MixedTrails: Bayesian Hypothesis Comparison on Heterogeneous Sequential Data

Martin Becker · Florian Lemmerich · Philipp Singer · Markus Strohmaier · Andreas Hotho

July 2017

Abstract Sequential traces of user data are frequently observed online and offline, e.g., as sequences of visited websites or as sequences of locations captured by GPS. However, understanding factors explaining the production of sequence data is a challenging task, especially since the data generation is often not homogeneous. For example, navigation behavior might change in different phases of browsing a website, or movement behavior may vary between groups of users. In this work, we tackle this task and propose *MixedTrails*, a Bayesian approach for comparing the plausibility of hypotheses regarding the generative processes of heterogeneous sequence data. Each hypothesis is derived from existing literature, theory or intuition and represents a belief about transition probabilities between a set of states that can vary between groups of observed transitions. For example, when trying to understand human movement in a city and given some observed data, a hypothesis assuming tourists to be more likely to move towards points of interests than locals, can be shown to be more plausible than a hypothesis assuming the opposite. Our approach incorporates such hypotheses as Bayesian priors in a generative mixed transition Markov chain model, and compares their plausibility utilizing Bayes factors. We discuss analytical and approximate inference methods for calculating the marginal likelihoods for Bayes factors, give guidance on interpreting the results, and illustrate our approach with several experiments on synthetic and empirical data from Wikipedia and Flickr. Thus, this work enables a novel kind of analysis for studying sequential data in many application areas.

1 Introduction

Sequential data over a discrete state space emerges in a variety of settings, including sequences of weather conditions [21], DNA sequences [54], Web navigation [44], or real-world travel sequences over locations [26,42]. Understanding the underlying processes that generate such sequences can be useful for a wide range of applications, such as improving network

Martin Becker and Andreas Hotho
Data Mining and Information Retrieval Group, University of Würzburg, Würzburg, Germany
E-mail: {becker, hotho}@informatik.uni-wuerzburg.de

Florian Lemmerich and Philipp Singer and Markus Strohmaier
GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
E-mail: {florian.lemmerich, philipp.singer, markus.strohmaier}@gesis.org

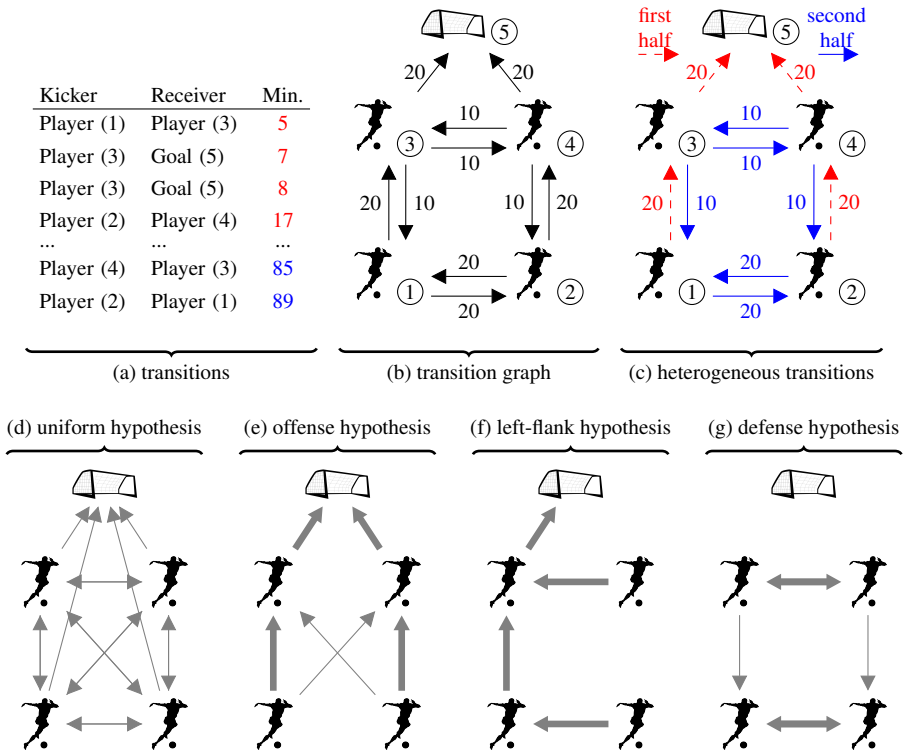


Fig. 1 Illustrating example. In this figure, we show an illustrating soccer example: We are interested in a team's strategy in a specific game. We start with data on passes and shots (a). Using a simple Markov chain, we can model these as transitions between states (b). The previously proposed HypTrails 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 appears to be a lot more plausible. MixedTrails enables the comparison of such hypotheses on heterogeneous data.

structures, predicting user clicks on websites, or enhancing recommendations, and has been a challenge and complex research objective in our community for years.

Background. The (first-order) Markov chain model is one of the most elementary, yet versatile, models for transitions between sequence states. It follows the Markovian assumption that the probability of the next state in a sequence depends exclusively on the current state. Building upon this basic model, the recently proposed HypTrails approach [52] allows to compare hypotheses about sequential data, where hypotheses represent beliefs in state transition probabilities that are derived from existing literature, theory, or intuition with regard to the respective application domain. For example, by studying Wikipedia user data, we found that the hypothesis that users preferably click on links at the top of a page provides a better explanation of user navigation than a hypothesis that assumes transitions to semantically similar pages [16].

Figure 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 [52], researchers can then express and compare hypotheses (d-g) about pass sequences by specifying different beliefs in transitions. For instance, a simple hypothesis could state 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. Given the transition data in Figure 1(a), the approach would rank the uniform hypothesis (d) as the most plausible one, as it resembles the overall data (b) best.

Problem and objectives. Simple Markov chain models, and consequently also the HypTrails approach, assume homogeneous sequence data. As such, they cannot take into account behavior stemming from several underlying processes. For instance, research on mobility has found starkly differing user groups such as tourists and locals [38], and there exist different phases of Web navigation with distinct patterns [63]. Reconsidering our soccer scenario of Figure 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 1 (e) and (g)) could provide a better explanation for our data, but cannot be compared with existing approaches.

To that end, our goal in this paper 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 in the context of smooth behavioral shifts or uncertain classifiers on the other hand.

Contributions. In this paper, 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 belief matrices and 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.

Overall, we present a novel approach for specifying and comparing hypotheses about heterogeneous sequence data that involve varying behavior in parts of the observed transitions. This will enable researchers and practitioners to perform a new kind of analysis on such data.

2 Background and Notation

In this section, we shortly introduce HypTrails [52] which our approach *MixedTrails* builds on and cover its building blocks, i.e., Markov chains and Bayesian model comparison. An overview of all important notations used throughout this article can be found in Appendix B.

