Mining Subgroups with Exceptional Transition Behavior

Florian Lemmerich
GESIS - Leibniz Institute for the Social Sciences & University of Koblenz-Landau
florian.lemmerich@gesis.org

Martin Becker
University of Würzburg
becker@informatik.uni-wuerzburg.de

Philipp Singer
GESIS - Leibniz Institute for the Social Sciences & University of Koblenz-Landau
philipp.singer@gesis.org

Denis Helic
Graz University of Technology
dhelic@tugraz.at

Andreas Hotho
University of Würzburg and L3S Hannover
hotho@informatik.uni-wuerzburg.de

Markus Strohmaier
GESIS - Leibniz Institute for the Social Sciences & University of Koblenz-Landau
markus.strohmaier@gesis.org

ABSTRACT

We present a new method for detecting interpretable subgroups with exceptional transition behavior in sequential data. Identifying such patterns has many potential applications, e.g., for studying human mobility or analyzing the behavior of internet users. To tackle this task, we employ exceptional model mining, which is a general approach for identifying interpretable data subsets that exhibit unusual interactions between a set of target attributes with respect to a certain model class. Although exceptional model mining provides a well-suited framework for our problem, previously investigated model classes cannot capture transition behavior. To that end, we introduce first-order Markov chains as a novel model class for exceptional model mining and present a new interestingness measure that quantifies the exceptionality of transition subgroups. The measure compares the distance between the Markov transition matrix of a subgroup and the respective matrix of the entire data with the distance of random dataset samples. In addition, our method can be adapted to find subgroups that match or contradict given transition hypotheses. We demonstrate that our method is consistently able to recover subgroups with exceptional transition models from synthetic data and illustrate its potential in two application examples. Our work is relevant for researchers and practitioners interested in detecting exceptional transition behavior in sequential data.

CCS Concepts: Information systems –> Data mining

Keywords: Subgroup Discovery; Exceptional Model Mining; Markov chains; Transitions

1. INTRODUCTION

Exceptional Model Mining [13, 29], a generalization of the classic subgroup discovery task [3, 25], is a framework that identifies patterns which contain unusual interactions between multiple target attributes. In order to obtain operationalizable insights, it emphasizes the detection of *easy-to-understand* subgroups, i.e., it aims to find exceptional subgroups with descriptions that are directly interpretable by domain experts. In general, exceptional model mining operates as follows: A target model of a given model class is computed once over the entire dataset, resulting in a set of model parameters. The same parameters are also calculated for each subgroup in a large (often implicitly specified) candidate set, using only the instances covered by the respective subgroup. A subgroup is considered as *exceptional* or *interesting* if its parameter values differ significantly from the ones of the overall dataset. While exceptional model mining has been implemented for a variety of model classes including classification [29], regression [12], Bayesian network [15] and rank correlation [11] models, it has not yet been applied with models for sequential data.

In this paper, we aim to apply exceptional model mining to discover interpretable subgroups with exceptional transition behavior. This enables a new analysis method for a variety of applications. As one example, assume a human mobility dataset featuring user transitions between locations. The overall transition model could for example show that people either move in their direct neighborhood or along main roads. Detecting subgroups with exceptional transition behavior goes beyond this simple analysis: It allows to automatically identify subgroups of people (such as “male tourists from France") or subsegments of time (such as “10 to 11 p.m.") that exhibit unusual movement characteristics, e.g., tourists moving between points-of-interest or people walking along well-lit streets at night. Other application examples could include subgroups of web-users with unusual navigation behavior or subgroups of companies with unusual development over time, cf. [24].

The main contribution of this paper is a new method that enables mining subgroups with exceptional transition behavior by introducing *first-order Markov chains as a novel model class for exceptional model mining*. Markov chains have been utilized for studying sequential data about, e.g., human navigation [35, 42] and mobility [20], meteorology [19] or economics [24, 51]. To apply exceptional model mining with this model, we derive an interestingness measure that quantifies the exceptionality of a subgroup's transition model. It measures how much the distance between the Markov transitions matrix of a subgroup and the respective matrix of the entire data deviates from the distance of random dataset samples. We also show how an adaptation of our approach allows to find subgroups specifically matching (or contradicting) given hypotheses about transition behavior (cf. [8, 41, 45]). This enables the use of exceptional model mining for a new type of studies, i.e., the detailed analysis of such hypotheses. We demonstrate the potential of the proposed approach with synthetic as well as real-world data.

KDD '16 August 13–17, 2016, San Francisco, CA, USA

DOI: aa.bbb/ccc_d

The remainder of this work is organized as following: We summarize our background in Section 2. Then, the main approach for mining subgroups with exceptional transition behavior is introduced in Section 3. Section 4 presents experiments and results. Finally, we discuss related work in Section 5, before we conclude in Section 6.

2. BACKGROUND

Our solution extends *Exceptional Model Mining* with first-order *Markov Chain Models*. In the following, we give a brief overview of both techniques.

2.1 Exceptional Model Mining

We formally define a dataset D as a set of data instances $i \in I$ described by a set of attributes A containing describing attributes $A_D \subset A$ and model attributes $A_M \subset A$. A *subgroup* consists of a *subgroup description* $p : D \rightarrow \{true, false\}$ that is given by a Boolean function, and a *subgroup cover* $c(p)$, i.e., the set of instances described by p, i.e., $c(p) = \{i \in I | p(i) = true\}$. In principle, our approach works with any *pattern description language* to describe subgroups. As a canonical choice, we focus in our experiments on conjunctions of selection conditions over individual describing attributes, i.e., attribute-value pairs in the case of a nominal attribute, or intervals in the case of numeric attributes. Hence, an example description of a subgroup could be: $Age < 18 \wedge Gender = Male$. Due to combinatorial explosion, a large number of subgroups can be formed even from comparatively few selection conditions.

From this large set of candidate subgroups, exceptional model mining identifies the ones that are unusual ("interesting") with respect to a target model class and the model attributes A_M. While for traditional subgroup discovery the target concept is given by a Boolean expression over a single attribute (e.g., *"class = good"*) and a subgroup is considered as interesting if the expressions holds more (or less) often than expected, exceptional model mining considers more complex target concepts. Given a specific *model class* (such as a correlation or a classification model), the *model parameters* for a subgroup can be computed depending on the instances of the respective subgroup. A subgroup is then considered as interesting if its model parameters deviate significantly from the parameters of the model that is derived from all dataset instances. For example, consider a study about the correlation (model class) between the two model attributes exam preparation time of students and the final score they achieve. A finding of exceptional model mining could be: *"While overall there is a positive correlation between the exam preparation time and the score ($\rho = 0.3$), the subgroup of males that are younger than 18 years show a negative correlation ($\rho = -0.1$)"*.

