovery and Data Mining*. 1996, pp. 249–271.

[283] Willi Klösgen and Jan M Zytkow. *Handbook of data mining and knowledge discovery*. 2002.

[284] Christian Körner, Dominik Benz, Markus Strohmaier, Andreas Hotho, and Gerd Stumme. "Stop Thinking, start Tagging - Tag Semantics emerge from Collaborative Verbosity". In: *International World Wide Web Conference*. 2010.

[285] Raymond Kosala and Hendrik Blockeel. "Web Mining Research: A Survey". In: *ACM SIGKDD Explorations Newsletter* 2.1 (2000), pp. 1–15.

[286] Beate Krause, Robert Jäschke, Andreas Hotho, and Gerd Stumme. "Logsonomy-social information retrieval with logdata". In: *Conference on Hypertext and Hypermedia*. 2008, pp. 157–166.

[287] John Kruschke. *Doing Bayesian Data Analysis (Second Edition)*. 2015.

[288] John K Kruschke. "Bayesian estimation supersedes the t-test". In: *Journal of Experimental Psychology: General* 142.2 (2013), pp. 573–603.

[289] John K Kruschke and Torrin M Liddell. "The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". In: *Psychonomic Bulletin & Review* (2017), pp. 1–29.

[290] Jouni Kuha. "AIC and BIC: Comparisons of Assumptions and Performance". In: *Sociological methods & research* 33.2 (2004), pp. 188–229.

[291] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[292] Ravi Kumar, Maithra Raghu, Tamás Sarlós, and Andrew Tomkins. "Linear Additive Markov Processes". In: *International Conference on World Wide Web*. 2017, pp. 411–419.

[293] Vesna Kumbaroska and Pece Mitrevski. "Behavioural–based modelling and analysis of Navigation Patterns across Information Networks". In: *Journal of Emerging Research and Solutions in ICT* 1.2 (2016), pp. 60–74.

[294] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. "Exploring universal patterns in human home-work commuting from mobile phone data". In: *PloS One* 9.6 (2014), e96180.

[295] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura. "Travel Route Recommendation Using Geotags in Photo Sharing Sites". In: *International Conference on Information and Knowledge Management (CIKM)*. 2010, pp. 579–588.

[296] Peter S Kutchukian, David Lou, and Eugene I Shakhnovich. "FOG: Fragment Optimized Growth algorithm for the de novo generation of molecules occupying druglike chemical space". In: *Journal of Chemical Information and Modeling* 49.7 (2009), pp. 1630–1642.

[297] Mei-Po Kwan. "GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns". In: *Geografiska Annaler: Series B, Human Geography* 86.4 (2004), pp. 267–280.

[298] Mei-Po Kwan and Jiyeong Lee. "Geovisualization of human activity patterns using 3D GIS: a time-geographic approach". In: *Spatially integrated social science* 27 (2004).

[299] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *International Conference on Machine Learning*. 2001, pp. 282–289.

[300] Daniel Lamprecht, Markus Strohmaier, Denis Helic, Csongor Nyulas, Tania Tudorache, Natalya F Noy, and Mark A Musen. "Using ontologies to model human navigation behavior in information networks: A study based on wikipedia". In: *Semantic Web* 6.4 (2015), pp. 403–422.

[301] Nada Lavrač, Branko Kavšek, Peter A. Flach, and Ljupco Todorovski. "Subgroup Discovery with CN2-SD". In: *Journal of Machine Learning Research* 5 (2004), pp. 153–188.

[302] Srivatsan Laxman, Vikram Tankasali, and Ryen W White. "Stream prediction using a generative model based on frequent episodes in event sequences". In: *International Conference on Knowledge Discovery and Data Mining*. 2008, pp. 453–461.

[303] Lucien Le Cam. "Maximum likelihood: an introduction". In: *International Statistical Review/Revue Internationale de Statistique* (1990), pp. 153–171.

[304] Everett S. Lee. "A Theory of Migration". In: *Demography* 3.1 (1966), pp. 47–57.

[305] Juhnyoung Lee, Mark Podlaseck, Edith Schonberg, and Robert Hoch. "Visualization and analysis of clickstream data of online stores for understanding web merchandising". In: *Data Mining and Knowledge Discovery* 5.1/2 (2001), pp. 59–84.

[306] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* (2014).

[307] Dennis Leman, Ad Feelders, and Arno J. Knobbe. "Exceptional Model Mining". In: *European Conference on Machine Learning and Knowledge Discovery in Databases*. 2008, pp. 1–16.

[308] Florian Lemmerich. "Novel techniques for efficient and effective subgroup discovery". PhD thesis. University of Würzburg, 2014.

[309] Florian Lemmerich and Martin Atzmueller. "Describing locations using tags and images: explorative pattern mining in social media". In: *Modeling and Mining Ubiquitous Social Media*. 2012, pp. 77–96.

[310] Florian Lemmerich, Martin Becker, and Martin Atzmueller. "Generic Pattern Trees for Exhaustive Exceptional Model Mining". In: *European Conference on Machine Learning and Knowledge Discovery in Databases*. 2012, pp. 277–292.

[311] Florian Lemmerich, Martin Becker, and Frank Puppe. "Difference-Based Estimates for Generalization-Aware Subgroup Discovery". In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2013, pp. 288–303.

[312] Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. "Mining Subgroups with Exceptional Transition Behavior". In: *International Conference on Knowledge Discovery and Data Mining*. 2016.

[313] Maxime Lenormand, Aleix Bassolas, and José J Ramasco. "Systematic comparison of trip distribution laws and models". In: *Journal of Transport Geography* 51 (2016), pp. 158–169.

[314] Maxime Lenormand, Sylvie Huet, Floriana Gargiulo, and Guillaume Deffuant. "A Universal Model of Commuting Networks". In: *PloS One* 7.10 (2012), e45985.

[315] Neal Lesh, Mohammed J Zaki, and Mitsunori Ogihara. "Mining Features for Sequence Classification". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 1999, pp. 342–346.

[316] Mark Levene, José Borges, and George Loizou. "Zipf's law for Web surfers". In: *Knowledge and Information Systems* 3.1 (2001), pp. 120–129.

[317] Fraser Lewis, Adam Butler, and Lucy Gilbert. "A unified approach to model selection using the likelihood ratio test". In: *Methods in Ecology and Evolution* 2.2 (2011), pp. 155–162.

[318] Jiuyong Li, Jixue Liu, Hannu Toivonen, Kenji Satou, Youqiang Sun, and Bingyu Sun. "Discovering Statistically Non-Redundant Subgroups". In: *Knowledge-Based Systems* 67 (2014), pp. 315–327.