Markov Chain Model. A Markov chain model M_{MC} [34,53] is a random process modeling a sequence of random variables $X_1, X_2, \ldots, X_l$ as transitions between a set of states

$S = \{s_1, s_2, \ldots, s_n\}$. In this work, we focus on first-order Markov chains, which describe a memoryless process meaning that the next state s_j in a sequence only depends on the current one s_{i_τ}, i.e.: $\Pr(X_{\tau+1} = s_j | X_1 = s_{i_1}, \ldots, X_\tau = s_{i_\tau}) = \Pr(X_{\tau+1} = s_j | X_\tau = s_i) = \theta_{i,j}$. The parameters of a Markov chain model M_{MC} are the transition probabilities $\theta_{i,j}$ between states s_i and s_j represented by a transition matrix $\boldsymbol{\theta} = (\theta_{i,j})$. As the model is stochastic, each row of the transition matrix sums to 1, i.e. $\forall i : \sum_j \theta_{i,j} = 1$. Thus, given a transition dataset D with transition counts $n_{i,j}$ between states s_i and s_j, the likelihood of observing these transitions is:

$$\Pr(D|\boldsymbol{\theta}, M_{MC}) = \prod_{t_k \in D} \theta_{i_k, j_k} = \prod_{s_i, s_j \in S} \theta_{i,j}^{n_{i,j}}$$

In the Bayesian setting, *prior* beliefs in transition probabilities are updated after observing data. The choice of the prior distribution over the transition probabilities $\boldsymbol{\theta}$ is crucial for calculating the posterior or examining the marginal likelihood. As detailed below, in this paper, we employ independent Dirichlet priors for each state i, i.e., $\boldsymbol{\theta}_{s_i} \sim Dir(\boldsymbol{\alpha}_{s_i})$ where $\boldsymbol{\alpha}_{s_i}$ are the parameters of the Dirichlet distribution.

Bayesian model comparison. Given a set of models $\{M_1, \ldots, M_m\}$ and some data D, Bayesian model comparison establishes a partial order on the set of models $M_i \sqsubseteq M_j \sqsubseteq M_k$ based on the marginal likelihood $Pr(D|M_i)$ of the data D given each model M_i. The marginal likelihood represents the plausibility of the model. The strength of evidence in favor of a model M_i compared to a model M_j can then be formally measured by a Bayes factor $B_{i,j}$ [33]. It represents the factor by which the prior odds in favor of one of two compared models change after seeing the data (posterior odds):

$$\underbrace{\frac{\Pr(M_i|D)}{\Pr(M_j|D)}}_{\text{posterior odds}} = B_{i,j} \cdot \underbrace{\frac{\Pr(M_i)}{\Pr(M_j)}}_{\text{prior odds}}, \text{ with } \underbrace{B_{i,j} = \frac{\Pr(D|M_i)}{\Pr(D|M_j)}}_{\text{Bayes factor}} \tag{1}$$

Bayes factors can also be utilized for conducting Bayesian hypotheses comparison if priors encode theory-induced hypotheses as advocated in [35,50,60]; we use this throughout this article. For judging significance, we refer to Kass and Raftery's interpretation table [33].

HypTrails. HypTrails [52] operationalizes Bayesian model comparison for hypotheses on Markov chain models M_{MC} in order to establish a partial order $\sqsubseteq$ on a set of hypotheses $\mathscr{H} = \{H_1, \ldots, H_n\}$ based on their plausibility given the data. A hypothesis H is expressed as a prior probability distribution $\Pr(\boldsymbol{\theta}|H, M_{MC})$ over all instances of transition probability matrices $\boldsymbol{\theta}$, which is required to compute the marginal likelihood used by the Bayes factor:

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

If we now assume all hypotheses to be equally likely a-priori (as often done in Bayesian model comparison), the Bayes factor directly implies the posterior probabilities, cf. the derivation of Bayes factor in [33].

To express a hypothesis H about transition probabilities $\Pr(\boldsymbol{\theta}|H, M_{MC})$, HypTrails uses Dirichlet priors, i.e., for each state s_i an individual Dirichlet prior $Dir(\boldsymbol{\alpha}_{s_i})$ is specified which defines beliefs about transition probabilities from that state s_i to all other states. The parameters $\boldsymbol{\alpha}_{s_i}$ are vectors of positive numbers, i.e., $\boldsymbol{\alpha}_{s_i} = (\alpha_{i,1}, \ldots \alpha_{i,n}),\ \alpha_{i,j} \in \mathbb{R}^+$. That is, given a hypothesis H and a fixed number of imaginary (pseudo) transitions originating from state s_i, $\alpha_{i,j} - 1$ denotes the expected number of observed transitions from state s_i to state s_j.

The process of expressing a belief as a formal hypothesis H and transforming it into prior parameters (pseudocounts) is called *elicitation*. For elicitation, HypTrails assumes a two step process: First, a transition probability distribution $\phi_{s_i} = (\phi_{i,1}, \ldots, \phi_{i,n})$ is specified for each state s_i, resulting in a stochastic transition matrix $\phi = (\phi_{i,j})$. Then, a concentration factor $\kappa \in \mathbb{N}_0^+$ is set in order to derive the hyperparameters $\alpha_{i,j}$ by calculating: $\alpha = \kappa \cdot \phi + 1$, where κ is proportional to the amount of pseudocounts we assign to each state [1]. 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. [52]). The higher we set the concentration factor κ, the more we "believe" in our hypothesis, i.e., we get higher marginal likelihood values if we are correct, but we are also penalized more if the hypothesis is off. The lower we set the concentration factor κ, the more "slack" we allow for our hypothesis, i.e., we are not as strongly penalized for errors, but we also cannot reach large marginal likelihood values if we are correct. Note that 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, or certainty in the belief. HypTrails somewhat alleviates this issue by formalizing the suggestion from [33] to compare several prior instantiations by using a range of concentration factors $\kappa = \{\kappa_1, \kappa_2, ...\}$. This allows for a structured and detailed comparison of hypotheses. Also see Section 3.5 for a discussion on alternative approaches.

3 MixedTrails: Bayesian Hypothises Comparison in Heterogeneous Sequence Data

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 3.1) and explain how hypotheses for heterogeneous sequence data are structured. Then, we introduce the Mixed Transition Markov Chain (MTMC) model (Section 3.2) — an extension of the basic Markov chain model — that allows to model such heterogeneous data. By incorporating hypotheses as elicited priors over the model parameters (Section 3.3), we can utilize Bayesian model comparison to make relative judgements about the plausibility of the given hypotheses. Finally, we derive an approach for model inference (Section 3.4) and give guidelines for interpreting the results (Section 3.5). For illustrative purposes, we will refer to the soccer example visualized in Figure 1.

3.1 Problem statement and approach

The goal of this paper is to compare hypotheses about *heterogeneous* sequence data. That is, considering a dataset of transitions $D = \{t_1, \ldots, t_m\}$ between a set of states $S = \{s_1, \ldots s_n\}$, we want to establish a partial ordering $\sqsubseteq$ on a set of given hypotheses $\mathscr{H} = \{H_1, H_2, \ldots\}$ that express *how* the observed transitions may have been generated. Extending HypTrails [52], we focus on transitions generated by several independent processes.

Hypotheses We describe a heterogeneous hypothesis $H = (\gamma, \phi)$ by two components. First, the *group assignment probabilities* γ associate each transition $t \in D$ in the dataset D with a

[1] Note that this is a slightly simplified version of the original Trial Roulette method from the HypTrails paper [52] regarding two aspects. First, we do not distribute chips but multiply by a concentration factor which is effectively equivalent and easier to compute. Second, we assume in this paper the same weight in each row of the Markov chain which makes formulating hypotheses and interpreting results easier. However, these simplifications are not required and reverting them is straightforward.