The goal of finding exceptional subgroups is accomplished by using a quality measure q that maps a subgroup to a real number (a score) based on the supposed interestingness of its model parameters and performing a search for the subgroups with the highest scores.

2.2 Markov Chain Models

In this paper, we introduce first-order Markov chains as a target concept for exceptional model mining. Markov Chains are stochastic systems modeling transitions between states $s_1, \ldots, s_m$. Each observed sequence of states corresponds to a sequence of assignments of random variables $X_1, \ldots, X_z$, $X_i \rightarrow \{s_1, \ldots, s_m\}$. The commonly employed *first-order* Markov chain model assumes that this process is memoryless, i.e., the probabilities of the next state at time $\tau+1$ only depend on the current state at time τ: $P(X_{\tau+1} = s_j | X_1 = s_{i_1}, \ldots, X_\tau = s_{i_\tau}) = P(X_{\tau+1} = s_j | X_\tau = s_{i_\tau})$, denoted in short as $P(s_j|s_i)$. First-order Markov chain modeling is an established and robust method that underlies many analyses and algorithms [22, 42], the most prominent example being Google's PageRank [35].

The parameters of a first-order Markov chain model can be specified by a stochastic transition matrix $T = (t_{ij})$ with matrix elements $t_{ij} = P(s_j|s_i)$ displaying the conditional probability for a transition from state i to state j; Thus, the sum of elements for each matrix row is 1. When working with datasets containing transitions, we can easily derive a stochastic transition matrix from a transition matrix containing counts for each transition by normalizing each row. Thus, we use the term transition matrix for both, stochastic matrices and count matrices, if specifics are clear from the context.

3. MAIN APPROACH

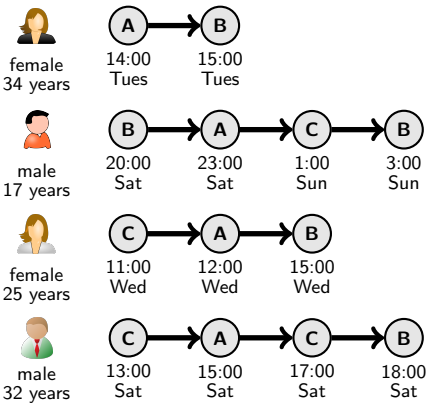
Given a set of state sequences and additional information on the sequences or parts of sequences, our main goal is to discover subgroups of transitions that induce exceptional transition models. We formalize this as an exceptional model mining task.

For that purpose, we first prepare the dataset D of transitions with model attributes A_M and describing attributes A_D (see Section 3.1). These allow to form a large set of candidate subgroup descriptions and filter the dataset accordingly. For each candidate subgroup g, we then determine the corresponding set of transitions and compute its transition matrix T_g. By comparing this matrix to a reference matrix T_D derived from the entire data, we can then calculate a score according to an interestingness measure q (see Section 3.2). In order to detect the subgroups with the highest scores, standard exceptional model mining search algorithms are utilized to explore the candidate space (see Section 3.3). The automatically found subgroups then should be assessed by human experts (see Section 3.4). In a variation of our approach, we do not use the transition matrix of the entire data T_D for comparison with the subgroup matrices T_g, but instead employ a matrix T_H that expresses a user-specified hypothesis. This allows for finding subgroups that specifically match or contradict this hypothesis (see Section 3.5).

3.1 Data Representation

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

Figure 1(a-b) illustrates such a preparation process for a simple example. It shows sequences of states (e.g., certain locations) that users have visited and some background knowledge, i.e., some user information and the time of each visit (Figure 1a). This information is integrated in a single data table (Figure 1b). It contains two columns for the transition model attributes A_M, i.e., for the source and the target state of each transition. Additional describing attributes A_D capture more information on these transitions. This includes information specific to a single transition such as the departure time at the source state but also information on the whole sequence that is projected to all of its transitions, e.g., user data or the sequence length. Example subgroup descriptions that can be expressed based on these attributes are *"all transitions by female users"*, *"all transitions on Saturdays"*, or combinations such as *"all transitions between 13:00h and 14:00h from users older than 30 years that visited at least three locations"*. As different types of information can be considered for the construction of the descriptive attributes, the approach is very flexible.



(a) Sequence data with background knowledge

A_M A_D

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

(b) Transition dataset

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

(c) Transition matrix T_D (entire dataset)

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

(d) Transition matrix $T_{Gender=f}$

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

(e) Transition matrix $T_{Weekday=Sat}$

Figure 1: ***Illustrative example.*** **Sequential data with background information (a) is initially transformed to a transition dataset with transition model attributes A_M and descriptive attributes A_D (b). To discover interesting subgroups, transition matrices for the dataset (c) and for the candidate subgroups, e.g., *Gender=f* (d) or *Weekday=Sat* (e), are computed and compared with each other.**

3.2 Interestingness Measurement

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

Distance measure and weighting. First, we compute the reference transition (count) matrix $T_D = (d_{ij})$ for the overall dataset D as exemplified in Figure 1c. To evaluate a subgroup g, all instances in the tabular dataset that match its subgroup description are identified and a transition matrix $T_g = (g_{ij})$ is determined accordingly (see, e.g., Figure 1d and Figure 1e). Then, a distance measure is employed to measure the difference of transition probabilities in these matrices. For both matrices T_D and T_g, each row i describes a conditional categorical probability distribution for the next state given that state s_i was observed directly before. In literature, several methods have been proposed to compare such distributions with each other. Here, we focus on the *total variation distance* δ_{tv}, also called *statistical distance* or (excluding the constant factor) *Manhattan distance*. For one row, this is computed as the sum of absolute differences between the normalized row entries, i.e., between transition probabilities:

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

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

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

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

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

Of course, there are also alternatives to the total variation distance measure that we can use, e.g., the *Kullback-Leibler divergence* $\delta_{kl}(g,D,i) = \sum_j g_{ij} \cdot \log \frac{g_{ij}}{d_{ij}}$ or the *Hellinger distance* $\delta_{hell}(g,D,i) = \frac{1}{\sqrt{2}} \sqrt{\sum_j (\sqrt{g_{ij}} - \sqrt{d_{ij}})^2}$. However, we focus in this paper on the weighted total variation as it naturally extends existing approaches for interestingness measures from classical pattern mining: it can be considered as an extension of the measure *multi-class weighted relative accuracy* for multi-class subgroup discovery [1]. Additionally, it can also be interpreted as a special case of *belief update* in a Bayesian approach as it has been proposed in [40] for traditional pattern mining. We provide a proof for this in the Appendix. Despite this focus, we also conducted a large set of experiments with all three distance measures in parallel with overall very similar results.

