Detecting Community Patterns Capturing Exceptional Link Trails

Martin Atzmueller
Research Center for Information System Design
University of Kassel, Germany
atzmueller@cs.uni-kassel.de

***Abstract*—We present a new method for detecting descriptive community patterns capturing exceptional (sequential) link trails. For that, we provide a novel problem formalization: We model sequential data as first-order Markov chain models, mapped to an attributed weighted network represented as a graph. Then, we detect subgraphs (communities) using exceptional model mining techniques: We target subsets of sequential transitions between nodes that are *exceptional* in that sense that they either conform strongly to a specific reference or show significant deviations, estimated by a quality measure. In particular, such a community is described by a community pattern composed of descriptive features (of the attributed graph) covering the respective community. We present a comprehensive modeling approach and discuss results of a case study analyzing data from two real-world social networks.**

1. Introduction

Sequential social data can be generated in various forms, e. g., by establishing connections within a social network, by observing alarm sequences in industrial plants, or by visiting specific locations in a location-based social network. Sequences specific for a certain actor can then be represented as sequential link trails, i. e., as transitions between states, denoted by nodes of a network (e. g., alarms, contacts, locations), that the respective actor is interacting with. The analysis of such (sequential) link trails has broad applicability, including the exploration of web navigation patterns [1], patterns in industrial networks, as well as mobility patterns [2], or the detection, analysis and explanation of anomalies.

Problem. We formalize the novel problem of detecting descriptive community patterns [3] in the context of *exceptional (sequential) link trails*: We aim to detect subgraphs of an attributed weighted graph, i. e., communities of sequential transitions (between nodes of a network) that are exceptional in that sense that they either conform to a specific behavior model or show significant deviations. Such a community is described by a specific community pattern which is composed of descriptive features that are common to all members of the respective community, i. e., covering all involved sequential transitions. We could detect, for example, that participants of a distributed event being interested in *classical and latin music* show a significantly deviating behavior than all participants.

Objectives. We tackle the problem of detecting descriptive community patterns capturing subsets of sequential (link) transitions that show an exceptional behavior compared to some reference behavior (model). We present a novel approach using first-order Markov chain models [1], [2] combined with exceptional model mining techniques [3], [4], [5] for that task. Further, we discuss estimation methods for ranking exceptional patterns, exemplified by two proposed quality measures, for a general solution to this problem.

Approach & Methods. Based on description-oriented community detection techniques, cf., [3], we investigate subsets of sequential transitions, i. e., sequences of states, captured by sequential trails in order to detect exceptional community patterns. We present a method based on the DASHTrails approach [2] for distribution-adapted modeling and comparison of hypotheses with sequential trail data, and propose suitable quality measure for estimating the exceptionality. Modeled as first order Markov chains, those patterns can then be identified that e. g., show the largest evidence concerning the observed data, i. e., the reference model given by all transitions. The approach proposed in this paper identifies community patterns capturing a set of sequential transitions that are exceptional compared to that reference model.

Contributions. Our contribution is summarized as follows:

1) We provide a novel problem formalization and present a framework for detecting descriptive community patterns capturing exceptional sequential trails compared to a reference model, as estimated by a quality function.
2) Based on first order Markov chain models and exceptional model mining techniques, we propose a flexible modeling approach, and show how to embed the recent DASHTrails [2] approach in our context of detecting exceptional community patterns. Furthermore, we present suitable quality measures for estimating their quality in order to generate a ranking of the patterns.
3) We demonstrate the applicability of our proposed framework and the presented measures using a case study on two real-world social network datasets.

Structure. Section 2 discusses related work. After that, Section 3 outlines the proposed approach. Next, Section 4 presents results of a case study utilizing two real-world social network datasets. Finally, Section 5 concludes with a discussion and interesting directions for future work.

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00 ©2016 IEEE

2. Related Work

In this section, we summarize related work on community detection, exceptional model mining and sequential analysis, and put our proposed approach into context.