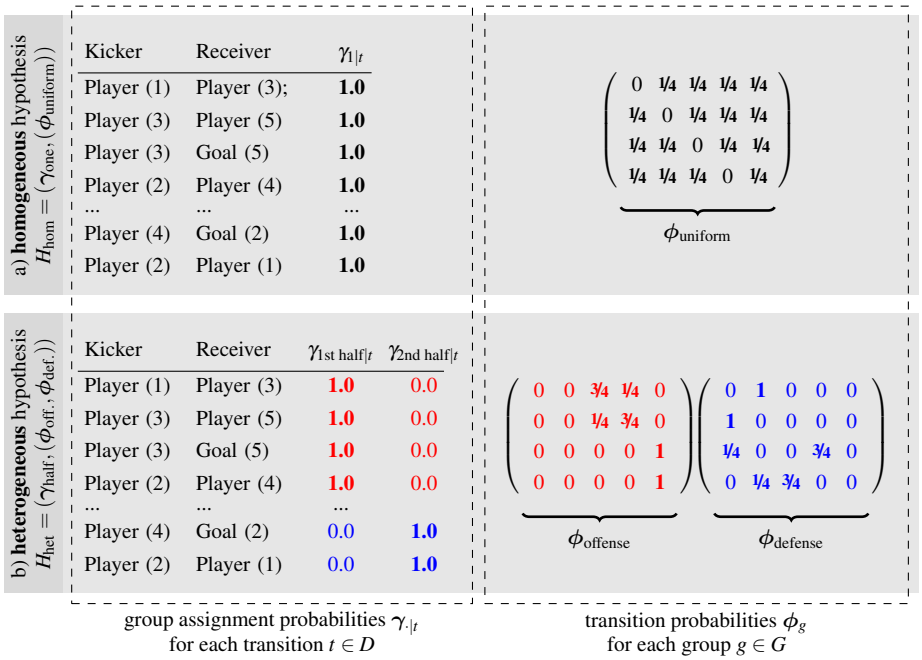


Fig. 2 Hypotheses for heterogeneous sequence data. In MixedTrails, we formulate hypotheses about heterogeneous sequence data. E.g., in the soccer example, we define two hypotheses: The homogeneous hypothesis H_{hom} (a) assumes that players just randomly pass the ball around; the heterogeneous hypothesis H_{het} (b) assumes an offensive strategy in the first half of the game and a defensive strategy in the second half, cf. Figure 1. This is formalized based on two components: *group assignment probabilities* γ, i.e., probability distributions over a set of groups for each transition, and a belief matrix of *group transition probabilities* ϕ_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.

probability distribution γ_t over a set of *groups* $G = \{g_1, \ldots, g_o\}$ defined by the corresponding hypothesis. We write all group assignment probabilities for a hypothesis as $\gamma = \{\gamma_t | t \in D\}$, with $\gamma_t = \{\gamma_{g|t} | g \in G\}$. Here, $\gamma_{g|t}$ is the probability that transition t belongs to group g. Second, the *group transition probabilities* ϕ 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 $\phi = (\phi_1, \ldots \phi_o)$, with $\phi_g = (\phi_{i,j|g})$, 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 transition 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 3.3) of the generative model MTMC (see Section 3.2) and calculating the marginal likelihood (Bayesian evidence).

Example For illustration, consider again the soccer game example from Figure 1. In the following, we specify two hypotheses for this scenario: a homogeneous one H_{hom} 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 ϕ_{uniform} as shown in Figure 2(a). Consequently, the corresponding group assignment probabilities γ_{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 (ϕ_{offense} and ϕ_{defense}), one for each halftime. Then, we assign each transition to the group (halftime) it belongs to via γ_{half}. Transitions are assigned to half-times without uncertainty, thus, the probabilities used are either 0 or 1. The resulting hypothesis is defined as $H_{\text{het}} = (\gamma_{\text{half}}, (\phi_{\text{offense}}, \phi_{\text{defense}}))$ as visualized in Figure 2(b). Now, our approach MixedTrails determines the marginal likelihood $\Pr(D|H_{\text{hom}})$ and $\Pr(D|H_{\text{het}})$ as a measure for the plausibility of the data under a hypothesis. Since $\Pr(D|H_{\text{het}}) > \Pr(D|H_{\text{hom}})$ (as demonstrated later, Section 3.5), we assert that explaining the data as a result of an offensive strategy in the first half of the game and a defensive strategy in the second half (H_{het}) is a more plausible hypothesis given the observed data.

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 can 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, 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 modelling higher order Markovian processes, i.e., by defining n^x groups (where n 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.Thus, even though there are some limitations and possible extensions (cf. Section 5), all in all, MixedTrails provides a very flexible and easy to use framework to model a very large and possibly complex set of hypotheses.

3.2 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 dataset to a group $g \in G = \{g_1, ..., g_o\}$, which is drawn from an individual categorical distribution with parameters $\gamma_t = (\gamma_{g_1|t}, \ldots, \gamma_{g_o|t})$, 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 $\theta_{s_i|g} = (\theta_{i,1|g}, \ldots, \theta_{i,n|g})$ 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 $\boldsymbol{\alpha}_{s_i|g} = (\alpha_{i,1|g}, \ldots, \alpha_{i,n|g})$. For shorter notation, we write the set of transition probabilities over all groups as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_o)$ and the set of transition probabilities over all states in a group as $\boldsymbol{\theta}_g = (\boldsymbol{\theta}_{s_1|g}, \ldots, \boldsymbol{\theta}_{s_n|g})$. Similarly, we denote the set of all hyperparameters for all Markov models, i.e., all Dirichlet parameters, as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_o)$, and the set of all hyperparameters for a single group as $\boldsymbol{\alpha}_g = (\boldsymbol{\alpha}_{s_1|g}, \ldots, \boldsymbol{\alpha}_{s_n|g})$. 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 $\text{dst}_k \sim Cat(\boldsymbol{\theta}_{\text{src}_k|z_k})$.

3.3 Eliciting priors from hypotheses

As mentioned in Section 3.1, MixedTrails elicits hypotheses as Bayesian priors for the MTMC model (see Section 3.2), which takes two independent sets 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 3.1, the parameters $\boldsymbol{\alpha}$ of the Dirichlet prior need to be *elicited* from the transition probabilities $\boldsymbol{\phi}$ defined by the hypothesis.

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, using the notion of *pseudo-observations*, cf. [52]. That is, for each group $g \in G$ and each state s_i, we set the Dirichlet parameters starting from an uninformed proto-prior and add κ transitions distributed as the hypothesis suggests for this group via $\boldsymbol{\phi}_g$. Formally, this is:

$$\alpha_{i,j|g} = \kappa \cdot \phi_{i,j|g} + 1. \tag{2}$$

Here, the number of pseudo-observations κ (also called concentration factor) reflects the strength of belief in the respective hypothesis. Different settings for the concentration parameter 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.