[319] Rui Li, Robert Perneczky, Alexander Drzezga, and Stefan Kramer. "Efficient Redundancy Reduced Subgroup Discovery via Quadratic Programming". In: *Journal of Intelligent Information Systems* 44.2 (2015), pp. 271–288.

[320] Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, and Tao Cheng. "Geospatial big data handling theory and methods: A review and research challenges". In: *Journal of Photogrammetry and Remote Sensing (ISPRS)* 115 (2016), pp. 119–133.

[321] Zhao Li and Jeff Tian. "Testing the Suitability of Markov Chains as Web Usage Models". In: *International Conference on Computer Software and Applications*. 2003, pp. 356–361.

[322] Xiao Liang, Jichang Zhao, Li Dong, and Ke Xu. "Unraveling the origin of exponential law in intra-urban human mobility". In: *Scientific Reports* 3 (2013), p. 2983.

[323] Xiao Liang, Xudong Zheng, Weifeng Lv, Tongyu Zhu, and Ke Xu. "The scaling of human mobility by taxis is exponential". In: *Physica A: Statistical Mechanics and its Applications* 391.5 (2012), pp. 2135–2144.

[324] Bing Liu, Bamshad Mobasher, and Olfa Nasraoui. "Web Usage Mining". In: *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. 2011, pp. 527–603.

[325] Hsing Liu, Ying-Hsing Chen, and Jiann-Shing Lih. "Crossover from exponential to power-law scaling for human mobility pattern in urban, suburban and rural areas." In: *European Physical Journal B–Condensed Matter* 88.5 (2015).

[326] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. "Personalized news recommendation based on click behavior". In: *International Conference on Intelligent User Interfaces*. 2010, pp. 31–40.

[327] Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. "A survey on information visualization: Recent advances and challenges". In: *The Visual Computer* 30.12 (2014), pp. 1373–1393.

[328] Yu Liu, Zhengwei Sui, Chaogui Kang, and Yong Gao. "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data". In: *PloS ONE* 9.1 (2014), e86026.

[329] Shiwei Lu, Zhixiang Fang, Xirui Zhang, Shih-Lung Shaw, Ling Yin, Zhiyuan Zhao, and Xiping Yang. "Understanding the representativeness of mobile phone location data in characterizing human mobility indicators". In: *International Journal of Geo-Information* 6.1 (2017).

[330] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson. "Approaching the limit of predictability in human mobility". In: *Scientific Reports* 3, 2923 (2013), p. 2923.

[331] Xin Lu, Linus Bengtsson, and Petter Holme. "Predictability of population displacement after the 2010 haiti earthquake". In: *National Academy of Sciences* 109.29 (2012), pp. 11576–11581.

[332] Xin Lu, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. "Photo2trip: Generating travel routes from geo-tagged photos for trip planning". In: *International Conference on Multimedia*. 2010, pp. 143–152.

[333] Yi Lu and Christie I Ezeife. "Position coded pre-order linked WAP-tree for web log sequential pattern mining". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2003, pp. 337–349.

[334] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. "Mining smart card data for transit riders' travel patterns". In: *Transportation Research Part C: Emerging Technologies* 36 (2013), pp. 1–12.

[335] Alan M MacEachren, Monica Wachowicz, Robert Edsall, Daniel Haug, and Raymon Masters. "Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods". In: *International Journal of Geographical Information Science* 13.4 (1999), pp. 311–334.

[336] David JC MacKay. *Information theory, inference and learning algorithms*. 2003.

[337] Alexander Maedche and Steffen Staab. "Ontology learning for the semantic web". In: *Intelligent Systems* 16.2 (2001), pp. 72–79.

[338] Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. "YAGO3: A knowledge base from multilingual Wikipedias". In: *Biennial Conference on Innovative Data Systems Research*. 2014.

[339] Heikki Mannila and Hannu Toivonen. "Discovering generalized episodes using minimal occurrences." In: *International Conference on Knowledge Discovery in Databases*. Vol. 96. 1996, pp. 146–151.

[340] Gary Marchionini. "Information-seeking strategies of novices using a full-text electronic encyclopedia". In: *Journal of the American Society for Information Science* 40.1 (1989), pp. 54–66.

[341] Andrey A Markov. "An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains". In: *Science in Context* 19.04 (2006), pp. 591–600.

[342] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. "HT06, tagging paper, taxonomy, Flickr, academic article, to read". In: *Conference on Hypertext and Hypermedia*. 2006, pp. 31–40.

[343] Nathan Marz. *Big data: principles and best practices of scalable realtime data systems*. 2013.

[344] A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. "Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows". In: *Physical Review E* 88.2 (2013), p. 022812.

[345] Joan T Matamalas, Manlio De Domenico, and Alex Arenas. "Assessing reliable human mobility patterns from higher order memory in mobile communications". In: *Journal of The Royal Society Interface* 13.121 (2016), p. 20160203.

[346] Adam Mathes. *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. 2004.

[347] Yasuko Matsubara, Yasushi Sakurai, and Christos Faloutsos. "Autoplait: Automatic mining of co-evolving time sequences". In: *International Conference on Management of Data*. 2014, pp. 193–204.

[348] Yasuko Matsubara, Yasushi Sakurai, Christos Faloutsos, Tomoharu Iwata, and Masatoshi Yoshikawa. "Fast mining and forecasting of complex time-stamped events". In: *International Conference on Knowledge Discovery and Data Mining*. 2012, pp. 271–279.

[349] Jean Damascène Mazimpaka and Sabine Timpf. "Trajectory data mining: A review of methods and applications". In: *Journal of Spatial Information Science* 2016.13 (2016), pp. 61–99.

[350] Sharon McDonald and Linda Spencer. "Gender differences in web navigation". In: *Women, Work and Computerization*. 2000, pp. 174–181.

[351] Michael G McNally. "The four-step model". In: *Handbook of Transport Modelling: 2nd Edition*. 2007, pp. 35–53.

[352] Allan DR McQuarrie and Chih-Ling Tsai. *Regression and time series model selection*. 1998.

[353] Sherry E Mead, Victoria A Spaulding, Richard A Sit, Beth Meyer, and Neff Walker. "Effects of age and training on world wide web navigation strategies". In: *Annual Meeting of the Human Factors and Ergonomics Society*. Vol. 41. 1. 1997, pp. 152–156.

[354] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. "Ranking web sites with real user traffic". In: *International Conference on Web Search and Data Mining*. 2008, pp. 65–76.

[355] Mark Meiss, John Duncan, Bruno Gonçalves, José J Ramasco, and Filippo Menczer. "What's in a session: Tracking individual behavior on the web". In: *Conference on Hypertext and Hypermedia*. 2009, pp. 173–182.