2.1. Community Detection

Communities and cohesive subgroups have been extensively studied in social sciences, e. g., using social network analysis methods [6]. Fortunato [7] presents a thorough survey on the state of the art community detection algorithms in graphs, focussing on detecting *disjoint* communities, e. g., [8], [9]. In contrast to those, our proposed approach detects overlapping community patterns, such that a node can be included in different community assignments. In general, overlapping communities allow an extended modeling of actor–actor relations in social networks: Nodes of a corresponding graph can then participate in multiple communities, e. g., [10], [11]. A general overview on algorithms for overlapping community detection is provided by Xie et al. [12] as comprehensive survey. In contrast to the algorithms and approaches discussed above, the proposed approach utilizes further *descriptive information* of *attributed graphs*, e. g., [13].

Attributed (or labeled) graphs as richer graph representations enable approaches that specifically exploit the descriptive information of the labels assigned to nodes and/or edges of the graph. Overall, there are several methods that consider community detection and description, i. e., that focus on generating *explicit descriptions connected with the graph structure*. Most methods aim at detecting dense structures based on quasi-cliques that somehow correlate with respective descriptive patterns, e. g., [14], [15], [16]. In [3], we focus on *description-oriented community detection* and present the COMODO algorithm using subgroup discovery techniques [4], [17]. For providing both structurally valid and interpretable communities we utilize the graph structure as well as additional descriptive features of the graph's nodes. Using additional descriptive features of the nodes contained in the network, we approach the task of identifying communities as sets of nodes together with a *description*, i. e., a logical formula on the values of the nodes' descriptive features. Such a *community pattern* then provides an intuitive description of the community, e. g., by an easily interpretable conjunction of attribute-value pairs. COMODO is able to identify communities according to standard community quality measures, while providing characteristic descriptions at the same time.

Here, in contrast, we do not only focus on the graph structure, but compare a weighted subgraph (modeling a subset of sequential transitions) to a reference model in order to identify conforming or deviating community patterns covering the respective subgraph. In this paper, we provide a general framework for obtaining the k-best community patterns capturing exceptional sequential link trails. To the best of the authors' knowledge, no community detection approach tackling this problem has been proposed so far.

2.2. Exceptional Model Mining

The detection and analysis of irregular or exceptional patterns, e. g.,, anomalies, in network-structured data is a novel research area, e. g., for identifying new and/or emerging behavior, or for identifying detrimental or malicious activities. The former can be used for deriving new information and knowledge from the data, for identifying events in time or space, or for identifying interesting, important or exceptional groups [18], [19].

In the context of descriptive pattern mining, the concept of *exceptional model mining* has recently been introduced [5]. It can be considered as a variant of subgroup discovery for detecting interesting subgroups, e. g., [4], [17], [20], [21], enabling more complex target properties. Essentially, exceptional model mining tries to identify interesting patterns with respect to a local model derived from *a set* of *target* attributes, e. g., a correlation or a linear regression model. The interestingness can be flexibly defined, e. g., by a significant deviation from a model that is derived from the total population. Possible applications include the identification of characteristic patterns [22], network analysis of node information [23], [24], [25], or descriptive community mining [3], [26], [27]. Here, we adapt exceptional model mining techniques to our sequential trail problem setting.

2.3. Sequential Analysis

A general view on modeling and mining of ubiquitous and social multi-relational data is given in [28] focusing on social interaction networks captured during certain events, e. g., during conferences. Here, contacts patterns, for example, and their underlying mechanisms, e. g., [29] are analyzed. Furthermore, [25] describe the dynamics of community structures and roles at conferences, while [30] focuses on their evolution. However, the analysis in these contexts targets aggregated sequential data. Navigational patterns, as sequential (link) patterns in online systems, have been analyzed and modeled, e. g., in [1], [31]. [32] defines a sequence-based representation of networks, where sequential patterns are used to characterize communities.

In contrast to that, our approach focuses on the detection of community patterns capturing sequential transitions. Similar to evidence networks in the context of social networks, e. g., [33], [34], we model transitions assuming a certain interpretation of the data towards a sequential representation. Then, we can identify exceptional patterns given by characterizing descriptions.

For comparing hypotheses and sequential trails, the HypTrails [35] algorithm has been applied to web data and recently to geo-spatial trajectory data [36]. In [2] we have presented the DASHTrails approach that incorporates probability distributions for deriving transitions. Using general weight-attributed network representations, we can infer transition matrices as *graph interpretations*, relying on Markov chain modeling [1], [37] and Bayesian inference [1], [38]. In this paper, we adapt and extend that for detecting community patterns capturing exceptional sequential link trails.