For example, consider the heterogeneous hypothesis $H_{\text{het}} = (\boldsymbol{\gamma}_{\text{half}}, (\boldsymbol{\phi}_{\text{offense}}, \boldsymbol{\phi}_{\text{defense}}))$ from Figure 2(b). 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 (2). 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' ($\gamma_{g|t_k} \cdot \gamma_{g'|t_k}$). This probability is then used as a weight for the respective belief matrix $\phi_{g'}$. Formally:

$$\alpha_{i,j|g} = \kappa \cdot \left(\frac{1}{Z_i} \cdot \sum_{t_k \in D} \left(\sum_{g' \in G} \gamma_{g|t_k} \cdot \gamma_{g'|t_k} \cdot \phi_{i,j|g'} \right) \right) + 1, \tag{3}$$

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 (2).

3.4 Model Inference

For comparing the plausibility of heterogeneous hypotheses, in MixedTrails, we determine the evidence (marginal likelihood) of the data under a hypothesis (cf. Section 3.1) based on the MTMC model as introduced in Section 3.2. 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 can be written as an integral over all parameter settings $\boldsymbol{\theta}$:

$$\Pr(D|H) = \int \underbrace{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})}_{\text{likelihood}} \underbrace{\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})}_{\text{prior}} \, d\boldsymbol{\theta} \tag{4}$$

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. 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 sampling.

Analytical solution. When ignoring the group assignment probabilities $\boldsymbol{\gamma}$ in Equation (4), the marginal likelihood of the MTMC model is equivalent to the homogeneous Markov chain model for which an analytical solution exists [53]. However, in our setting, we need to aggregate over all possible instantiations $\omega \in \Omega$ of group assignments. 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 [53] for details). Overall, we obtain the following formula (for an in-depth derivation see Appendix A):

$$\Pr(D|H) = \sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{s_i \in S} \frac{B(\boldsymbol{n}_{s_i|g,\omega} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})}, \tag{5}$$

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 (5) simplifies to:

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

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 [4].

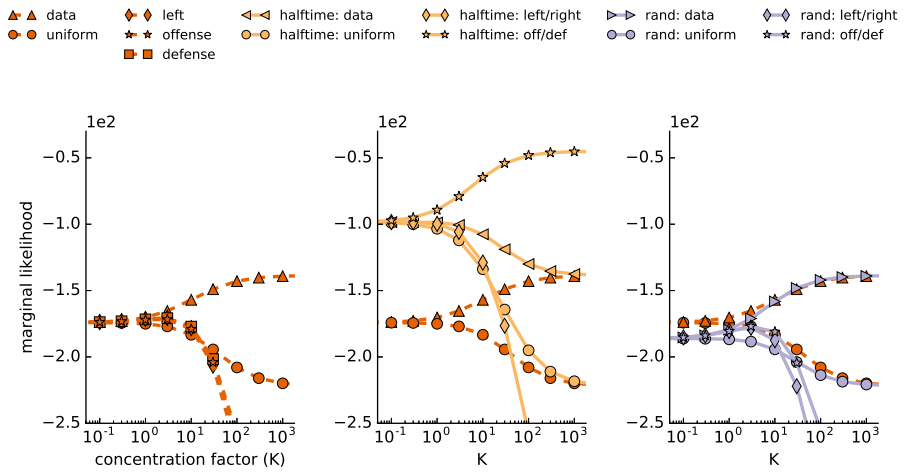
Approximation. For the general, probabilistic case, calculating the marginal likelihood of an MTMC model analytically with Equation (5) is computationally intractable. Therefore, we show how we can efficiently approximate it by direct sampling. According to the formula, the overall marginal likelihood 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 3.2). 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:

$$\Pr(D|H) \approx \frac{1}{|\Omega'|} \sum_{\omega \in \Omega'} \underbrace{\prod_{g \in G} \prod_{s_i \in S} \frac{B(\boldsymbol{n}_{s_i|g,\omega} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})}}_{\Pr(D|\boldsymbol{\alpha},\omega)}$$

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 in only a few hours on a regular desktop machine.

3.5 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 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,



(a) Homogeneous hypotheses (b) Heterogeneous hyp. based on a split according to halftimes (c) Heterogeneous hypotheses based on a random split

Fig. 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 for increasing strengths of belief κ. 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 groups leads to less plausible explanations (c).

given the observed transitions from Figure 1(a-c), we aim to compare the plausibility of the different beliefs in transition probabilities from Figure 1(d-g) utilizing the marginal likelihood as elaborated in Section 3.4. 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 half-time of the passes/shots, and (c) a completely random group assignment. The hypotheses are formulated analogously to the examples covered in Section 3.3. The results are shown in Figure 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; each line represents one given hypothesis; solid lines refer to heterogeneous hypotheses and dashed lines to homogeneous hypotheses. In general, higher values of the marginal likelihood 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*. This means that even if one hypothesis outperforms all other hypotheses under consideration, this does not necessarily mean that it models the data well. However, the goal of this paper 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 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 with 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. [16]). 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 [33]. This means that conclusions should only be drawn for sections of the marginal likelihood plots where the values are farther apart than 10. In these cases, the change of the posterior is to be interpreted as "decisive". Consequently, we only draw conclusions from such decisive results in this manuscript.

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 3(b)). 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 hypotheses.

Results on homogeneous hypotheses. Figure 3(a) shows results for the homogeneous hypotheses. As expected, the data "hypothesis", which uses 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 analogously be obtained by the HypTrails approach [52].

Results on heterogeneous hypotheses: the split. However, our approach MixedTrails enables us to also compare more fine-grained, *heterogeneous* hypotheses. Figure 3(b) features four heterogeneous hypotheses (solid lines) that assign the data deterministically into two groups, i.e., the first and the second half-time. 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 curve. For higher values of κ, the marginal likelihoods diverge: The offense/defense hypothesis, i.e., in the first half-time players behave as the offense belief suggests, and in the second halftime as the defense belief suggests (see Figure 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 better."* 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 3(c) shows the same four hypotheses, but assigns transition to two groups randomly (*rand*). 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. Overall, these examples give a broad overview of possible MixedTrails results. More examples are covered in Section 4.

4 Experiments

In this section, we demonstrate the applicability and benefits of our approach with experiments on synthetic and real-world datasets. An open source Python implementation[2] as well as the datasets are freely available[3]. Conclusions from the experimental results drawn in the text rely on results that are "decisive" with respect to the established interpretation table given in [33], cf. Section 3.5.

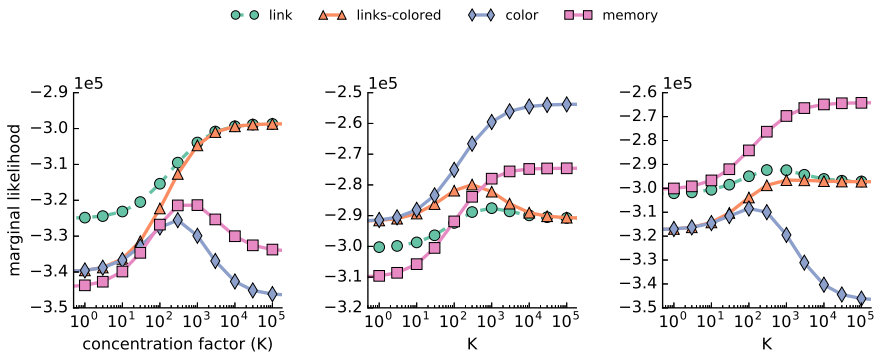
4.1 Synthetic Datasets: Deterministic Group Assignments

First, 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 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 [3] with 100 nodes and 10 edges for each new node. Each node has a

[2] `http://dmir.org/mixedtrails`

[3] The scripts for generating the synthetic data are included in the code, the Wikispeedia data set (cf. Section 4.3) is accessible online and the Flickr data (cf. Section 4.4) is available via e-mail to Martin Becker.