Comparison with random samples. The measure ω_{tv} describes a weighted distance between transition matrices. Yet, it is heavily influenced by the number of transitions covered by a subgroup. For example, small subgroups might be over-penalized by small weighting factors w_i, while very large subgroups can be expected to reflect the distribution of the overall dataset more precisely. Thus, using ω_{tv} directly as an interestingness measure does not consistently allow for identifying subgroups that actually influence transition behavior in presence of noise attributes, cf. Section 4.2.

To account for these effects, we propose a sampling-based normalization procedure. First, we compute the weighted distance $\omega_{tv}(g,D)$ of the subgroup g to the reference matrix as described before. Then, we draw a set of r random sample transition datasets $R = \{R_1, \ldots, R_r\}$, $R_i \subset D$ from the overall dataset D, each containing as many transitions as the evaluated subgroup g. Then, we compute the weighted distances $\omega_{tv}(R_i)$ for each of these samples, and build a *distribution of false discoveries* (cf. [14]) from the obtained scores. In particular, we compute the mean value $\mu(\omega_{tv}(R_1,D), \ldots, \omega_{tv}(R_r,D))$ and the sample standard deviation $\sigma(\omega_{tv}(R_1,D), \ldots, \omega_{tv}(R_r,D))$ for the distances of the random samples. A subgroup is considered as interesting if the distance of the

subgroup strongly deviates from the distances of the random samples. We quantify this by a (marginally adapted) *z-score*, which we will use as the interestingness measure q in our approach:

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

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

The rationale for using sampling *without replacement* is that the subgroup itself also cannot contain multiple instances of the same transition. As a consequence, even subgroups selected by random noise attributes would appear to be exceptional compared to samples with replacement in some settings. Sampling without replacement is also equivalent to randomizing (shuffling) the entries of the column for the target state as it has been suggested in pattern mining for the statistical validation of results patterns [14], see also [21].

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

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

Significance. To ensure that our findings are not only caused by random fluctuations in the data, the z-score q_{tv} that we compute as interestingness score can be used as a test statistic for a z-test on statistical significance. Yet, this test requires a normal distribution of the weighted distances $\omega_{tv}(R_i,D)$ obtained from the samples. Although in many practical situations the distribution of the sampled distances is *approximately* normally distributed, this does not necessarily hold in all cases. We thus propose a two-step approach to assess statistical significance of the results. First, we use a normality test such as the *Shapiro-Wilk-Test* [39] on the set of distance scores obtained for the sample set R. If the test does not reject the assumption of normality, a p-value can be directly computed from the z-score. If normality is rejected, a substantially larger set of random samples can be drawn to compute the *empirical p-value* of a specific subgroup [21], i.e., the fraction of samples that show a more extreme distance score than the subgroup. Although this is computationally too expensive to perform for every single candidate subgroup, it can be used for confirming significance for the most interesting subgroups in the result set.

For both methods one must consider the *multiple comparison problem* [23]: if many different subgroups are investigated (as it is usually done in data mining), then some candidates will pass standard significance tests with unadapted significance values by pure chance. Therefore an appropriate correction such as *Bonferroni correction* [16] or *layered critical values* [47] must be applied.

Estimate the effect of limited sample numbers. Determining the interestingness score $q_{tv}(g,D)$ requires to choose a number of random samples r. While fewer samples allow faster computation, results might get affected by random outliers in drawn samples. To estimate the potential error in the score computation caused by the limited number of samples, we employ a *bootstrapping approach* [17]: we perform additional sampling on the weighted distances of the original samples $S = \{\omega_{tv}(R_1,D),\ldots,\omega_{tv}(R_r,D)\}$. From this set, we repeatedly draw (e.g., 10,000 times) *"bootstrap replications"*, i.e., we draw r distance values by sampling *with* replacement from S and compute the subgroup score q_{tv} for each replication. The standard deviation of the replication scores provides an approximation of the standard error compared to an infinitely large number of samples, cf. [18]. In other words, we estimate how precise we compute the interestingness score q_{tv} with the chosen value of r compared to an infinite number of samples. If the calculated standard error is high compared to the subgroup score, re-computation with a higher number of samples is recommended.

3.3 Subgroup Search

To detect interesting subgroups, we enumerate all candidate subgroups in the search space in order to find the ones with the highest scores. For this task, a large variety of mining algorithms has been proposed in the pattern mining literature featuring exhaustive as well as heuristic search strategies, e.g., depth-first search [25], best-first search [46, 52], or beam-search [27, 44]. In this paper, we do not focus on efficient algorithms for exceptional model mining, but apply a depth-first mining algorithm as a standard solution.

Candidate evaluation in our approach is computationally slightly more expensive than for traditional subgroup discovery. That is, the runtime complexity for determining the score of a single subgroup in our implementation is $O(r \cdot (N + S^2))$ for a dataset with N transitions, S different states, and a user chosen parameter of r samples: selecting the set of instances from a subgroup as well as drawing a stratified sample requires $O(N)$ operations per subgroup or sample. The transition matrices for each of these transition sets can also be build in linear time. The weighted distance for each of the r samples and the subgroup can then be determined in $O(S^2)$ as a constant number of operations is required for each of the S^2 matrix cells.

A typical problem in pattern mining is redundancy, i.e., the result set often contains several similar subgroups. for example, if the subgroup *male* induces an exceptional transition model and thus achieves a high score, then also the subgroup *males older than 18* can be expected to feature a similarly unusual model and receive a high score—even if age does not influence transition behavior at all. A simple, but effective approach to reduce redundancy in the result set is to adapt a *minimum improvement constraint* [7] as a filter criterion. To that end, we remove a subgroup from the result set if the result contains also a generalization, i.e., a subgroup described by a subset of conditions, with a similar (e.g., less than 10% difference) or higher score.

3.4 Subgroup Assessment

Automatic discovery algorithms with the proposed interestingness measure can detect subgroups with exceptional transition models. Yet, to interpret the results, manual inspection and assessment of the top findings is crucial as they allow users to identify in what aspects

the found "interesting" subgroups differ from the overall data. For that purpose, a comparison between the subgroup transition matrix and the reference matrix is required. Yet, manual comparison can be difficult for large matrices (state spaces). Therefore, we recommend to assess subgroups with summarizing *key statistics*, such as the number of transitions in a subgroup, the weighted distance ω_{tv} between subgroup and reference transition matrices, its unweighted raw distance $\Delta_{tv} = \sum_i \delta_{tv}(sg, D, i)$ or the distribution of source and target states. Additionally, *exemplification*, e.g., by displaying representative sequences, and *visualizations* are helpful tools for subgroup inspection. In that regard, we propose a graph-based visualization to get a quick overview of the differences between subgroup and reference transition matrices, see Figure 3 for an example. Here, each state is represented as a node and directed edges represent the differences of transition probabilities between the states. The width of an edge represents the amount of change in the transition probability, the color indicates if it is a decrease or increase. Edges without significant differences can be removed from the graph to increase visibility. In addition to that, application-specific visualizations often allow a natural view on the data, see Figure 2 for an example featuring geo-spatial data.