3. Method

We first provide an overview on the proposed approach. Then, we present modeling and estimation in detail.

3.1. Overview

In the following, we provide an overview on the proposed approach for detecting community patterns capturing exceptional (sequential) link trails. Our subject of analysis is given by an attributed graph that models the link trails in the following way: Nodes of the graph denote actors of a social network, e. g., users of a social system or locations in a location-based social network. The edges of the graph model the links between the nodes – as we will see below as transitions between these. As a simple example, we consider a set of users and a set of locations. Each user visits a sequence of locations – in a location-based social network. Then, we are interested in modeling these sequences (of locations), and in detecting exceptional groups of transitions (between locations) w.r.t. users and their properties, respectively. At a music event festival, for example, possible characterizing factors describing certain users groups could be specific music genres. Here, exceptional patterns could include, for example, users being interested in *rock music* and *dance* visiting only a very specific selection of performances in characteristic sequences, compared to the behavior of all users covering the total set of sequential link trails.

For modeling sequences of such actors we resort to a Markov chain approach, and model sequences as first-order Markov chains. Essentially, this comes down to transitions between individual states (corresponding to nodes of a network) where links between nodes make up the respective transitions. The weights of these links are then given according to the respective transition probabilities (observed in the Markov chain). Adapting the modeling principles of the DASHTrails approach that we have presented in [2] to our network formalism, we model transition matrices according to sequential link trails in a first-order Markov chain representation. We assume a discrete set of states Ω corresponding to the nodes of the network (without loss of generality $\Omega = \{1, \ldots, n\}$, $n \in \mathbb{N}, |\Omega| = n$). Then, assuming a certain *network interpretation* of the weights of the edges, we construct transitions between states. Using a *transition modeling function* $\tau : \Omega \times \Omega \to \mathbb{R}$., transitions between sequential states $i, j \in \Omega$ are captured by the elements m_{ij} of the transition matrix M, i. e., $m_{ij} = \tau(i, j)$.

For incorporating properties (or features) into the network (graph), we include edge labels. That is, the edges of the graph (modeling specific transitions between nodes) are labeled according to descriptive properties, e. g., capturing properties of the specific sequences the transitions were derived from. Then, using a specific set of labels we can *select* a set of edges, i. e., all edges having the respective label set, inducing a subgraph, i. e., a community. A community pattern is then given by the respective label set and its corresponding (induced) subgraph, covering a subset of nodes and transitions (i. e., edges), respectively.

3.2. Modeling Patterns

Let L denote a set of *labels*, e. g., binary features. Intuitively, in our context a *pattern* is made up of a collection of labels that are being combined in a conjunction. The pattern can then also be interpreted as a predicate that is true for an object, if the pattern covers the object, i. e., all the labels contained in the pattern are also contained in the description of the object. Then, for a specific *community pattern* P, a community C_P is the set of all objects (e. g., nodes/edges) that are covered by that pattern. In our context, a pattern covers a set of edges, inducing a subgraph.

It is easy to see, that a pattern describes a fixed set of objects (community), while a community can also be described by a set of patterns, if there are different options for covering the objects contained in the community. In the following, we define these concepts more formally.

Definition 1. A (complex) *community pattern* P is given by a set of basic community patterns $P = \{l_1, \ldots, l_m\}$, where $l_i \in L$, $i \in [1; m]$, which is interpreted as a conjunction, i.e., $P(I) = l_1 \wedge \ldots \wedge l_m$, with $length(P) = m$.

We call a pattern P' a *superpattern* (or *refinement*) of a *subpattern* P, iff $P \subset P'$.

Definition 2. A *community (extension)*

$$C_P \coloneqq ext(P) \coloneqq \{o \in O | P(o) = true\}$$

is the set of all objects o from a given universal set O which are covered by the community pattern P.

As search space for description-oriented community detection the set of all possible patterns 2^L is used, that is, all combinations of the basic patterns contained in L. Typically, exceptional model mining approaches apply a general-to-specific search strategy, such that search traverses (complex) patterns and according superpatterns recursively. A similar strategy can be applied for description-oriented community detection, e. g., using the COMODO algorithm [3] as described below, also in our context. For mining community patterns we utilize both the link structure of the attributed weighted graph, as well as its descriptive information, i. e., the label information of the attributed graph.