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

Fig. 4 Synthetic data results. 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. For details on interpreting the plots, see Section 4.1.

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 the mechanism for the respective dataset. She stops after ten steps. 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 transition 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 $H_{\text{link}} = (\gamma_{\text{link}}, \phi_{link})$, which expresses the belief that all transition are randomly chosen from the link network, thus $\phi_{link} = (\boldsymbol{\theta}_{link})$; secondly, the color-preference hypothesis $H_{\text{color}} = (\gamma_{\text{color}}, \phi_{\text{color}})$ 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: $\phi_{\text{color}} = (\boldsymbol{\theta}_{\text{red}}, \boldsymbol{\theta}_{\text{blue}})$; and thirdly, the memory hypothesis $H_{\text{mem}} = (\gamma_{\text{mem}}, \phi_{\text{mem}})$ 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: $\phi_{\text{mem}} = (\boldsymbol{\theta}_{\text{red}}, \boldsymbol{\theta}_{\text{blue}}, \boldsymbol{\theta}_{\text{link}})$. To illustrate how our approach copes with groups that introduce unnecessary complexity, we add a fourth hypothesis $H_{\text{link-color}} = (\gamma_{\text{color}}, (\boldsymbol{\theta}_{link}, \boldsymbol{\theta}_{link}))$ 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. For the link dataset D_{link}, see Figure 4(a), we find that the

homogeneous hypothesis reflects the data very well and thus achieves the 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. Both heterogeneous hypotheses show increasing ML for increasing κ at first since the hypotheses carry some information, i.e., which network links are contained in the data. With increasing concentration, 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 includes transitions to red and blue nodes in more equal proportions for each source state.

Next, we consider the color dataset D_{color}, see Figure 4(b). 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 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 a correct ordering on the given hypotheses based on the processes used to generate the data.

Finally, we consider the memory dataset D_{mem}. Here we can observe that — as expected — the memory hypothesis H_{mem} performs best for all values of κ. The 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 $H_{\text{link-colord}}$ 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.

4.2 Synthetic Datasets: 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 4.4.

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 "mixed 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

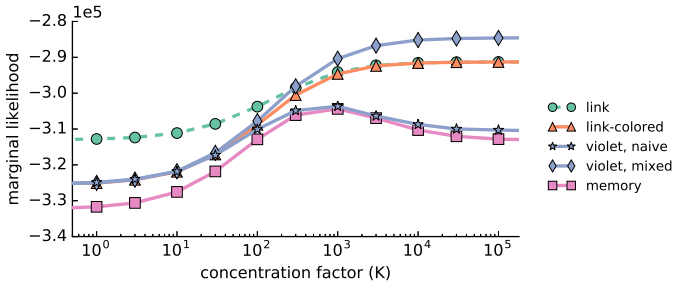


Fig. 5 Probabilistic group assignments on synthetic data. The *violet, mixed* hypothesis, using probabilistic group assignment probabilities, is the most plausible one for increasing concentration factors as it directly models the processes underlying the data. The *violet, naive* hypothesis illustrates 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 4.2.

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.

Hypotheses. As hypotheses, we define H_{link}, $H_{\text{link-colored}}$ and H_{memory} analogously to Section 4.1. In addition, we introduce a hypothesis $H_{\text{violet}} = (\gamma_{\text{violet}}, \phi_{\text{violet}})$ specifically tailored to violet walkers. Thus, we define the group dependent transition probabilities as $\phi_{\text{violet}} = (\boldsymbol{\theta}_{\text{red}}, \boldsymbol{\theta}_{\text{blue}})$. 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: $\gamma_{g|t_w} = (s_w, 1 - s_w)$, where t_w represents a transition from a specific walker w.

Results. The results are shown in Figure 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 3.3 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 from Figure 5, 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 3.3 is an integral part of the approach.

4.3 The Wikispeedia dataset

Wikispeedia [63] is a game in which players aim to find a short 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, the authors 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"[4]). A corresponding dataset [64] is freely available[5]. It consists of the plain text of each article, the link network, and a set of click sequences (including back clicks) created by humans playing the game. Like West et al. [63], 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 a mean length of 5.6.

Hypotheses. To investigate the hypothesis from [63], we consider two transition probability matrices: ϕ_{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 ϕ_{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 sublinear scaling to the tf values. [6] For comparison, we additionally consider the link matrix ϕ_{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 similar to Figure 2: $H_{\text{link}} = (\gamma_{\text{one}}, \phi_{link})$, $H_{\text{deg}} = (\gamma_{\text{one}}, \phi_{deg})$, $H_{\text{sim}} = (\gamma_{\text{one}}, \phi_{sim})$. Furthermore, $H_{\text{deg,sim}}$ 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. In particular, 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 γ_{phases}. The heterogeneous hypotheses are then defined as: $H_{\text{deg,sim}} = (\gamma_{\text{phases}}, (\phi_{deg}, \phi_{sim}))$ and $H_{\text{deg,sim}} = (\gamma_{\text{phases}}, (\phi_{sim}, \phi_{deg}))$ assuming the degree and the similarity transition probability matrices to explain the "initial phase", respectively.

Results. Figure 6 shows that, as literature hypothesized, the heterogeneous hypothesis

[4] available at `schools-wikipedia.org` (version of 2007)

[5] `https://snap.stanford.edu/data/wikispeedia.html`

[6] Differing from our approach, West et al. [64] 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.

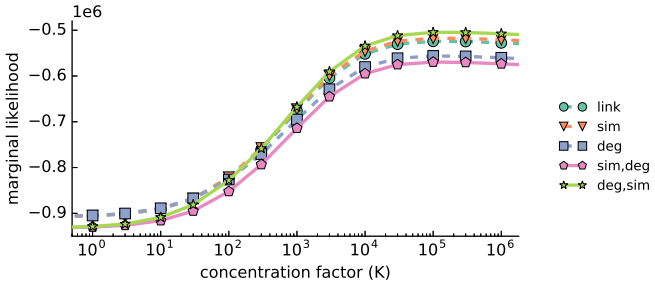


Fig. 6 Wikispeedia results. For the game Wikispeedia, players try to quickly navigate from one article to another using the underlying link structure of Wikipedia. One hypothesis (*deg,sim*) is that players will first navigate to articles with a large degree, and then "home-in" on their target using similarity based navigation. The graph shows the results of modeling this heterogeneous hypothesis using the MTMC model by splitting each click sequence after 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 (at concentration factor $\kappa = 0$) is not inherently advantageous. For details, see Section 4.3. Note that, while the differences visually appear to be marginal in the plot, they are decisive (cf. Section 3.5).

$H_{\text{deg,sim}}$ 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. 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 ML 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.

Overall, this example demonstrates the applicability of MixedTrails to a real world scenario. We also see that a more fine-grained hypothesis may explain observed sequential data better than using a single, overly general hypothesis.