3.5 Variation: User-Defined Hypotheses

In addition to comparing subgroups to the overall dataset, our approach can also detect subgroups that specifically contradict or match a user-defined hypothesis. Following the concepts of [41], we can express such a hypothesis as a *belief matrix* $T_H = (h_{ij})$, where higher values h_{ij} indicate a stronger belief in transitions from state s_i to state s_j. An example of a hypothesis considering the example dataset of Figure 1 could be stated as: $\left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{smallmatrix}\right)$. This hypothesis formalizes a belief that users from state A (first row) will always go to state B, and users from the states B and C will proceed to any of the three states with equal probability.

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

4. EXPERIMENTS

We demonstrate the potential of our approach with synthetic and empirical data. With synthetic data, we show that our approach is able to recover (combinations of) conditions that determine the transition probabilities in presence of noise attributes. With empirical data, we illustrate possible application scenarios and findings.

For all our experiments, we used all attribute-value pairs of nominal attributes and intervals of numeric attributes obtained by equal-frequency discretization in 5 groups to construct subgroup descriptions. For the interestingness measure, we used $r = 1,000$ random samples. If not stated otherwise, we focused on subgroups with simple descriptions, i.e., no combinations of selection conditions were considered. For the empirical studies, we confirmed our top results to be statistically significant on an $\alpha = 0.01$ level using the procedure presented in Section 3.2. Our implementation, an extension of the VIKAMINE data mining environment [4], and the synthetic datasets are publicly available[1].

[1] https://bitbucket.org/florian_lemmerich/subtrails

4.1 Synthetic I: Random transition matrices

We start with a synthetic dataset directly generated from two transition matrices of first-order Markov chain models.

Experimental setup. We created two 5×5 matrices of transition probabilities by inserting uniformly distributed random values in each cell and normalizing the matrices row-wise. Then, for each generated instance, one of the matrices was chosen based on two attributes, a ternary attribute A and a binary attribute B. If both attributes take their first values, i.e., $A = A1$ and $B = B1$, then transitions were generated from the first matrix, otherwise from the second matrix. For each combination of values, we generated $10,000$ transitions, resulting in 60,000 transitions overall. For each transition, we additionally generated random values for 20 binary noise attributes, each with an individual random probability for the value *true*. We employed our approach with a maximum search depth of 2 selectors to find subgroups with different transition models compared to the overall dataset. Our approach should then detect the subgroup $A = A1 \wedge B = B1$ as the most relevant one.

Results. The top-5 result subgroups are displayed in Table 1. It shows the number of covered transitions, the score of the interestingness measure q_{tv} including the standard error of its computation estimated by bootstrapping ($\pm$), the weighted total variation ω_{tv} between the subgroup and the reference transition matrix, and its unweighted counterpart Δ_{tv}. Result tables for the following experiments will be analogously structured.

We observe that our approach successfully recovered the subgroup with transitions that were generated from a different probability matrix, i.e., the subgroup $(A = A1 \wedge B = B1)$. This subgroup receives the best score q_{tv} by a wide margin. The subgroup with the next highest score $(A = A1)$ is a generalization of this subgroup. Since it contains transitions generated from both matrices in a different mixture, it also features indeed an unusual transition model compared to the entire dataset. In the same way, the next subgroups all feature the attributes A and B that actually influence the transition behavior, and none of the noise attributes. These top subgroups all pass a Bonferroni-adjusted statistical significance test as described in Section 3.2 with an empirical p-value of $p \ll 10^{-10}$, while all subgroups containing only noise attributes (not among the shown top subgroups) do not pass such a test with a critical value of $\alpha = 0.05$.

4.2 Synthetic II: Random walker

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

Experimental setup. First, we generated a scale-free network consisting of $1,000$ nodes (states) with a Barabasi-Albert model [6]. That is, starting with a small clique of nodes, new nodes with degree 10 were inserted to the graph iteratively using preferential attachment. Then, we assigned one of ten colors randomly to each node. On this network, we generated $200,000$ sequences of random walks with five transitions each, resulting in $1,000,000$ transitions overall. For each sequence, we randomly assigned a walker type. With

Table 1: Top subgroups for random transition matrix data. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv} and the unweighted total variation Δ_{tv}.

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

a probability of 0.8, the walk was purely *random*, i.e., given the current node of the walker, the next node is chosen with uniform probability among the neighbouring nodes. Otherwise, the walk was *homophile*, i.e., it chose nodes of the same color twice as likely. For each transition, the resulting dataset contains the source node, the target node, the type of the respective walker (random or homophile), and additionally the values for 20 binary noise attributes, which were assigned with an individual random probability each.

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

Results. Table 2 presents the results for the three experiments. As intended, exceptional model mining identified the subgroups that influence the transition behavior as the top subgroups for all three tasks. In the first experiment (see Table 2a) both subgroups described by the *type* attribute are top-ranked. For the second experiment (see Table 2b), the subgroup *type=random* receives the highest score. As any random subgroup contains transitions from homophile as well as non-homophile walkers, this subgroup is by construction the one that exposes the least homophile behavior. Its complement subgroup *type=homophile* does not contradict our hypothesis and thus does not appear in the top subgroups. By contrast and as expected, the subgroup *type=homophile* receives the highest score in the third experiment that searches for subgroups matching the homophile hypothesis, while *type=random* is not returned as a top result, cf. Table 2c. For all three experiments, the statistical significance of the top subgroups described by the *type* attribute was decisive ($p \ll 10^{-10}$), while the top findings for the noise attributes were not significant at the Bonferroni-adjusted $\alpha = 0.05$ level.

Table 2: Top subgroups for the random walker datasets. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv} and the unweighted total variation Δ_{tv}.

(a) Comparison to the overall dataset

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	37.63 ± 2.27	51,929	125.96
Type = Random	799,085	34.25 ± 2.21	51,929	31.73
Noise9 = True	318,165	2.47 ± 0.21	51,358	77.94
Noise9 = False	681,835	2.42 ± 0.19	51,358	36.27
Noise2 = False	18,875	1.85 ± 0.14	14,844	394.51

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

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Random	799,085	31.35 ± 2.70	1,554,130	981.38
Noise4 = True	519,130	2.58 ± 0.20	1,008,912	981.25
Noise2 = False	18,875	2.44 ± 0.18	37,057	987.49
Noise1 = True	469,290	2.40 ± 0.19	912,032	981.26
Noise0 = False	476,590	2.24 ± 0.17	926,241	981.26

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

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	12.00 ± 0.77	389,841	981.04
Noise4 = False	480,870	2.61 ± 0.22	934,190	981.20
Noise19 = False	657,235	2.33 ± 0.17	1,276,868	981.20
Noise1 = False	530,710	1.95 ± 0.16	1,031,101	981.20
Noise0 = True	523,410	1.77 ± 0.16	1,016,899	981.21

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

We also experimented extensively with different parametrizations (e.g., different walker type probabilities or different numbers of colors in the network). Consistently, we were able to identify the two subgroups *type=random* and *type=homophile* as the top subgroups.