For ranking a specific pattern, we utilize a quality measure that estimates the interestingness of the pattern.

Definition 3. A *quality measure* $q \colon 2^L \to \mathbb{R}$ maps every pattern in the search space to a real number that reflects the interestingness of a pattern (or the extension of a pattern, respectively).

The result of top-k community detection is the set of the top-k patterns $P_1, \ldots, P_k$, where $P_i \in 2^L$ with the highest interestingness are selected according to the applied quality measure. Since the patterns can contain redundancy in the descriptions, typically redundancy management is applied [4], [17]. A simple but quite effective approach utilizes a minimal improvement filter [39]: A pattern is then removed from the result set, if that set contains a corresponding subpattern, i. e., a pattern that is described by a subset of labels, with a similar quality – within a certain interval – e. g., up to a 1% lower value, or a higher quality.

3.3. Modeling Sequential Trails

In the following, we describe our modeling method for capturing sequential trails in the form of attributed networks (modeled as attributed graphs). We start with a description of modeling the complete network before we tackle the issue of comparing subgraphs induced by community patterns.

Overall, for modeling we map a set of sequential trails to a transition matrix using principles of first-order Markov chain modeling. That transition matrix can then also be interpreted as a weighted adjacency matrix of an (attributed) graph, where the individual values of an entry (i, j) correspond to the weight of the link between nodes i and j; at the same time, this can be interpreted as the transition probability between two states i and j. For our attributed graph model, we label the links according to the descriptive information of the sequential trail. Then, we identify exceptional community patterns based on the labels and structure of the contained links using exceptional model mining.

Reference Model. As outlined above, we derive transition matrices (modeling transitions between states) for a sequential trail using a certain *transition modeling function* $\tau : \Omega \times \Omega \rightarrow \mathbb{R}$. Using τ, we can model transition matrices corresponding to the *observed data*. Concerning all sequential trails that make up our reference model, we construct an according matrix M^N with $m^N_{ij} = \tau(i, j)$: For those observed sequences we can simply construct transition matrices counting the transitions between the individual states. Then, $\tau(i, j) = |suc(i, j)|$, where $suc(i, j)$ denotes the successive sequences from state i to state j contained in the sequence. For constructing more complex transition matrices from a probability distribution over events or subsets, for example, we need to apply a more complex modeling approach. We refer to [2], [35] for more details on modeling and inference, respectively. For comparing the model to matrices induced by community patterns, we can either provide the matrix itself, or use an adapted (e. g., normalized) matrix depending on the requirements of the applied quality measure. We can assess, for example, the model and the community pattern using an approach based on comparing network structures. Furthermore, we can apply a Bayesian approach and compare the model to induced hypotheses.

Community Pattern. A community pattern P induces a subgraph (community) C_P given a set of labels P, selecting all links that are covered, i. e., that share a label contained in P. Then, all transitions in the matrix M^N are selected (corresponding to a set of links of the network) that are covered by the pattern P. Using that, we construct an according transition pattern matrix M^P based on the respective counts of the covered transitions. Intuitively, the matrix M^P can then be regarded as some kind of "projection" of matrix M^N given the pattern P using our modeling approach. In the simplest case, we can just transfer the weighted links of the subgraph C_P. Now, given the (row-normalized, where required) transition matrix for P we need to rank it in relation to the network data and other community patterns. For that, we apply a quality measure as described below.

3.4. Quality Measures

For ranking a set of community patterns, we propose two quality measures for our modeling context.