4.4 The Flickr dataset

Finally, we investigate geo-spatial trails obtained from the photo-sharing platform Flickr[7].

Dataset. As data in this setting we employ a dataset from previous work [38]. It contains all Flickr photos from the years 2010 to 2014 with geo-spatial information (i.e., latitude and longitude) at street-level accuracy in Manhattan. We mapped each photo according to its geo-location to one of the 288 census tracts (administrative units) that we use as state space in our model (see also [22]). Then, for each user, we built a sequence of different tracts she has taken photos at (excluding self-transitions). Thus, we know the start and end date for each user sequence. The final dataset contains 386,981 transitions overall.

Hypotheses. In previous research [5], we found that a combination of spatial proximity to the current state and to points of interest (PoIs) is the best explanation for the transitions

[7] `https://www.flickr.com/`

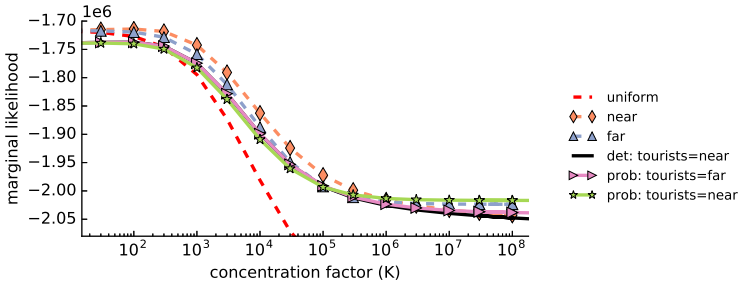


Fig. 7 Flickr results. We model the navigation behavior between tracts in Manhattan based on photo trails on the social photo-sharing platform Flickr. Overall, we have not found hypotheses explaining the data well as indicated by the strongly decreasing marginal likelihoods. However, those we evaluated 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, MTMC allows for modelling uncertain classification of tourists which covers the underlying processes better than a deterministic group assignment (*det: tourists=near*).

of Flickr users. However, in some settings, proximity to the current state is more relevant, while in others, larger, spatial variances lead to better results. Accordingly, we built two different transition probability matrices that we call ϕ_{near} and ϕ_{far}, which feature different parametrizations of the proximity/POI hypothesis. In particular, we set the influence radius of PoIs to $400m$ and the standard deviation of the proximity factor to $2.5km$ (ϕ_{near}) and $5.0km$ (ϕ_{far}). For details, we refer to [5].

In this paper, we aim to extend the previous study by taking into account whether a user is a *tourist* or a *local*. To classify users as tourists or locals, we use the time difference between their first and their last photo in the data, cf. [15]. In that regard, we consider different group assignments: a) a baseline γ_{one} that puts all transitions into one group, b) deterministic grouping γ_{det} by defining tourists as users with a trail duration of 21 or less days, and c) 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 γ_{prob}.

We combine these three group assignments and transition probability matrices to form five hypotheses: (i) $H_{\text{near}} = (\gamma_{\text{one}}, \phi_{\text{near}})$, (ii) $H_{\text{far}} = (\gamma_{\text{one}}, \phi_{\text{far}})$, (iii) $H_{\text{det: tourist=near}} = (\gamma_{\text{det}}, (\phi_{\text{near}}, \phi_{\text{far}}))$, (iv) $H_{\text{prob: tourist=near}} = (\gamma_{\text{prob}}, (\phi_{\text{near}}, \phi_{\text{far}}))$, and (v) $H_{\text{prob: tourist=far}} = (\gamma_{\text{prob}}, (\phi_{\text{far}}, \phi_{\text{near}}))$. 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 tourist 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 shows the results. Obviously, the uniform hypothesis is substantially less plausible than all proximity/PoI-based hypotheses. Among the latter, we see that for smaller concentration 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, i.e., by using probabilistic group assignments and expressesing the belief that tourists take their next photo at a more near-by location with a close PoI while locals choose locations with higher distance more often. By contrast, a deterministic split γ_{det} does not cover the uncertainty of classifying tourists and locals.

Overall, this example illustrates how MixedTrails with probabilistic group assignments enables more fine-grained analyses of sequential data.

5 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.

Top-down vs. bottom-up. MixedTrails is a *top-down* approach — also called a *deductive* approach in certain contexts [58,31,8] — meaning that it takes a set of hypotheses based on *ideas* and *theories* from the application domain as input and compares them using some 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., [67,17,37]. Note, that these methods usually do not yield directly interpretable results. If they do (e.g., [17]), they can be used in a *bottom-up* setting — sometimes also called an *inductive* [58,8] setting —, which takes the opposite approach than MixedTrails: bottom-up methods use observations to extract patterns or regularities from which *new* hypotheses or theories can be derived. The same is true for other specialized approaches, e.g., for segmentation, labeling, or clustering [46,9,62,59,19]: while they can be facilitated or extended to *uncover new* interesting heterogeneous patterns by examining their latent structures, applying them results in a bottom-up approach which *yields new* hypotheses instead of comparing existing ones.

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 (e.g., stickiness [19,65]), as for example possible in Markov switching processes such as the Hidden Markov model. 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 modelled. 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 a 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 likelihood ratio test). However, by also investigating lower concentration factors, we obtain additional information on the quality of the group assignments (cf. Section 3.5). 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 in 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 3.4 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 [4] may be useful for larger datasets. We have also experimented with other methods for approximating the marginal likelihood such as [14], 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 random, some of them would appear to be statistically significantly better than the baseline, cf. [7]. 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 [27,23]. Nonetheless, exploring the challenges of multiple comparisons will be an issue that we will study more in-depth in future work.

6 Related work

In this section, we provide an overview of related work on Markov chain models, their applications (focusing on the Web context), and respective extensions. Further discussions and elaborations of related work have been captured throughout the course of this work.

(First-order) Markov chain models have been first introduced by A. A. Markov in 1913 [39]. Several adaptations of the Markov chain model have been proposed, such as the so-called higher-order Markov chain models [53], Hidden Markov models [48], or mixtures of Markov models [47,55]. Historically, Markov chain models have been applied in many diverse settings due to their simplicity and generality. Examples include textual data modeling [39], weather data modeling [21], C.E. Shannon's take on information theory [51], or the application of modeling Web navigation leading to the PageRank algorithm [44]. A historical summary of Markov chain models and their applications can be found in [30].

In this work, we have focused on applications in the Web context.[8] This line of research has been tackled in a multitude of studies. For example, early work by Catledge and Pitkow [11] investigated human navigation on WWW pages. Subsequent studies have further demonstrated that Web navigation is guided by certain regularities [32,45,13,63,61]. Prominent theories are, for example, that humans prefer to transition between semantically similar concepts [12,10,63], or the so-called *information foraging theory* [45,13] postulating that human behavior in an information environment on the Web is guided by *information scent*. Among many others, further studies have focused on sequence prediction [37,1,43,18], the recommendation of travel routes [15], search trails [66], or the study of music sequences [2, 17].

Motivated by this large array of hypotheses about sequential behavior, HypTrails [52] was proposed for comparing the plausibility of hypotheses on sequential data. MixedTrails, as introduced in this paper, builds on HypTrails and addresses one of its main issues, namely allowing to model and compare hypotheses about *heterogeneous* sequence data.