[356] Sandro Meloni, Nicola Perra, Alex Arenas, Sergio Gómez, Yamir Moreno, and Alessandro Vespignani. "Modeling human mobility responses to the large-scale spreading of infectious diseases". In: *Scientific Reports* 1 (2011), p. 62.

[357] David R. Millen and Jonathan Feinberg. "Using social tagging to improve social navigation". In: *Workshop on the Social Navigation and Community based Adaptation Technologies*. 2006.

[358] Craig S. Miller and Roger W. Remington. "Modeling information navigation: Implications for information architecture". In: *Human-Computer Interaction* 19.3 (2004), pp. 225–271.

[359] Julien Mineraud, Oleksiy Mazhelis, Xiang Su, and Sasu Tarkoma. "A gap analysis of Internet-of-Things platforms". In: *Computer Communications* 89 (2016), pp. 5–16.

[360] Alan Mislove, Hema Swetha Koppula, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Growth of the flickr social network". In: *Workshop on Online Social Networks*. 2008, pp. 25–30.

[361] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. "Discovery and evaluation of aggregate usage profiles for web personalization". In: *Data mining and knowledge discovery* 6.1 (2002), pp. 61–82.

[362] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. "Using sequential and non-sequential patterns in predictive web usage mining tasks". In: *International Conference on Data Mining*. 2002, pp. 669–672.

[363] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. 1781.

[364] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. "WhereNext: A location predictor on trajectory pattern mining". In: *International Conference on Knowledge Discovery and Data Mining*. 2009, pp. 637–646.

[365] Alan L Montgomery, Shibo Li, Kannan Srinivasan, and John C Liechty. "Modeling online browsing and path analysis using clickstream data". In: *Marketing Science* 23.4 (2004), pp. 579–595.

[366] Carl Mooney and John Roddick. "Sequential pattern mining – approaches and algorithms". In: *ACM Computing Surveys* 45.2 (2013), 19:1–19:39.

[367] Denton E Morrison and Ramon E Henkel. *The significance test controversy: A reader*. 2006.

[368] M Mun, Deborah Estrin, Jeff Burke, and Mark Hansen. "Parsimonious mobility classification using GSM and WiFi traces". In: *Workshop on Embedded Networked Sensors (HotEmNets)*. 2008.

[369] Michal Munk, Jozef Kapusta, and Peter Švec. "Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor". In: *Procedia Computer Science* 1.1 (2010), pp. 2273–2280.

[370] Kevin P Murphy. "Dynamic bayesian networks: Representation, inference and learning". PhD thesis. University of California, Berkeley, 2002.

[371] Mirco Musolesi and Cecilia Mascolo. "Designing mobility models based on social network theory". In: *SIGMOBILE Mobile Computing and Communications Review* 11.3 (2007), pp. 59–70.

[372] Jerzy Neyman and Egon S Pearson. "On the problem of the most efficient tests of statistical hypotheses". In: *Breakthroughs in Statistics*. 1992, pp. 73–108.

[373] Thomas Niebler, Martin Becker, Daniel Zoller, Stephan Doerfel, and Andreas Hotho. "FolkTrails: Interpreting navigation behavior in a social tagging system". In: *International on Conference on Information and Knowledge Management*. 2016, pp. 2311–2316.

[374] Thomas Niebler, Daniel Schlör, Martin Becker, and Andreas Hotho. "Extracting Semantics from unconstrained navigation on wikipedia". English. In: *KI - Künstliche Intelligenz* 30.2 (2016), pp. 163–168.

[375] Thomas Niebler, Philipp Singer, Dominik Benz, Christian Körner, Markus Strohmaier, and Andreas Hotho. "How tagging pragmatics influence tag sense discovery in social annotation systems". In: *Advances in Information Retrieval*. 2013, pp. 86–97.

[376] Allen E Nix and Michael D Vose. "Modeling genetic algorithms with markov chains". In: *Annals of Mathematics and Artificial Intelligence* 5.1 (1992), pp. 79–88.

[377] Anastasios Noulas. "Human urban mobility in location-based social networks: Analysis, models and applications". PhD thesis. University of Cambridge, 2013.

[378] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. "A tale of many cities: Universal patterns in human urban mobility". In: *PLOS ONE* 7.5 (2012), pp. 1–10.

[379] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. "Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining". In: *Journal of Machine Learning Research* 10 (2009), pp. 377–403.

[380] R Nuzzo. "Scientific method: statistical errors". In: *Nature* 506.7487 (2014), pp. 150–152.

[381] Vicki L O'Day and Robin Jeffries. "Orienteering in an information landscape: how information seekers get from here to there". In: *Conference on Human Factors in Computing Systems*. 1993, pp. 438–445.

[382] J Oakley. "Eliciting univariate probability distributions". In: *Rethinking risk measurement and reporting* 1 (2010).

[383] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. "Web page revisitation revisited: implications of a long-term click-stream study of browser usage". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pp. 597–606.

[384] Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Carlos Sarraute, Jorge Brea, and Ignacio Alvarez-Hamelin. "On the regularity of human mobility". In: *Pervasive and Mobile Computing* 33 (2016), pp. 73–90.

[385] Christopher Olston and Ed H Chi. "ScentTrails: Integrating browsing and searching on the Web". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 10.3 (2003), pp. 177–197.

[386] Francois Pachet, Pierre Roy, and Gabriele Barbieri. "Finite-length Markov processes with constraints". In: *Twenty-Second International Joint Conference on Artificial Intelligence*. 2001.

[387] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*. Publication 1999-66. Stanford InfoLab, 1999.

[388] Vasyl Palchykov, Marija Mitrović, Hang-Hyun Jo, Jari Saramäki, and Raj Kumar Pan. "Inferring human mobility using communication patterns". In: *arXiv preprint arXiv:1404.7675* (2014).

[389] Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. "Human mobility modelling: Exploration and preferential return meet the gravity model". In: *Procedia Computer Science* 83 (2016), pp. 934–939.

[390] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. "Mining sequential patterns by pattern-growth: The prefixspan approach". In: *IEEE Transactions on Knowledge and Data Engineering* 16.11 (2004), pp. 1424–1440.

[391] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Hua Zhu. "Mining access patterns efficiently from web logs". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2000, pp. 396–407.

[392] Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. "Smart card data use in public transit: A literature review". In: *Transportation Research Part C: Emerging Technologies* 19.4 (2011), pp. 557–568.

[393] Chengbin Peng, Xiaogang Jin, Ka-Chun Wong, Meixia Shi, and Pietro Liò. "Collective human mobility pattern from taxi trips in urban area". In: *PloS ONE* 7.4 (2012), e34487.