4.3 Flickr

In addition to the synthetic datasets, we present illustrative examples with empirical data; we start with data from Flickr.

Experimental setup. For this dataset, we crawled all photos with street-level accurate geo-spatial information (i.e., latitude and longitude) in Manhattan from the years 2010 to 2014 on Flickr. Each photo was mapped according to its geo-location to one of the 288 census tracts (administrative units) that we use as states in our model, cf. also [20]. Based on this information, we built sequences of different tracts (i.e., no self-transitions) that users have taken photos at. Additionally, we elicited a wide range of describing attributes for each transition, i.e., the number of photos the respective user has uploaded from Manhattan, the number of views the source photo of the transition received on Flickr, as well as the month, the weekday and the hour this photo was taken. We added two more features based on the user origin, that is, the tourist status and the country of origin. We considered a user to be a tourist if the time from her first to her last photo does not exceed 21 days, cf. [10]. Country information of a user was derived from the location field in her user profile by extracting the country using a combination of querying

Table 3: Top subgroups for the Flickr dataset. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv} and the unweighted total variation Δ_{tv}.

(a) Comparison to the overall dataset

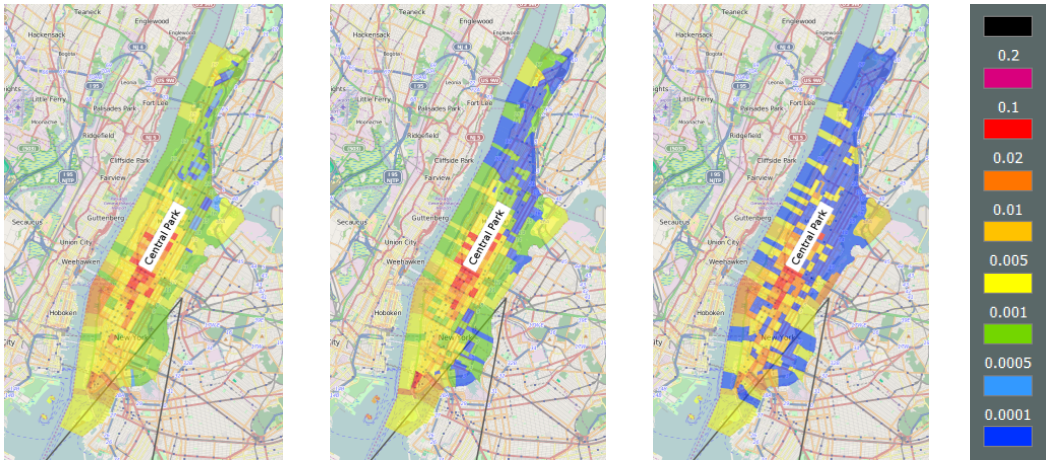
Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Photos > 714	76,859	103.83 ± 2.41	42,277	106.68
# Photos $\leq$ 25	78,254	88.83 ± 2.07	37,555	141.78
Tourist = True	76,667	75.42 ± 1.79	33,418	148.64
Tourist = False	310,314	75.00 ± 1.60	33,418	16.92
Country = US	163,406	64.47 ± 1.39	44,822	70.97
# Photos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

(b) Comparison to the *Proximate-PoI* hypothesis, contradicting

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

(c) Comparison to the *Proximate-PoI* hypothesis, matching

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



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

Figure 2: ***Exceptional transition behavior of Flickr users.*** **The figure shows transition probabilities from Central Park to other tracts in Manhattan for (a) the entire dataset, (b) the subgroup *tourists*, and (c) the subgroup *transitions during night time (22–23 h)*. Cool colors (blue, green) represent small, warm colors (orange, red) high transition probabilities, see the legend on the right hand side.**

GeoNames[2] and specialized regular expressions. The country information was only available for about half of the users. Overall, our dataset contained 386,981 transitions and allowed to construct 163 selection conditions.

In a first experiment, we aimed at discovering subgroups with different transition models compared to the entire data. Additionally, we investigated an existing hypothesis about the trails derived from Flickr photos, that is, the *Proximate-PoI* hypothesis. This hypothesis has been shown to be one of the best hypotheses for explaining movements in Flickr photo data [8]. It expresses that users at a certain location are most likely to go to another location that is (a) nearby and (b) contains important points of interest (PoI), such as tourist attractions or transportation hubs. To construct this hypothesis, the locations of points of interest have been extracted from DBPedia [28], see [8] for further construction details.

Results. Table 3 reports our results: the most exceptional subgroups in comparison with the overall data (see Table 3a) describe transitions by users that take either very many (more than 714) or very few (less than 25) photos. We explain this by the fact that users with overall fewer photos are more likely to travel a longer distance before taking another picture, resulting in more long distance transitions. The next two subgroups *Tourist=True* and *Tourist=False* suggest that tourists continue their trip to different locations than locals, e.g., as they are more interested in touristic attractions. Further top subgroups with deviating transition models involve the number of views pictures receive on Flickr and the country of origin.

Table 3b and Table 3c display the top subgroups that contradict the Proximate-PoI hypothesis, respectively match it. We observe that users with small amounts of pictures and non-tourists do not move as the investigated hypothesis suggests. Also, night time mobility (roughly 21h – 1h, see the result table for exact subgroup ordering) does not match this hypothesis, e.g., due to the closing of touristic attractions at night times. By contrast, tourists and users with many pictures as well as transitions at midday are especially consistent with the Proximate-PoI hypothesis.

Although we discover these exceptional subgroups from the large set of candidates automatically, it has to be investigated post-hoc *how* the transition models deviate. In that direction, we studied the subgroup *Tourist=True* in detail. For that purpose, we first computed the source state with the most unusual distribution of target states, i.e., the row that contributes the highest value to the weighted total variation q_{tv}. For the tourist subgroup, this state (tract) corresponds to the central park. We then visualized the transition probabilities for this single state for the entire dataset and the subgroup with the *VizTrail* visualization system [9], see Figure 2. It can be observed that tourists are less likely to move to the northern parts of Manhattan, but are more likely to take their next picture in the city center or at the islands south of Manhattan. For a second investigated subgroup, i.e., the subgroup of transitions between 22h and 23h, this effect is even more pronounced as almost no transitions from the central park to the northern or north-eastern tracts can be observed. Note, that this visualization only covers the transition probabilities of a single state, not the overall transition matrix used for detecting interesting subgroups.