QAP. The quadratic assignment procedure [40] (QAP) is a standard approach for comparing network structures, e. g., using a graph correlation measure: For comparing two graphs G_1 and G_2, it estimates the correlation of the respective adjacency matrices M_1 and M_2 and tests that graph level statistic against a QAP null hypothesis [40]. QAP compares the observed graph correlation of (G_1, G_2) to the distribution of the respective resulting correlation scores obtained on repeated random row and column permutations of the adjacency matrix of G_2. As a result, we obtain a correlation and a statistical significance level according to the randomized distribution scores. For deriving a quality measure based on QAP and graph correlation, we compare the reference matrix M_N and the matrix M_P for pattern P:

$$q_Q(P) = QAP(M_N, M_P) = \frac{cov(M_N, M_P)}{\sqrt{var(M_N) \cdot var(M_P)}},$$

where M_N is the transition matrix induced by the reference model (as described above), M_P is the transition matrix induced by pattern P, cov indicates the covariance of the matrices, and $var(M) = cov(M, M)$ the variance, e. g., cf., [40] for more details on QAP.

Bayesian Estimation. Using our Markov chain modeling formalism, we can also take a Bayesian modeling view and utilize the community pattern for constructing a *hypothesis*, in order to check how well it explains the behavior of the transitions of the reference model. On the one hand, if the hypothesis does not explain the behavior of the model well, then we observe a *deviating* behavior. On the other hand, if the hypothesis explains the transitions contained in the model well, then we observe *conforming* behavior.

For estimation, we apply the core inference step of DASHTrails [2] on our first-order Markov chain model [35]. As an input, we provide a hypothesis constructed using the row-normalized transition matrix covered by the pattern, containing the transitional information (frequencies) of transitions between the respective states. In principle, in this step we can also include further transformations of the matrix, if required. In addition, we utilize the according transition matrix of the reference model. Following [35], we elicit a conjugate Dirichlet prior given the data (matrix) and finally obtain the evidence using marginal likelihood estimation. Here, the evidence denotes the probability of the model (data) given a specific hypothesis. Thus, this can also be interpreted as the relative plausibility of a hypothesis. Then, the hypotheses can be ranked in terms of their evidence.

A central aspect of the method is an additional parameter (b) indicating the *belief* in a given hypothesis: The higher b the higher the belief in the respective hypothesis matrix. Given a lower value of b the hypothesis is assigned more tolerance, such that other (but similar) parameter configurations become more probable. We then assess the performance

of a hypothesis with increasing b, typically relative to the uniform hypothesis (as a baseline) and further hypotheses.

For obtaining a quality measure, we first estimate the evidence of the hypothesis constructed using pattern P relative to the reference model for a given b:

$$q_E(P, b) = Evidence(M_P, M_N, b) = ML(M_N|M_P^b)\,,$$

where $ML(M_N|M_P^b)$ computes the marginal likelihood (evidence) of the data (M_N) given a hypothesis M_P^b (derived from M_P with belief b), with the transition matrix M_P for the community pattern P and the normalized matrix M_N of the reference model. We refer to [35], [38] for a derivation of $ML(M_N|M_P^b)$ and M_P^b.

Since q_E depends on a parameter b we also need to obtain an overall picture for different values of b indicating different beliefs in our reference model. Typically, different values $b = 1, \ldots, n$ are provided in order to show the trends of the evidence computation. For more details, we refer to [2]. For obtaining a comprehensive view on q_E for a set of values for b we can now combine the different contributions of the q_E values. In order to do that, we compute the *evidence area under the curve* q_C, similar to the *area under the curve (AUC)*, for predictive applications [41]. The quality measure q_C (for a given n) is defined as follows:

$$q_C(P) = q_E(P, 1) + \sum_{b=2}^{n} \frac{q_E(P, b) + q_E(P, b-1)}{2}\,,$$

where M_P denotes the transition matrix for the community pattern P and M_N the normalized matrix of the reference model.

For the assessment of $q_C(P)$ for a community pattern P we can compare it relative to other patterns. Furthermore, we also take into account a random baseline, i. e., the uniform hypothesis (square matrix, all entries being 1). Then, a conforming hypothesis should exhibit large evidence values, i. e., a large q_C value (and accordingly large q_E values). In addition, it should also be "well away" from the random baseline, cf., [1], [35]. For the uniform hypothesis we can obtain according values for the evidence area under the curve analogously, as described above. For different patterns, q_C and q_E can be compared using Bayes factors analysis [42] in order to identify significant differences.

A further option for (interactive) assessment is given by a visualization of the obtained evidences. Here, we plot the distinct evidence values of the respective hypotheses. In addition, we can also plot the evidence area under the curve values into a single plot for a comprehensive visual overview on the relations between the different hypotheses, the pattern hypothesis, and the (uniform) baseline.