[394] Géraldine Pflieger, Céline Rozenblat, Diana Mok, Barry Wellman, and Juan Carrasco. "Does distance matter in the age of the Internet?" In: *Urban Studies* 47.13 (2010), pp. 2747–2783.

[395] Santi Phithakkitnukoon and Zbigniew Smoreda. "Influence of social relations on human mobility and sociality: a study of social ties in a cellular network". In: *Social Network Analysis and Mining* 6.1 (2016), pp. 1–9.

[396] Santi Phithakkitnukoon, Zbigniew Smoreda, and Patrick Olivier. "Socio-geography of human mobility: A study using longitudinal mobile phone data". In: *PloS ONE* 7.6 (2012), e39253.

[397] Frederick Kin Hing Phoa and Juana Sanchez. "Modeling the browsing behavior of world wide web users". In: *Open Journal of Statistics* 3.2 (2013), pp. 145–154.

[398] Byron J. Pierce, Stanley R. Parkinson, and Norwood Sisson. "Effects of Semantic Similarity, Omission Probability and Number of Alternatives in Computer Menu Search". In: *International Journal of Man-Machine Studies* 37.5 (1992), pp. 653–677.

[399] Juho Piironen and Aki Vehtari. "Comparison of Bayesian predictive methods for model selection". In: *Statistics and Computing* (2016), pp. 1–25.

[400] Michał Piórkowski. "Sampling Urban Mobility Through On-line Repositories of GPS Tracks". In: *1st ACM International Workshop on Hot Topics of Planet-Scale Mobility Measurements*. 2009, 1:1–1:6.

[401] Peter L. T. Pirolli and Stuart K. Card. "Information Foraging". In: *Psychological Review* 106.4 (1999), pp. 643–675.

[402] Peter L. T. Pirolli and James E. Pitkow. "Distributions of surfers' paths through the World Wide Web: Empirical characterizations". In: *World Wide Web* 2.1-2 (1999), pp. 29–45.

[403] Yao Jean Marc Pokou, Philippe Fournier-Viger, and Chadia Moghrabi. "Authorship Attribution Using Small Sets of Frequent Part-of-Speech Skip-grams." In: *International Florida Artificial Intelligence Research Society Conference*. 2016, pp. 86–91.

[404] Jay M Ponte and W Bruce Croft. "Text segmentation by topic". In: *International Conference on Theory and Practice of Digital Libraries*. 1997, pp. 113–125.

[405] Carsten Stig Poulsen. "Mixed Markov and latent Markov modelling applied to brand choice behaviour". In: *International Journal of Research in Marketing* 7.1 (1990), pp. 5–19.

[406] Daniel Preoţiuc-Pietro and Trevor Cohn. "Mining user behaviours: a study of check-in patterns in location based social networks". In: *Annual ACM Web Science Conference*. 2013, pp. 306–315.

[407] Semi Purhonen, Jukka Gronow, and Keijo Rahkonen. "Nordic democracy of taste? Cultural omnivorousness in musical and literary taste preferences in Finland". In: *Poetics* 38.3 (2010), pp. 266–298.

[408] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. 2014.

[409] Ahmed El-Rabbany. *Introduction to GPS: the global positioning system*. 2002.

[410] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[411] Lawrence R Rabiner and Biing-Hwang Juang. "An introduction to hidden Markov models". In: *ASSP Magazine, IEEE* 3.1 (1986), pp. 4–16.

[412] Adrian E Raftery. *Bayes factors and BIC: comment on Weakliem*. Tech. rep. 347. Department of Statistics, University of Washington, 1998.

[413] Adrian E. Raftery. "A Model for High-Order Markov Chains". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 47.3 (1985), pp. 528–539.

[414] A Rajimol and G Raju. "Web access pattern mining–a survey". In: *Data Engineering and Management*. 2012, pp. 24–31.

[415] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. "Are Call Detail Records Biased for Sampling Human Mobility?" In: *Mobile Computing and Communications Review* 16.3 (2012), pp. 33–44.

[416] Ernest George Ravenstein. "The Laws of Migration". In: *Journal of the Royal Statistical Society* 52.2 (1889), pp. 241–305.

[417] Ernest George Ravenstein. "The laws of migration". In: *Journal of the Statistical Society of London* 48.2 (1885), pp. 167–235.

[418] Jun Rekimoto, Takashi Miyaki, and Takaaki Ishizawa. "LifeTag: WiFi-based continuous location logging for life pattern analysis". In: *International Conference on Location and Context Awareness*. 2007, pp. 35–49.

[419] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. "Factorizing Personalized Markov Chains for Next-basket Recommendation". In: *International Conference on World Wide Web*. 2010, pp. 811–820.

[420] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. "On the Levy-walk Nature of Human Mobility". In: *IEEE/ACM Transactions on Networking* 19.3 (2011), pp. 630–643.

[421] Matthew J Roorda and Eric J Miller. "Assessing transportation policy using an activity-based microsimulation model of travel demand". In: *Transportation Specialty Conference*. 2006, TR-136–1 - TR-136–10.

[422] S.M. Ross. *Stochastic processes*. 1996.

[423] Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. "Bayesian t tests for accepting and rejecting the null hypothesis". In: *Psychonomic Bulletin & Review* 16.2 (2009), pp. 225–237.

[424] Richard Royall. *Statistical evidence: a likelihood paradigm*. 1997.

[425] Earl R Ruiter. "Toward a better understanding of the intervening opportunities model". In: *Transportation Research* 1.1 (1967), pp. 47–56.

[426] Narayanan Sadagopan and Jie Li. "Characterizing Typical and Atypical User Sessions in Clickstreams". In: *International Conference on World Wide Web*. 2008, pp. 885–894.

[427] Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. "Tracking human mobility using wifi signals". In: *PloS One* 10.7 (2015), e0130824.

[428] F Sardà, F Maynou, and Ll Talló. "Seasonal and spatial mobility patterns of rose shrimp Aristeus antennatus in the Western Mediterranean: results of a long-term study". In: *Marine Ecology Progress Series* 159 (1997), pp. 133–141.

[429] Ramesh R. Sarukkai. "Link prediction and path analysis using Markov chains". In: *Computer Networks* 33.1 (2000), pp. 377–386.

[430] Lisa Sattenspiel and Klaus Dietz. "A structured epidemic model incorporating geographic mobility among regions". In: *Mathematical Biosciences* 128.1 (1995), pp. 71–91.

[431] Maged El-Sayed, Carolina Ruiz, and Elke A Rundensteiner. "FS-Miner: efficient and incremental mining of frequent sequence patterns in web logs". In: *International Workshop on Web Information and Data Management*. 2004, pp. 128–135.