4.4 LastFM

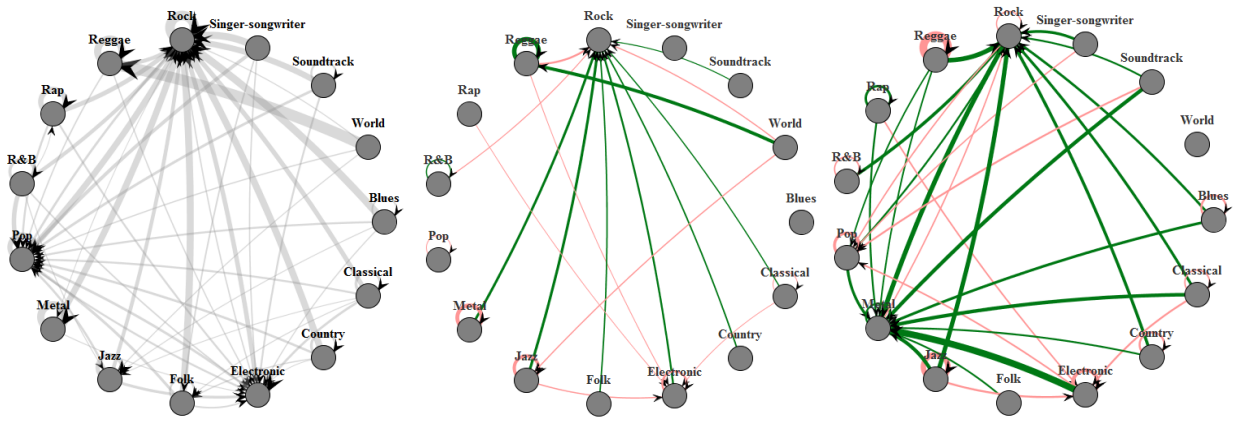
Additionally, we analyzed data from the LastFM music service.

Experimental setup. We used the *1K listening data*[3] containing the full listening history of 1,000 LastFM users featuring more than 19,000,000 tracks (songs) by more than 170,000 artists. With this data, we studied sequences of music genres (such as *rock*, *pop*, *rap*, *classical*, etc.) of songs that users listened to, focusing on a list of 16 main genres. Since genre information is difficult to obtain on a track-level, we labeled each track with the best fitting genre for the respective artist as obtained by the EchoNest API[4]. In doing so, we could determine genres for more than 95% of the tracks. We then constructed genre transitions for each user based on the sequence of tracks she had listened to. We filtered subsequent tracks of the same artist to remove cases where the user listened to all songs of a single album. Additionally, we removed all transitions with unknown source or target state (genre). Thus, we obtained a dataset of 9,020,396 transitions between tracks. Background knowledge includes user information about age, gender, origin and the year of signup to LastFM, and the point in time the source song of the transition was played, i.e., the hour of the day, the weekday, the

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

[3]http://ocelma.net/MusicRecommendationDataset/index.html

[4]http://developer.echonest.com/



(a) All transitions (b) Users from the United States (c) Users from Finland

Figure 3: *Exceptional transition models of LastFM users.* The figures show transitions between music genres: stronger arrows represent higher transition probabilities. (a) shows all transitions in the data, (b) and (c) illustrate the differences of transition models in two exceptional subgroups. Green arrows indicate that transitions are more probable in the subgroup than in the overall dataset, red arrows the contrary. E.g., it can be observed in (b) that users from the US are more likely to listen to *Reggae* after *World* music; (c) shows that Finnish users have higher transition probabilities to *Metal*. Insignificant differences are removed for visibility.

month and the year. This allowed to generate 86 selection conditions. On this data, we applied our approach twice to find subgroups with exceptional transition behavior: once for subgroups described by a single selection condition only, and once including combinations of two selection conditions (search with depth 2).

Results. Results for single selection conditions are displayed in Table 4a. We can see that the country of users is an important factor: the majority of top subgroups is described by this attribute. Specifically, users from the United States, from Finland, and from Argentina exhibit transitions between music genres that are unusual compared to the entire data. By contrast, date and time influence the transitions between genres only little and do not describe any of the top subgroups. For subgroups described by combinations of conditions, see Table 4b, we can see that users with a high number of tracks show unusual transition behavior, especially if they signed up to the system early, or if they are in a certain age group.

Figure 3 visualizes differences in transition behavior in comparison to the overall dataset for two top-subgroups. Here, each node represents a state (genre). The first graph gives an impression on the transition probabilities in the entire dataset. Stronger arrows represent higher probabilities. We omit probabilities below 0.1. The next two graphs show *deviations* from these probabilities in the subgroups *country=US* and *country=Finland*. Green arrows indicate transitions that are more likely in the subgroup than in the overall data, red arrows imply less likely transitions. Stronger arrows represent higher deviations; small deviations (< 0.05) are omitted.

We observe that users from the US and Finland deviate from the overall behavior in characteristic ways. For example, users from the US tend to skip to *Rock* more often, while the same is true for users from Finland with regard to *Metal*. Also, we observe interesting dynamics between genres that go beyond the different target state distributions: for example, users from the US are more likely to listen to *Reggae* after a song from the *World* genre, while the preference for *Rock* decreases in this case. We can also see, that although *Rock* is overall more popular in the US, it follows a track of *Reggae* less likely than in the entire data.

Table 4: Top subgroups for the LastFM dataset. For each subgroup, we show the number of instances covered by this subgroup, the interestingness score q_{tv}, the weighted total variation ω_{tv} and the unweighted total variation Δ_{tv}.

(a) Single selection conditions only

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Country = US	2,576,652	427.23 ± 9.38	326,435	1.44
Country = Finland	384,214	408.37 ± 8.89	132,378	3.05
Country = Argentina	174,140	360.22 ± 8.62	84,285	4.54
# Tracks > 79277	1,803,363	355.27 ± 8.39	249,634	1.61
Country = Poland	378,003	346.09 ± 7.37	122,155	3.35

(b) Including combinations of selection conditions

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
# Tracks >=79277 $\wedge$ signup=2005	617,245	458.22 ± 5.70	186,048	3.38
# Tracks >=79277 $\wedge$ age = [23–24]	155,998	431.82 ± 5.50	89,769	4.67
# Tracks >=79277 $\wedge$ signup=2006	658,135	428.08 ± 5.74	182,690	3.38
Country = US	2,576,652	427.23 ± 9.38	326,435	1.44
Country = Finland	384,214	408.37 ± 8.89	132,378	3.05

5. RELATION TO STATE-OF-THE-ART

Mining patterns in sequential data has a long history in data mining. However, large parts of research have been dedicated to the tasks of finding frequent subsequences efficiently, see for example [2, 34, 50]. Other popular settings are sequence classification [31, 49] and sequence labeling [26]. Unlike this work, these methods do not aim to detect subgroups with unusual transition behavior.