4. Results

In the following, we describe two case studies using real-world social network datasets focusing on location-based social networks. We first describe the applied datasets before we present the results of our experiments. After that, we provide a detailed discussion.

4.1. Datasets

We applied two location-based social network datasets. The first dataset is given by a bimodal network of user–performance visits at a distributed event, where a timestamp is assigned to each link accordingly. Then, we can construct sequential (visit) trails for each user given these timestamps.

The second dataset considers environmental noise measurements, in a bimodal network of user–location relations, where here a timestamp is assigned to each measurement (of environmental noise) as well. We can accordingly construct sequential (measurement) trails for each user given the respective measurements and timestamps.

For both datasets, we construct transition matrices as discussed above, and label the transitions of each trail given the respective properties of the users, i. e., interests and tags, respectively. In the resulting attributed graphs a link between two nodes indicates a transition between the respective locations with a probability according to the link's weight.

4.1.1. LNM 2013. The Lange Nacht der Musik (Long Night of Music; LNM), e. g., [43], [44], is an annual cultural event that is organized in the city of Munich. At one evening in May, a diverse range of pubs, discotheques and clubs, and other cultural venues, such as churches and museums are hosting various musical performances. On May 11th 2013, approximately 20,000 people visited a total of 212 available performances at 113 distinct locations, that were dispersed across the city. For supporting the participants, a targeted app for event planning was offered in the app store; a total of 1159 users downloaded and used that app. For location-based assessment, also GPS (Global Positioning System) of the app data was logged for tracking users' actual (time-based) visits. As not all users had their GPS enabled, only the visits of 111 out of the 1159 visitors could be reconstructed. For more details on the dataset and its collection, we refer to [43], [44].

Given that data, we construct a bipartite graph, creating an edge between user u and performance h, if u attended h. We assign the timestamp when user u entered performance h to that edge. There are 245 nodes (111 users and 134 performances) – as the only non-singleton component [44]. Then, we can naturally collect sequential trails by ordering the sets of edges for each user. Thus, the dataset contains 111 trails, with a mean length of about 5 performances. The transitions (links) of these trails are then labeled with labels using categories and descriptions about the performances.

Concerning the descriptive information, i. e., the labels that are assigned to the edges of our final attributed graph modeling the location-based social network, we extracted descriptive information from the textual information of the event: This included *genre* categories as well as descriptions (free text) of the individual performances. We applied typical data preprocessing steps such as stemming and stop word removal, e. g., [45]. On average, 45.50 features are assigned to each trail. We furthermore filtered words below a minimal frequency threshold $\tau = 5$ reducing a total number of 2767 to 180 descriptive words (features) in total.

4.1.2. EveryAware – *WideNoise Plus*. As our second real-world dataset, we furthermore utilize data from the EveryAware[1] project, e. g., [46]. Specifically, we focus on collectively organized noise measurements collected using the *WideNoise Plus* application between December 14, 2011 and June 6, 2014, cf., [47]. *WideNoise Plus* allows the collection of noise measurements using smartphones. It includes sensor data from the microphone given as noise level in dB(A), the location from the GPS-, GSM-, and WLAN-sensor represented as latitude and longitude coordinate, as well as a timestamp. In addition, tags can be assigned to the recording. We collected data from all around the world using iOS and Android devices. The data are stored and processed using the EveryAware backend which is based on the UBICON software platform [46], [48].

The applied dataset contains 6,069 data records, i. e., noise measurements of 635 users (i. e., 635 trails, with an average trail length of about 10) and 2,009 distinct tags: The available tagging information was cleaned such that only tags with a length of at least three characters were considered. Only data records with valid tag assignments were included. Furthermore, we applied stemming and split multi-word tags into distinct single word tags. Each trail contains 34.31 tags on average. We also filtered words below a minimal frequency threshold $\tau = 5$ reducing the total number of tags to 288 features. Concerning the GPS data we identified 249 locations in a grid-based approach.