[432] Aju Thalappillil Scaria, Rose Marie Philip, Robert West, and Jure Leskovec. "The last click: Why users give up information network navigation". In: *International Conference on Web Search and Data Mining*. 2014, pp. 213–222.

[433] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. "Socio-spatial properties of online location-based social networks." In: *International Conference on Web and Social Media* 11 (2011), pp. 329–336.

[434] Stuart Schechter, Murali Krishnan, and Michael D. Smith. "Using path profiles to predict HTTP requests". In: *Computer Networks and ISDN Systems* 30.1 (1998), pp. 457–467.

[435] C M Schneider, V Belik, T Couronné, Z Smoreda, and M C González. "Unravelling daily human mobility motifs". In: *Journal of The Royal Society Interface* 10.84 (2013).

[436] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger. "Understanding Online Social Network Usage from a Network Perspective". In: *Conference on Internet Measurement Conference*. 2009, pp. 35–48.

[437] Steffen Schnitzer, Christoph Rensing, Sebastian Schmidt, Kathrin Borchert, Matthias Hirth, and Phuoc Tran-Gia. "Demands on task recommendation in crowdsourcing platforms - The worker's perspective". In: *Conference of Recommender Systems*. 2015.

[438] Thimo Schulze, Stefan Seedorf, David Geiger, Nicolas Kaufmann, and Martin Schader. "Exploring task properties in crowdsourcing-an empirical study on mechanical turk." In: *European Conference on Information Systems*. 2011.

[439] Gideon Schwarz et al. "Estimating the dimension of a model". In: *The annals of Statistics* 6.2 (1978), pp. 461–464.

[440] Daniel Schweizer, Michael Zehnder, Holger Wache, Hans-Friedrich Witschel, Danilo Zanatta, and Miguel Rodriguez. "Using consumer behavior data to reduce energy consumption in smart homes: Applying machine learning to save energy without lowering comfort of inhabitants". In: *International Conference on Machine Learning and Applications*. 2015, pp. 1123–1129.

[441] Mary C Seiler and Fritz A Seiler. "Numerical recipes in C: the art of scientific computing". In: *Risk Analysis* 9.3 (1989), pp. 415–416.

[442] R. Sen and M. Hansen. "Predicting web users' next access based on log data". In: *Journal of Computational Graphics and Statistics* 12.1 (2003), pp. 143–155.

[443] Claude Elwood Shannon. "A mathematical theory of communication". In: *SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[444] Samuel Sanford Shapiro and Martin B Wilk. "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3/4 (1965), pp. 591–611.

[445] Li Shen and Peter R Stopher. "Review of GPS travel survey and GPS data-processing methods". In: *Transport Reviews* 34.3 (2014), pp. 316–334.

[446] Galit Shmueli et al. "To explain or to predict?" In: *Statistical Science* 25.3 (2010), pp. 289–310.

[447] Ben Shneiderman. "The eyes have it: A task by data type taxonomy for information visualizations". In: *Symposium on Visual Languages*. 1996, pp. 336–343.

[448] Börkur Sigurbjörnsson and Roelof Van Zwol. "Flickr tag recommendation based on collective knowledge". In: *International Conference on World Wide Web*. 2008, pp. 327–336.

[449] Avi Silberschatz and Alexander Tuzhilin. "What makes patterns interesting in knowledge discovery systems". In: *Transactions on Knowledge and Data Engineering* 8 (1996), pp. 970–974.

[450] Filippo Simini, Marta C. Gonzalez, Amos Maritan, and Albert-Laszlo Barabasi. "A universal model for mobility and migration patterns". In: *Nature* 484.7392 (2012), pp. 96–100.

[451] Filippo Simini, Amos Maritan, and Zoltán Néda. "Human mobility in a continuum approach". In: *PloS ONE* 8.3 (2013), e60069.

[452] David W. Sims, Emily J. Southall, Nicolas E. Humphries, Graeme C. Hays, Corey J. A. Bradshaw, Jonathan W. Pitchford, Alex James, Mohammed Z. Ahmed, Andrew S. Brierley, Mark A. Hindell, David Morritt, Michael K. Musyl, David Righton, Emily L. C. Shepard, Victoria J. Wearmouth, Rory P. Wilson, Matthew J. Witt, and Julian D. Metcalfe. "Scaling laws of marine predator search behaviour". In: *Nature* 451.7182 (2008), pp. 1098–1102.

[453] Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. "HypTrails: A bayesian approach for comparing hypotheses about human trails on the web". In: *International Conference on World Wide Web*. 2015, pp. 1003–1013.

[454] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. "Detecting memory and structure in human navigation patterns using markov chain models of varying order". In: *PloS ONE* 9.7 (2014), e102070.

[455] Philipp Singer, Thomas Niebler, Markus Strohmaier, and Andreas Hotho. "Computing semantic relatedness from human navigational paths: A case study on wikipedia". In: *International Journal on Semantic Web and Information Systems* 9.4 (2013), pp. 41–70.

[456] Roger W Sinnott. "Virtues of the haversine". In: *Sky and Telescope* 68.2 (1984), pp. 158–159.

[457] Alina Sirbu, Martin Becker, Saverio Caminiti, Bernard De Baets, Bart Elen, Louise Francis, Pietro Gravino, Andreas Hotho, Stefano Ingarra, Vittorio Loreto, Andrea Molino, Juergen Mueller, Jan Peters, Ferdinando Ricchiuti, Fabio Saracino, Vito D. P. Servedio, Gerd Stumme, Jan Theunis, Francesca Tria, and Joris Van den Bossche. "Participatory patterns in an international air quality monitoring initiative". In: *PLoS ONE* 10.8 (2015), e0136763.

[458] Larry A. Sjaastad. "The costs and returns of human migration". In: *Journal of Political Economy* 70.5 (1962), pp. 80–93.

[459] Aidan Slingsby, Jason Dykes, Jo Wood, and Keith Clarke. "Interactive tag maps and tag clouds for the multiscale exploration of large spatio-temporal datasets". In: *International Conference on Information Visualization*. 2007, pp. 497–504.

[460] Kate A Smith and Alan Ng. "Web page clustering using a self-organizing map of user navigation patterns". In: *Decision Support Systems* 35.2 (2003), pp. 245–256.

[461] Richard L Smith, Jonathan A Tawn, and Stuart G Coles. "Markov chain models for threshold exceedances". In: *Biometrika* 84.2 (1997), pp. 249–268.

[462] Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. "Modelling the scaling properties of human mobility". In: *Nature Physics* 6.10 (2010), pp. 818–823.

[463] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. "Limits of predictability in human mobility". In: *Science* 327.5968 (2010), pp. 1018–1021.

[464] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4 (2002), pp. 583–639.