Our solution is based on exceptional model mining, a generalization of subgroup discovery [3, 25]. This classical data mining task aims at finding descriptions of data subsets that show an unusual statistical distribution of a target concept. Traditional subgroup discovery focuses on a single attribute as a target concept. Exceptional model mining [13, 29] was proposed as a generalized framework that facilitates more complex target concepts over multiple target attributes. For exceptional model mining, different model classes, e.g., classification [29] and regression models [12], Bayesian networks [15] as well as advanced mining algorithms [30, 43, 44] have been proposed. No models featuring sequential data have been explored for exceptional model mining so far.

We have presented an approach to detect subgroups with exceptional transition models, i.e., subgroups that show unusual distributions of the target states in first order Markov chain models.

The results from our approach may correlate with subgroups that could also be obtained by multi-class subgroup discovery [1] that investigates the distribution of target states. However, such a static analysis aims to achieve a different goal than our analysis of behavior dynamics and will not capture all subgroups with exceptional transition models. For example, in the random walker synthetic dataset (see Section 4.2) the distribution of target states is approximately uniform for all subgroups by construction, also for the ones that influence the transition behavior. As a consequence and in contrast to our method, a static analysis could not recover the exceptional subgroups. Furthermore, the task of finding subgroups that match or contradict a hypothesis of dynamic state transitions (e.g., as demonstrated in the Flickr example, see Section 4.3) cannot be formulated as a more traditional subgroup discovery task.

Our interestingness measure is inspired by previous methods. The weighted distance measure can be considered as an adaptation of the multi-class weighted relative accuracy [1] or as a special case of the Bayesian belief update [40]. The randomization/sampling processes to capture significant differences of subgroups also builds upon previous approaches. In that direction, Gionis et al. [21] utilize swap randomization to construct alternative datasets in order to ensure the statistical significance of data mining results. For subgroup discovery, analyzing a distribution of false discoveries obtained by randomization has been proposed to assess subgroups and interestingness measures [14]. We have extended these methods to exceptional model mining with complex targets and have used it directly in the interestingness measure for the subgroup search.

For modeling sequential processes, Markov chains have been used in a wide variety of applications ranging from user navigation [36, 42] to economical settings and meteorological data [19]. The *mixed markov model* extension [37] of classical Markov chains features separate transition matrices for "segments" of users, but these segments are not interpretable, i.e., have no explicit descriptions. The work maybe closest to ours is [38], where the authors detect outliers of user sessions with respect to their probability in a Markov-chain model; outliers are then manually categorized into several interpretable groups. By contrast, our solution allows to identify descriptions of groups that show unusual transition behavior automatically from large sets of candidate subgroups.

Recently, also the comparison of hypotheses about Markov chain models has been popularized [8, 41, 45]. The approach proposed in this paper enables extensions to this line of research with more fine-grained analyses: we cannot only compare hypotheses against each other, but also identify (sets of) conditions under which a given hypothesis is matched or contradicted.

6. CONCLUSION

In this paper, we have introduced first-order Markov chains as a novel model class for exceptional model mining in sequence data with background knowledge. This enables a novel kind of analysis: it allows to detect subgroups that exhibit exceptional transition behavior, i.e., induce different transition models compared to the entire dataset. In addition, we have presented a variation of the standard task that compares subgroups against user-defined hypotheses, enabling a detailed analysis of given hypotheses about transition behavior. We have illustrated the potential of our approach by applying it to both synthetic and empirical data. For synthetic data, the proposed method successfully recovered exceptional transitions from artificial noise attributes.

In the future, we aim to improve and extend our approach in several directions. First, the proposed interestingness measure is currently based on individual transitions. As a consequence, few very long sequences (e.g., of very active users) can strongly influence the results. To avoid dominance of such sequences, a weighting of the transition instances according to the overall activity could be applied in future extensions, cf. [5]. In addition, we intend to investigate ways of speeding-up the mining process, e.g., by optimistic estimate pruning [48] or by using advanced data structures [30], and more sophisticated options to reduce redundancy, cf. [32, 33]. Finally, the generalization of the proposed model class to Markov chains of higher order or — even further — the substitution with more advanced sequential models could add additional expressiveness.

Acknowledgements. This work was partially funded by the German Science Fund project "PoSTs II" and the Austrian Science Fund project "Navigability of Decentralized Information Networks".

APPENDIX

In Bayesian statistics, one's current beliefs H are expressed by Bayesian probabilities over parameters θ. Given new information I, the *prior* belief $P(\theta|H)$ is updated to a *posterior* belief $P(\theta|H,I)$. Here, we show that the total variation measure $\omega_{tv} = \sum_i \left(\sum_j g_{ij} \cdot \sum_j \left| \frac{g_{i,j}}{\sum_j g_{i,j}} - \frac{d_{i,j}}{\sum_j d_{i,j}} \right| \right)$ (see Section 3.2) is order-equivalent to the *amount of Bayesian belief update* in the setting of Markov chain models if the reference matrix T_D is used as a very strong prior. That means that both measures ultimately imply the same ranking of subgroups. The belief update implied by a subgroup g is defined by the difference between the prior distribution and the posterior distribution after observing the instances covered by g. The amount of belief update was proposed in [40] as an interestingness measure for pattern mining in traditional settings.

As [41] suggests, we can elicit the matrix of a Dirichlet prior $T_A = (a_{ij})$ through a reference matrix of (pseudo-)observations $T_D = (d_{ij})$ using the formula $a_{ij} = (k \cdot d_{ij}) + 1$. Here, k specifies the strength of the belief expressed by the prior. It is updated to a posterior according to observed transitions in a subgroup g given in a transition matrix $T_g = (g_{ij})$. In this context, according to [42], the expected probabilities $E[p_{ij}](X)$ for a state transition from state s_i to state s_j in the prior are $\frac{a_{ij}}{\sum_j a_{ij}}$ and the expected probabilities in the posterior are $c_i \cdot \frac{g_{ij}}{\sum_j g_{ij}} + (1 - c_i) \cdot \frac{a_{ij}}{\sum_j a_{ij}}$, with $c_i = \frac{\sum_j g_{ij}}{\sum_j (g_{ij} + a_{ij})}$. To determine the overall belief update BU for all state transitions, we compute the absolute difference between the posterior and the prior for each cell and aggregate over all cells in the matrix:

$$
\begin{aligned}
BU(H,D) &= \sum_i \sum_j \left| \left(c_i \frac{g_{ij}}{\sum_j g_{ij}} + (1 - c_i) \frac{a_{ij}}{\sum_j a_{ij}} \right) - \frac{a_{ij}}{\sum_j a_{ij}} \right| \\
&= \sum_i \sum_j \left| c_i \cdot \left(\frac{g_{ij}}{\sum_j g_{ij}} - \frac{a_{ij}}{\sum_j a_{ij}} \right) \right| \\
&= \sum_i c_i \cdot \sum_j \left| \left(\frac{g_{ij}}{\sum_j g_{ij}} - \frac{a_{ij}}{\sum_j a_{ij}} \right) \right| \\
&= \sum_i \frac{1}{\sum_j (g_{ij} + a_{ij})} \sum_j g_{ij} \sum_j \left(\left| \frac{g_{ij}}{\sum_j g_{ij}} - \frac{a_{ij}}{\sum_j a_{ij}} \right| \right)
\end{aligned}
$$