The model we employ in this paper is related to previously proposed extensions of Markov chains. One prominent example are mixture models [55]. In that direction, the *Mixed Markov chain* model has been studied by Poulsen [47] in the context of customer behavior segmentation. Poulsen, however, defines group memberships on a sequence level, not on a transition level sacrificing some of the expressiveness incorporated into MixedTrails. Similar group memberships are used by Rendle et al. who factorize Markov chains [49] and by Gupta et al. who reconstruct mixtures of Markov chains [28]. To our knowledge, these models have not been employed for the comparison of hypotheses so far. Additionally, the expressiveness of these models is limited, i.e., not all group assignments of the hypotheses featured in this paper could be expressed with these models. Another set of Markov chain extensions related to our approach is the class of Markov switching processes [48,19] which model observations dependent on hidden Markovian dependency structures. Some classic instances in this class are the Hidden Markov Model (HMM) [48], the Factorial HMM [24] or the the Auto-Regressive HMM [29] (also see [41] for further extensions). There are also methods based on, or related to, these methods which are used for prediction, clustering or segmentation [18,20,40,25], including, e.g., Bayesian nonparametric methods [57,19] 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 (as opposed to comparing *existing* ones). While, e.g, Hidden Markov models have been applied to compare streaky behavior with a baseline model [65], to best of

[8] Note that our approach can also be applied to very different settings in a straight-forward manner.

the authors knowledge, there are no general approaches to apply Markov switching processes in a top-down manner in the context of background data. For a more detailed discussion on the difference between top-down and bottom-up approaches please see Section 5.

Statistical methods for comparing the fits of different Markov chain models have been summarized in [53] and include likelihood ratio tests, information-theoretic AIC, BIC or DIC approaches, or the Bayes factor. These methods have been utilized, e.g., for comparing the fit of nested, higher-order Markov chain models that relax the basic assumption of the Markovian property and allow for longer memory fits. In this work, we focus on comparing fits by using marginal likelihoods and Bayes factors [56]; these have the advantage of an automatic built-in Occam's razor balancing the goodness of fit with complexity [33]. Additionally, instead of only using a flat Dirichlet prior, we also utilize the sensitivity of the marginal likelihood on the prior for comparing theory-induced hypotheses within the Bayesian framework—as advocated, e.g., in [50, 60, 35]—following the HypTrails approach elaborated in [52]. 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. This is in line with a general trend towards Bayesian methods for data analysis [36, 6].

7 Conclusions

In this paper, we have introduced MixedTrails, 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 we gave application examples with real-world data. Overall, this work enables a novel kind of analysis for studying sequence data in many application areas.

In the future, we may explore our method in additional real-world applications, such as investigating the movement of (groups of) Flickr users, cf. [5], or studying groups of editors in Wikipedia. Furthermore, more complex priors or hierarchical models may allow for more powerful ways of expressing hypotheses.

8 Acknowledgements

This work was partially funded by the BMBF project Kallimachos and the DFG German Science Fund research projects PoSTs II and p2map.

References

1. Akinori Asahara, Kishiko Maruyama, Akiko Sato, and Kouichi Seto. Pedestrian-movement prediction based on mixed markov-chain model. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 25–33. ACM, 2011.
2. Claudio Baccigalupo and Enric Plaza. Case-based sequential ordering of songs for playlist recommendation. In *European Conference on Case-Based Reasoning*, pages 286–300. Springer, 2006.
3. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

4. 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 *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 17–18, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
5. 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 Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu, editors, *Social Informatics*, pages 227–244, Cham, 2015. Springer International Publishing.
6. Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1026–1034, 2014.
7. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.
8. Amy Blackstone. *Sociological Inquiry Principles: Qualitative and Quantitative Methods*. Flat World Knowledge, Irvington, NY, USA, 2012.
9. David M Blei and Pedro J Moreno. Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–348. ACM, 2001.
10. Duncan P. Brumby and Andrew Howes. Good enough but i'll just check: Web-page search as attentional refocusing. In *International Conference on Cognitive Modeling*, pages 46–51, 2004.
11. Lara D Catledge and James E Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.
12. Matthew Chalmers, Kerry Rodden, and Dominique Brodbeck. The order of things: activity-centred information access. *Computer Networks and ISDN Systems*, 30(1):359–367, 1998.
13. Ed H. Chi, Peter L. T. Pirolli, Kim Chen, and James Pitkow. Using information scent to model user information needs and actions and the web. In *Conference on Human Factors in Computing Systems*, pages 490–497. ACM, 2001.
14. Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.
15. Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, and Cong Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 35–44, New York, NY, USA, 2010. ACM.
16. Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. What makes a link successful on wikipedia? In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 917–926, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.
17. 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*, 2016.
18. Flavio Figueiredo, Bruno Ribeiro, Jussara M Almeida, and Christos Faloutsos. Tribeflow: Mining & predicting user trajectories. In *Proceedings of the 25th International Conference on World Wide Web*, pages 695–706. International World Wide Web Conferences Steering Committee, 2016.
19. Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Bayesian nonparametric methods for learning markov switching processes. *IEEE Signal Processing Magazine*, 27(6):43–54, 2010.
20. Sylvia Frühwirth-Schnatter and Sylvia Kaufmann. Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89, 2008.
21. KR Gabriel and J Neumann. A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375):90–95, 1962.
22. Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and i will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL '10, pages 34–41, New York, NY, USA, 2010. ACM.
23. Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.
24. Zoubin Ghahramani, Michael I Jordan, and Padhraic Smyth. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.
25. Sharon Goldwater and Tom Griffiths. A fully bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-association for computational linguistics*, volume 45, page 744. Citeseer, 2007.

26. Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
27. Steven N Goodman. Multiple comparisons, explained. *American journal of epidemiology*, 147(9):807–812, 1998.
28. Rishi Gupta, Ravi Kumar, and Sergei Vassilvitskii. On mixtures of markov chains. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 3441–3449. Curran Associates, Inc., 2016.
29. James D Hamilton. Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1-2):39–70, 1990.
30. Brian Hayes et al. First links in the markov chain. *American Scientist*, 101(2):92–97, 2013.
31. Norman Herr. The sourcebook for teaching science strategies, activities and instructional resources. *San Francisco, CA: John*, 2008.
32. Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, Mar 1998.
33. Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.
34. John G Kemeny, James Laurie Snell, et al. *Finite markov chains*, volume 356. van Nostrand Princeton, NJ, 1960.
35. John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, Boston, 2015.
36. John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.
37. 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*, pages 453–461. ACM, 2008.
38. Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Mining subgroups with exceptional transition behavior. In *KDD '16: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
39. Andrey A Markov. An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(04):591–600, 2006. Originally published in 1913.
40. Yasuko Matsubara, Yasushi Sakurai, and Christos Faloutsos. Autoplait: Automatic mining of co-evolving time sequences. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 193–204. ACM, 2014.
41. Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.
42. Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLOS ONE*, 7(5):1–10, 05 2012.
43. Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 1038–1043, Washington, DC, USA, 2012. IEEE Computer Society.
44. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
45. Peter L. T. Pirolli and Stuart K. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.
46. Jay M Ponte and W Bruce Croft. Text segmentation by topic. In *International Conference on Theory and Practice of Digital Libraries*, pages 113–125. Springer, 1997.
47. Carsten Stig Poulsen. Mixed markov and latent markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, 7(1):5–19, 1990.
48. Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.
49. Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *International Conference on World wide web*, pages 811–820. ACM, 2010.
50. Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, 16(2):225–237, 2009.
51. Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
52. Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Intl. Conference on World Wide Web*, 2015.

53. Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PLOS ONE*, 9(7):e102070, 2014.
54. Lloyd M Smith, Jane Z Sanders, Robert J Kaiser, Peter Hughes, Chris Dodd, Charles R Connell, Cheryl Heiner, SB Kent, and Leroy E Hood. Fluorescence detection in automated dna sequence analysis. *Nature*, 321(6071):674–679, 1985.
55. Richard L Smith, Jonathan A Tawn, and Stuart G Coles. Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268, 1997.
56. Christopher C. Strelioff, James P. Crutchfield, and Alfred W. Hübler. Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Phys. Rev. E*, 76:011106, Jul 2007.
57. Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
58. William Trochim. *Research methods knowledge base, 2nd Edition*. Atomic Dog Publishing, Cincinnati, OH, USA, 2001.
59. 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 *ICSLP*, 1998.
60. Wolf Vanpaemel. Prior sensitivity in theory testing: An apologia for the bayes factor. *Journal of Mathematical Psychology*, 54(6):491–498, 2010.
61. 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 Conference on Information & Knowledge Management*. ACM, 2014.
62. Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.
63. Robert West and Jure Leskovec. Human wayfinding in information networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 619–628. ACM, 2012.
64. 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*, pages 1598–1603, 2009.
65. Ruud Wetzels, Darja Tutschkow, Conor Dolan, Sophie van der Sluis, Gilles Dutilh, and Eric-Jan Wagenmakers. A bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, 72:200–209, 2016.
66. 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*, pages 587–594. ACM, 2010.
67. Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd international conference on World wide web*, pages 783–794. ACM, 2014.

A Derivation of the marginal likelihood of MTMC

Given the generative process from Section 3.2 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:

$$\begin{aligned}\Pr(D|H) &= \int \underbrace{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})}_{\text{likelihood}} \underbrace{\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})}_{\text{prior}} \, d\boldsymbol{\theta} \\ &= \int \underbrace{\prod_{t_k \in D} \sum_{g \in G} \gamma_{g|t_k} \theta_{i_k,j_k|g}}_{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})} \underbrace{\prod_{g \in G} \Pr(\boldsymbol{\theta}_g|\boldsymbol{\alpha}_g)}_{\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})} \prod_{g \in G} d\boldsymbol{\theta}_g \end{aligned} \tag{6}$$

To solve this integral we take a similar path as in the homogeneous case (cf. [52]). 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:

$$\begin{aligned}\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma}) &= \prod_{t_k \in D} \sum_{g \in G} \gamma_{g|t} \theta_{i_k,j_k|g} \\ &= \sum_{\substack{\omega \in \Omega \\ \Omega = \{\{(t_1,g_1),\ldots,(t_m,g_m)\} | (g_1,\ldots,g_m) \in G^{|D|}\}}} \prod_{(t_k,g_k) \in \omega} \gamma_{g_k|t_k} \theta_{i_k,j_k|g_k} \\ &= \sum_{\omega \in \Omega} \underbrace{\prod_{(t_k,g_k) \in \omega} \gamma_{g_k|t_k}}_{p_\omega} \prod_{(t_k,g_k) \in \omega} \theta_{i_k,j_k|g_k} \\ &= \sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{s_i,s_j \in S} \theta_{i,j|g}^{n_{i,j|g,\omega}} \end{aligned} \tag{7}$$

Here, each ω represents a single, fixed group assignment of the set of transitions in D where the set of all possible group assignments ω is defined as $\Omega = \{\{(t_1,g_1),\ldots,(t_m,g_m)\} | (g_1,\ldots,g_m) \in G^{|D|}\}$. 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 in [52].

We now substitute the likelihood $\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})$ in Equation (6) with this reformulated likelihood (Equation (7)) 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 $t_k \in D$:

$$\begin{aligned}\Pr(D|H) &= \int \underbrace{\sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{s_i,s_j \in S} \theta_{i,j|g}^{n_{i,j|g,\omega}}}_{\Pr(D|\boldsymbol{\theta},\boldsymbol{\gamma})} \underbrace{\prod_{g \in G} \prod_{s_i \in S} \frac{1}{B(\boldsymbol{\alpha}_{s_i|g})} \prod_{s_j \in S} \theta_{i,j|g}^{\alpha_{i,j|g}-1}}_{\Pr(\boldsymbol{\theta}_g|\boldsymbol{\alpha}_g)} \prod_{g \in G} d\boldsymbol{\theta}_g \\ &= \sum_{\omega \in \Omega} p_\omega \prod_{g \in G} \prod_{s_i \in S} \frac{1}{B(\boldsymbol{\alpha}_{s_i|g})} \int \prod_{s_j \in S} \theta_{i,j|g}^{n_{i,j|g,\omega}+\alpha_{i,j|g}-1} d\boldsymbol{\theta}_g \\ &= \sum_{\omega \in \Omega} \theta_\omega \underbrace{\prod_{g \in G} \prod_{s_i \in S} \frac{B(\boldsymbol{n}_{s_i|g,\omega} + \boldsymbol{\alpha}_{s_i|g})}{B(\boldsymbol{\alpha}_{s_i|g})}}_{\Pr(D_{g|\omega}|\boldsymbol{\alpha}_g)} \end{aligned}$$

This concludes the derivation of the marginal likelihood formula in Equation (5).

B Notation overview

The following table provides an overview of all important notations used throughout the article.

S	set of all states $S = \{s_1, \ldots, s_n\}$
D	set of observed transitions $D = \{t_1, ..., t_m\}$
G	set of all groups $G = \{g_1, \ldots, g_o\}$
src_k, dst_k	the source state src_k and the destination state dst_k of transtion t_k
i_k, j_k	the index of the source state i_k and the destination state j_k of transtion t_k
$\gamma_{g\|t}$	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 for 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 for group g, i.e., $\boldsymbol{\theta}_g = \{\boldsymbol{\theta}_{s_i\|g} \mid s_i \in S\}$
$\boldsymbol{\theta}$	transition probabilities for all groups $\boldsymbol{\theta} = \{\boldsymbol{\theta}_g \| g \in G\}$
ϕ	belief in transition probabilities (from a hypothesis)
$\phi_{i,j\|g}$	belief (from a hypothesis) in the probability of a transition from state s_i to state s_j for group g
$\alpha_{i,j\|g}$	Dirichlet parameter ($\in \mathbb{N}$) 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 $\Omega = \{\{(t_1, g_1), ..., (t_m, g_m)\} \| (g_1, ..., g_m) \in G^{\|D\|}\}$
ω	a fixed group assignment $\omega \in \Omega$ for each transition in 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 $\omega \in \Omega$
$\boldsymbol{n}_{g,\omega}$	the matrix $\boldsymbol{n}_{g,\omega} = (n_{i,j\|g,\omega})$ holds the number of transitions in dataset D between all states given group $g \in G$ and group assignment $\omega \in \Omega$