[465] Ramakrishnan Srikant and Rakesh Agrawal. "Mining quantitative association rules in large relational tables". In: *Acm Sigmod Record*. Vol. 25. 2. 1996, pp. 1–12.

[466] Ramakrishnan Srikant and Rakesh Agrawal. "Mining sequential patterns: Generalizations and performance improvements". In: *Advances in Database Technology (EDBT'96)* (1996), pp. 1–17.

[467] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web usage mining: Discovery and applications of usage patterns from web data". In: *ACM SIGKDD Explorations Newsletter* 1.2 (2000), pp. 12–23.

[468] Steffen Staab and Rudi Studer. *Handbook on ontologies*. 2010.

[469] Hendrik Stange, Thomas Liebig, Dirk Hecker, Gennady Andrienko, and Natalia Andrienko. "Analytical workflow of monitoring human mobility in big event settings using bluetooth". In: *International Workshop on Indoor Spatial Awareness*. 2011, pp. 51–58.

[470] Mario Stanke and Stephan Waack. "Gene prediction with a hidden markov model and a new intron submodel". In: *Bioinformatics* 19.2 (2003), pp. 215–225.

[471] Kay M Stanney and Gavriel Salvendy. "Information visualization; assisting low spatial individuals with information access tasks through the use of visual mediators". In: *Ergonomics* 38.6 (1995), pp. 1184–1198.

[472] Mervyn Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the royal statistical society. Series B (Methodological)* (1974), pp. 111–147.

[473] Samuel A Stouffer. "Intervening opportunities: A theory relating mobility and distance". In: *American Sociological Review* 5.6 (1940), pp. 845–867.

[474] Christopher C Strelioff, James P Crutchfield, and Alfred W Hübler. "Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling". In: *Physical Review E* 76.1 (2007), p. 011106.

[475] Qiang Su and Lu Chen. "A method for discovering clusters of e-commerce interest patterns using click-stream data". In: *Electronic Commerce Research and Applications* 14.1 (2015), pp. 1–13.

[476] Chih Hua Tai, De-Nian Yang, Lung Tsai Lin, and Ming Syan Chen. "Recommending personalized scenic itinerary with geo-tagged photos". In: *International Conference on Multimedia and Expo*. 2008, pp. 1209–1212.

[477] Koh Takeuchi, Ryota Tomioka, Katsuhiko Ishiguro, Akisato Kimura, and Hiroshi Sawada. "Non-negative multiple tensor factorization". In: *International Conference on Data Mining*. 2013, pp. 1199–1204.

[478] Peiyi Tang, Markus P Turkia, and Kyle A Gallivan. "Mining web access patterns with first-occurrence linked WAP-trees." In: *International Conference on Software Engineering and Data Engineering*. 2007, pp. 247–252.

[479] V Tanuja and P Govindarajulu. "A survey on trajectory data mining". In: *International Journal of Computer Science and Security* 10.5 (2016), pp. 195–214.

[480] Linda Tauscher and Saul Greenberg. "How people revisit web pages: Empirical findings and implications for the design of history systems". In: *International Journal of Human-Computer Studies* 47.1 (1997), pp. 97–137.

[481] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. "The perfect search engine is not enough: a study of orienteering behavior in directed search". In: *Conference on Human Factors in Computing Systems*. 2004, pp. 415–422.

[482] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical dirichlet processes". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581.

[483] Isabelle Thomas, Ann Verhetsel, Hakim Hammadou, and Dries Vanhofstraeten. "Distance decay in activity chains analysis. A Belgian case study". In: *Proceedings of the 43rd Congress of the European Regional Science Association: "Peripheries, Centres, and Spatial Development in the New Europe"*. 2003.

[484] Waldo Tobler. "Migration: Ravenstein, Thornthwaite, and Beyond". In: *Urban Geography* 16.4 (1995), pp. 327–343.

[485] John Tolle. "Transactional log analysis: Online catalogs". In: *Annual International Conference on Research and Development in Information Retrieval*. 1983, pp. 147–160.

[486] H. Tong. "Determination of the order of a markov chain by Akaike's information criterion". In: *Journal of Applied Probability* 12.3 (1975), pp. 488–497.

[487] Jameson L Toole, Carlos Herrera-Yaqüe, Christian M Schneider, and Marta C González. "Coupling human mobility and social ties". In: *Journal of The Royal Society Interface* 12.105 (2015), p. 20141128.

[488] Jameson L Toole, Yves-Alexandre de Montjoye, Marta C González, and Alex S Pentland. "Modeling and understanding intrinsic characteristics of human mobility". In: *Social Phenomena*. 2015, pp. 15–35.

[489] Christoph Trattner, Philipp Singer, Denis Helic, and Markus Strohmaier. "Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks". In: *International Conference on Knowledge Management and Knowledge Technologies*. 2012, p. 14.

[490] Michele Trevisiol, Luca Chiarandini, Luca Maria Aiello, and Alejandro Jaimes. "Image ranking based on user browsing behavior". In: *International Conference on Research and Development in Information Retrieval (SIGIR)*. 2012, pp. 445–454.

[491] William Trochim. *Research methods knowledge base, 2nd Edition*. 2001.

[492] Dieter Uckelmann, Mark Harrison, and Florian Michahelles. "An architectural approach towards the future internet of things". In: *Architecting the Internet of Things*. 2011, pp. 1–24.

[493] Matthijs Van Leeuwen and Arno Knobbe. "Diverse Subgroup Set Discovery". In: *Data Mining and Knowledge Discovery* 25.2 (2012), pp. 208–242.

[494] Paul Van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. "Text segmentation and topic tracking on broadcast news via a hidden Markov model approach". In: *International Conference on Spoken Language Processing (ICSLP)*. 1998.

[495] Wolf Vanpaemel. "Constructing informative model priors using hierarchical methods". In: *Journal of Mathematical Psychology* 55.1 (2011), pp. 106–117.

[496] Wolf Vanpaemel. "Prior sensitivity in theory testing: An apologia for the Bayes factor". In: *Journal of Mathematical Psychology* 54.6 (2010), pp. 491–498.

[497] Wolf Vanpaemel and Michael D Lee. "Using priors to formalize theory: Optimal attention and the generalized context model". In: *Psychonomic Bulletin & Review* 19.6 (2012), pp. 1047–1056.

[498] Quang H Vuong. "Likelihood ratio tests for model selection and non-nested hypotheses". In: *Econometrica: Journal of the Econometric Society* (1989), pp. 307–333.

[499] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia". In: *International Conference on Web and Social Media (ICWSM)*. 2015, pp. 454–463.