Now, assume that we have a very strong belief in the prior, i.e., $k \to \infty$ and thus $a_{ij} \gg g_{ij}$. Then, the right hand sum converges to the total variation δ_{tv} between the observed transition matrix T_g and the reference matrix T_D. The factor $\sum_j g_{ij}$ corresponds to the weights w_i. The additional factor $\frac{1}{\sum_j (a_{ij,} + g_{ij})}$ is approximately constant across all subgroups if $a_{ij} \gg g_{ij}$ since a_{ij} is independent from the evaluated subgroup. Overall, the weighted total variation ω_{tv} describes the amount of belief update a subgroup induces to a prior that reflects a very strong belief in the transition probabilities given by the reference matrix T_D.

References

[1] T. Abudawood and P. Flach. Evaluation measures for multi-class subgroup discovery. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50. 2009.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In *International Conference on Data Engineering*, 1995.

[3] M. Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

[4] M. Atzmueller and F. Lemmerich. VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2012.

[5] M. Atzmueller and F. Lemmerich. Exploratory pattern mining on social media using geo-references and social tagging information. *International Journal of Web Science*, 2(1–2):80–112, 2013.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[7] R. J. Bayardo Jr. Efficiently mining long patterns from databases. In *ACM Sigmod Record*, volume 27, pages 85–93, 1998.

[8] M. Becker, P. Singer, F. Lemmerich, A. Hotho, D. Helic, and M. Strohmaier. Photowalking the city: Comparing hypotheses about urban photo trails on Flickr. In *Social Informatics*, 2015.

[9] M. Becker, P. Singer, F. Lemmerich, A. Hotho, D. Helic, and M. Strohmaier. Viztrails: An information visualization tool for exploring geographic movement trajectories. In *Conference on Hypertext and Social Media*, 2015.

[10] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Conference on Hypertext and Hypermedia*, 2010.

[11] L. Downar and W. Duivesteijn. Exceptionally monotone models: the rank correlation model class for exceptional model mining. In *International Conference on Data Mining*, pages 111–120, 2015.

[12] W. Duivesteijn, A. Feelders, and A. Knobbe. Different slopes for different folks: mining for exceptional regression models with cook's distance. In *International Conference on Knowledge Discovery and Data Mining*, 2012.

[13] W. Duivesteijn, A. Feelders, and A. Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, pages 1–52, 2015.

[14] W. Duivesteijn and A. Knobbe. Exploiting false discoveries–statistical validation of patterns and quality measures in subgroup discovery. In *International Conference on Data Mining*, 2011.

[15] W. Duivesteijn, A. Knobbe, A. Feelders, and M. Van Leeuwen. Subgroup discovery meets bayesian networks–an exceptional model mining approach. In *International Conference on Data Mining*, 2010.

[16] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[17] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, pages 1–26, 1979.

[18] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[19] K. R. Gabriel and J. Neumann. A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375).

[20] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Show me how you move and i will tell you who you are. In *International Workshop on Security and Privacy in GIS and LBS*, pages 34–41, 2010.

[21] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *Transactions on Knowledge Discovery from Data (TKDD)*, 1(3):14, 2007.

[22] S. Gomez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. Discrete-time markov chain approach to contact-based disease spreading in complex networks. *Europhysics Letters*, 89(3):38009, 2010.

[23] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[24] G. G. Judge and E. R. Swanson. Markov chains: basic concepts and suggested uses in agricultural economics. *Australian Journal of Agricultural Economics*, 6(2):49–61, 1962.

[25] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. 1996.

[26] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[27] N. Lavrač, B. Kavšek, P. A. Flach, and L. Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5:153–188, 2004.

[28] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2014.

[29] D. Leman, A. Feelders, and A. J. Knobbe. Exceptional model mining. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2008.

[30] F. Lemmerich, M. Becker, and M. Atzmueller. Generic pattern trees for exhaustive exceptional model mining. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012.

[31] N. Lesh, M. J. Zaki, and M. Ogihara. Mining features for sequence classification. In *International Conference on Knowledge Discovery and Data Mining*, 1999.

[32] J. Li, J. Liu, H. Toivonen, K. Satou, Y. Sun, and B. Sun. Discovering statistically non-redundant subgroups. *Knowledge-Based Systems*, 67:315–327, 2014.

[33] R. Li, R. Perneczky, A. Drzezga, and S. Kramer. Efficient redundancy reduced subgroup discovery via quadratic programming. *Journal of Intelligent Information Systems*, 44(2):271–288, 2015.

[34] C. H. Mooney and J. F. Roddick. Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys*, 45(2):19, 2013.

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[36] P. L. Pirolli and J. E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2(1-2):29–45, 1999.

[37] C. S. Poulsen. Mixed markov and latent markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, 7(1):5–19, 1990.

[38] N. Sadagopan and J. Li. Characterizing typical and atypical user sessions in clickstreams. In *International WWW Conference*, 2008.

[39] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[40] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *Transactions on Knowledge and Data Engineering*, 8:970–974, 1996.

[41] P. Singer, D. Helic, A. Hotho, and M. Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *International Conference on World Wide Web*, 2015.

[42] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PLoS ONE*, 9(7):e102070, 2014.

[43] M. van Leeuwen. Maximal exceptions with minimal descriptions. *Data Mining and Knowledge Discovery*, 21(2):259–276, 2010.

[44] M. van Leeuwen and A. Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2):208–242, 2012.

[45] S. Walk, P. Singer, L. E. Noboa, T. Tudorache, M. A. Musen, and M. Strohmaier. Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In *14th International Semantic Web Conference*, pages 551–568.

[46] G. I. Webb. Opus: An efficient admissible algorithm for unordered search. *Journal of AI Research*, 3(1):431–465, 1995.

[47] G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2–3):307–323, 2008.

[48] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, 1997.

[49] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *ACM SIGKDD Expl. Newsletter*, 12(1):40–48, 2010.

[50] M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1-2):31–60, 2001.

[51] A. Zimmermann, T. Heckelei, et al. Differences of farm structural change across european regions. *Discussion Paper-Food and Resource Economics, Institute for Food and Resource Economics, University of Bonn*, (2012: 4), 2012.

[52] A. Zimmermann and L. D. Raedt. Cluster-grouping: From subgroup discovery to clustering. *Machine Learning*, 77(1):125–159, 2009.