[500] Nicole Wagner, Khaled Hassanein, and Milena Head. "The impact of age on website usability". In: *Computers in Human Behavior* 37 (2014), pp. 270–282.

[501] Simon Walk, Philipp Singer, Lisette Espın Noboa, Tania Tudorache, Mark A Musen, and Markus Strohmaier. "Understanding How Users Edit Ontologies: Comparing Hypotheses About Four Real-World Projects". In: *International Conference on The Semantic Web (ISWC)*. Vol. 9366. 2015, pp. 551–568.

[502] Simon Walk, Philipp Singer, and Markus Strohmaier. "Sequential Action Patterns in Collaborative Ontology-Engineering Projects: A Case-Study in the Biomedical Domain". In: *International Conference on Information and Knowledge Management*. 2014, pp. 1349–1358.

[503] Simon Walk, Philipp Singer, Markus Strohmaier, Denis Helic, Natalya F. Noy, and Mark A. Musen. "How to apply Markov chains for modeling sequential edit patterns in collaborative ontology-engineering projects". In: *International Journal of Human-Computer Studies* 84 (2015), pp. 51–66.

[504] Hanna M Wallach. "Topic modeling: beyond bag-of-words". In: *International Conference on Machine Learning*. 2006, pp. 977–984.

[505] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. "Human mobility, social ties, and link prediction". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2011, pp. 1100–1108.

[506] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y Zhao. "You Are How You Click: Clickstream Analysis for Sybil Detection." In: *Usenix Security*. Vol. 14. 2013, pp. 241–256.

[507] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. "Unsupervised clickstream clustering for user behavior analysis". In: *Conference on Human Factors in Computing Systems (CHI)*. 2016, pp. 225–236.

[508] Ke Wang, Yabo Xu, and Jeffrey Xu Yu. "Scalable sequential pattern mining for biological sequences". In: *International Conference on Information and Knowledge Management (CIKM)*. 2004, pp. 178–187.

[509] Shaojung Sharon Wang and Michael A Stefanone. "Showing off? Human mobility and the interplay of traits, self-disclosure, and Facebook check-ins". In: *Social Science Computer Review* 31.4 (2013), pp. 437–457.

[510] Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui. "Regularity and conformity: Location prediction using heterogeneous mobility data". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2015, pp. 1275–1284.

[511] Sumio Watanabe. "A widely applicable Bayesian information criterion". In: *Journal of Machine Learning Research* 14.Mar (2013), pp. 867–897.

[512] Christopher Watson. "Trends in world urbanisation". In: *International Conference on Urban Pests*. 1993.

[513] David L Weakliem. "A critique of the Bayesian information criterion for model selection". In: *Sociological Methods & Research* 27.3 (1999), pp. 359–397.

[514] Geoffrey I. Webb. "Layered Critical Values: a Powerful Direct-Adjustment Approach to Discovering Significant Patterns". In: *Machine Learning* 71.2–3 (2008), pp. 307–323.

[515] Geoffrey I. Webb. "OPUS: An Efficient Admissible Algorithm for Unordered Search." In: *Journal of Artificial Intelligence Research* 3.1 (1995), pp. 431–465.

[516] Ingmar Weber, Emilio Zagheni, et al. "Studying inter-national mobility through IP geolocation". In: *International Conference on Web Search and Data Mining*. 2013, pp. 265–274.

[517] Harald Weinreich, Hartmut Obendorf, Eelco Herder, and Matthias Mayer. "Off the Beaten Tracks: Exploring Three Aspects of Web Navigation". In: *International Conference on World Wide Web*. 2006, pp. 133–142.

[518] Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. "The impact of biases in mobile phone ownership on estimates of human mobility". In: *Journal of the Royal Society Interface* 10.81 (2013), p. 20120986.

[519] Robert West and Jure Leskovec. "Human wayfinding in information networks". In: *International Conference on World Wide Web*. 2012, pp. 619–628.

[520] Robert West, Joelle Pineau, and Doina Precup. "Wikispeedia: an online game for inferring semantic distances between concepts". In: *Proceedings of the 21st International Joint Conference on Artifical intelligence*. 2009, pp. 1598–1603.

[521] Ruud Wetzels, Darja Tutschkow, Conor Dolan, Sophie van der Sluis, Gilles Dutilh, and Eric-Jan Wagenmakers. "A Bayesian test for the hot hand phenomenon". In: *Journal of Mathematical Psychology* 72 (2016), pp. 200–209.

[522] Michelle J. White. "Sex Differences in Urban Commuting Patterns". In: *The American Economic Review* 76.2 (1986), pp. 368–372.

[523] Ryen W White and Steven M Drucker. "Investigating behavioral variability in web search". In: *International Conference on World Wide Web*. 2007, pp. 21–30.

[524] Ryen W White and Jeff Huang. "Assessing the scenic route: measuring the value of search trails in web logs". In: *Conference on Research and Development in Information Retrieval*. 2010, pp. 587–594.

[525] James A Whittaker and Michael G Thomason. "A Markov chain model for statistical software testing". In: *IEEE Transactions on Software Engineering* 20.10 (1994), pp. 812–824.

[526] Peter Widhalm, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C González. "Discovering urban activity patterns in cell phone data". In: *Transportation* 42.4 (2015), pp. 597–623.

[527] Sarah E Wiehe, Aaron E Carroll, Gilbert C Liu, Kelly L Haberkorn, Shawn C Hoch, Jeffery S Wilson, and J Dennis Fortenberry. "Using GPS-enabled cell phones to track the travel patterns of adolescents". In: *International Journal of Health Geographics* 7.1 (2008), p. 22.

[528] Samuel S Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62.

[529] David A Williams. "Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures". In: *Biometrics* (1970), pp. 23–32.

[530] Michael J Wills. "A flexible gravity-opportunities model for trip distribution". In: *Transportation Research Part B: Methodological* 20.2 (1986), pp. 89–111.

[531] Jean Wolf. "Using GPS Data Loggers To Replace Travel Diaries In the Collection of Travel Data". PhD thesis. Georgia Institute of Technology, 2000.

[532] Stefan Wrobel. "An Algorithm for Multi-relational Discovery of Subgroups". In: *European Symposion on Principles of Data Mining and Knowledge Discovery (PKDD)*. 1997, pp. 78–87.

[533] Fang Wu and Bernardo A Huberman. "Novelty and collective attention". In: *Proceedings of the National Academy of Sciences* 104.45 (2007), pp. 17599–17601.

[534] Jianguo Wu. "Scale and scaling: a cross-disciplinary perspective". In: *Key topics in landscape ecology*. 2007.

[535] L Wu, Y Zhi, Z Sui, and Y Liu. "Intra-urban human mobility and activity transition: evidence from social media check-in data". In: *PLoS One* 9.5 (2014), e97010.

[536] Ellery Wulczyn. "Wikipedia Navigation Vectors". In: *figshare* (2017).

[537] Ellery Wulczyn and Dario Taraborelli. "Wikipedia Clickstream". In: *figshare* (2015).

[538] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. "Inferring social ties between users with human location history". In: *Journal of Ambient Intelligence and Humanized Computing* 5.1 (2014), pp. 3–19.

[539] Kexin Xie, Ke Deng, and Xiaofang Zhou. "From trajectories to activities: a spatio-temporal join approach". In: *Proceedings of the 2009 International Workshop on Location Based Social Networks*. 2009, pp. 25–32.

[540] Zhengzheng Xing, Jian Pei, and Eamonn Keogh. "A Brief Survey on Sequence Classification". In: *ACM SIGKDD Explorations Newsletter* 12.1 (2010), pp. 40–48.

[541] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Umeshwar Dayal. "From user access patterns to dynamic hypertext linking". In: *Computer Networks and ISDN Systems* 28.7-11 (1996), pp. 1007–1014.

[542] Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal, and Jie Liu. "Fast App Launching for Mobile Devices Using Predictive User Context". In: *International Conference on Mobile Systems, Applications, and Services*. 2012, pp. 113–126.

[543] Xiao-Yong Yan, Xiao-Pu Han, Bing-Hong Wang, and Tao Zhou. "Diversity of individual mobility patterns and emergence of aggregated scaling laws". In: *Scientific Reports* 3 (2013), p. 2678.

[544] Xiao-Yong Yan, Chen Zhao, Ying Fan, Zengru Di, and Wen-Xu Wang. "Universal predictability of mobility patterns in cities". In: *Journal of The Royal Society Interface* 11.100 (2014), p. 20140834.

[545] Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. "Finding progression stages in time-evolving event sequences". In: *International Conference on World Wide Web*. 2014, pp. 783–794.

[546] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. "Mining individual life pattern based on location history". In: *International Conference on Mobile Data Management: Systems, Services and Middleware (MDM)*. 2009, pp. 1–10.

[547] Josh Jia-Ching Ying, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S Tseng. "Semantic trajectory mining for location prediction". In: *International Conference on Advances in Geographic Information Systems*. 2011, pp. 34–43.

[548] Jungkeun Yoon, Brian D. Noble, Mingyan Liu, and Minkyong Kim. "Building Realistic Mobility Models from Coarse-grained Traces". In: *International Conference on Mobile Systems, Applications and Services*. 2006, pp. 177–190.

[549] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. "Task recommendation in crowdsourcing systems". In: *Workshop on Crowdsourcing and Data Mining*. 2012, pp. 22–26.

[550] Osmar R Zaiane, Man Xin, and Jiawei Han. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs". In: *Research and Technology Advances in Digital Libraries (ADL)*. 1998, pp. 19–29.

[551] Mohammed J Zaki. "SPADE: An Efficient Algorithm for Mining Frequent Sequences". In: *Machine Learning* 42.1-2 (2001), pp. 31–60.

[552] Mohammed Javeed Zaki. "Scalable algorithms for association mining". In: *IEEE Transactions on Knowledge and Data Engineering* 12.3 (2000), pp. 372–390.

[553] Christina Zarcadoolas, Mercedes Blanco, John F Boyer, and Andrew Pleasant. "Unweaving the Web: an exploratory study of low-literate adults' navigation skills on the World Wide Web". In: *Journal of Health Communication* 7.4 (2002), pp. 309–324.

[554] Wilbur Zelinsky. "The Hypothesis of the Mobility Transition". In: *Geographical Review* 61.2 (1971), pp. 219–249.

[555] Chao Zhang, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta. "Splitter: Mining fine-grained sequential patterns in semantic trajectories". In: *Proceedings of the VLDB Endowment (PVLDB)* 7.9 (2014), pp. 769–780.

[556] Zengbin Zhang, Lin Zhou, Xiaohan Zhao, Gang Wang, Yu Su, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. "On the validity of geosocial mobility traces". In: *Workshop on Hot Topics in Networks*. 2013, 11:1–11:7.

[557] Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao, and Sasu Tarkoma. "Explaining the power-law distribution of human mobility through transportation modality decomposition". In: *Scientific Reports* 5 (2015), p. 9136.

[558] Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. "Urban Human Mobility Data Mining: An Overview". In: *Big Data* (2016), pp. 1911–1920.

[559] Vincent W. Zheng, Yu Zheng, Xing Xie, and Qiang Yang. "Collaborative Location and Activity Recommendations with GPS History Data". In: *International Conference on World Wide Web*. 2010, pp. 1029–1038.

[560] Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. "Mining Travel Patterns from Geotagged Photos". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), 56:1–56:18.

[561] Yu Zheng. "Trajectory data mining: an overview". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 6.3 (2015), p. 29.

[562] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. "Understanding Mobility Based on GPS Data". In: *International Conference on Ubiquitous Computing*. 2008, pp. 312–321.

[563] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. "U-Air: When Urban Air Quality Inference Meets Big Data". In: *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 2013, pp. 1436–1444.

[564] Yu Zheng and Xing Xie. "Learning travel recommendations from user-generated GPS traces". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.1 (2011), 2:1–2:29.

[565] Yu Zheng, Xing Xie, and Wei-Ying Ma. "GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory." In: *IEEE Data Engineering Bulletin* 33.2 (2010), pp. 32–39.

[566] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. "Mining Interesting Locations and Travel Sequences from GPS Trajectories". In: *International Conference on World Wide Web*. 2009, pp. 791–800.

[567] Jianhan Zhu, Jun Hong, and John G Hughes. "Using markov chains for link prediction in adaptive web sites". In: *Soft-Ware 2002: Computing in an Imperfect World*. 2002, pp. 60–73.

[568] Sabrina Ziebarth, Irene-Angelica Chounta, and Heinz Ulrich Hoppe. "Resource Access Patterns in Exam Preparation Activities". In: *Design for Teaching and Learning in a Networked World*. 2015, pp. 497–502.

[569] Albrecht Zimmermann and Luc De Raedt. "Cluster-Grouping: From Subgroup Discovery to Clustering". In: *Machine Learning* 77.1 (2009), pp. 125–159.

[570] George Kingsley Zipf. "The $(P_1P_2)/D$ Hypothesis: On the Intercity Movement of Persons". In: *American Sociological Review* 11.6 (1946), pp. 677–686.

[571] I. Zukerman, D. W. Albrecht, and A. E. Nicholson. "Predicting Users' Requests on the WWW". In: *International Conference on User Modeling (UM)*. 1999, pp. 275–284.