4.2. Case Studies

Below, we present the results of our case studies on the LNM 2013 and the *WideNoise Plus* datasets. For the pattern detection step, we applied an adapted version of the COMODO algorithm for description-oriented community detection. It aims at discovering the top-k communities (described by community patterns) with respect to a number of standard community evaluation functions [3]. The method is based on a generalized subgroup discovery approach [26], [49] adapted to attributed graph data for detecting description-oriented community patterns. In our setting, COMODO works on the respective transition matrices. The descriptive information is provided in the form of an edge dataset, where a set of labels is assigned to each edge (indicating the incident edges). Then, each row of the dataset contains the edge, as well as the set of labels. Using that data representation, we can apply the proposed quality measures in order to mine the top-k community patterns.

We applied both the q_C and the q_Q measures for evaluation: We considered q_Q as a baseline, since it relies on a well-established technique for detecting associations between adjacency matrices. Therefore, for community detection we applied the proposed q_C measure, and put it into relation to q_Q concerning ranking consistency, and its effectiveness w.r.t. the identification of conforming and deviating patterns. In addition, we also outline a detailed view on the results of q_C using the respective values of q_E.

1. http://www.everyaware.eu

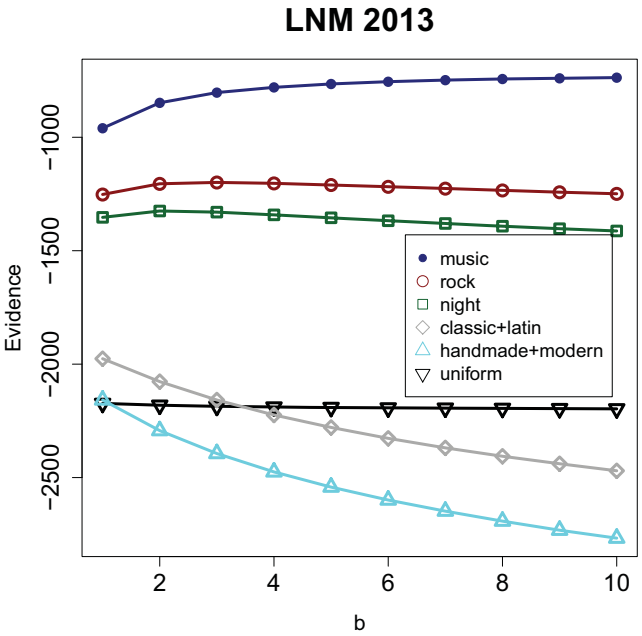

Figure 1: Detailed view on the quality of the community patterns in Table 1. We show the individual quality values obtained using q_E for increasing b (degree of belief in the pattern).

Table 1 shows exemplary exceptional conforming and deviating patterns for the LNM 2013 dataset, using q_C ($n = 10$) and q_Q as quality measures; in addition, it shows the sizes of the covered subsets. When constructing the hypotheses given the community patterns, we applied normalization using the distribution-adapted modeling approach discussed above; transitions were modeled respecting self loops, and reset state, cf., [35]. Figure 1 provides a detailed view on the performance of the patterns regarding q_E (and thus also q_C), and allows a fine-grained analysis for comparing the different patterns. Table 2 and Figure 2 show the respective results for the *WideNoise Plus* dataset.

TABLE 1: Exemplary exceptional conforming/deviating community patterns for LNM 2013. Patterns #1-#3 tend rather to conform to the reference model, while patterns #4-#5 show a deviating behavior.

#	q_C	q_Q	Size	Description
1	-7992	0.99	434	*music*
2	-12240	0.86	310	*rock*
3	-13632	0.81	277	*night*
4	-22478	0.48	73	*classic* $\wedge$ *latin*
5	-24996	0.20	20	*handmade* $\wedge$ *modern*

TABLE 2: Exemplary exceptional conforming/deviating community patterns for *WideNoise Plus*. Patterns #1-#3 tend rather to conform to the reference model (especially #1 and #2), while patterns #4-#5 (increasingly) show a deviating behavior.

#	q_C	q_Q	Size	Description
1	-42326	0.94	5078	*traffic*
2	-61574	0.89	3990	*car*
3	-65589	0.76	3326	*noise*
4	-90381	0.43	707	*bird* $\wedge$ *courtyard*
5	-110520	0.24	600	*background* $\wedge$ *quiet*

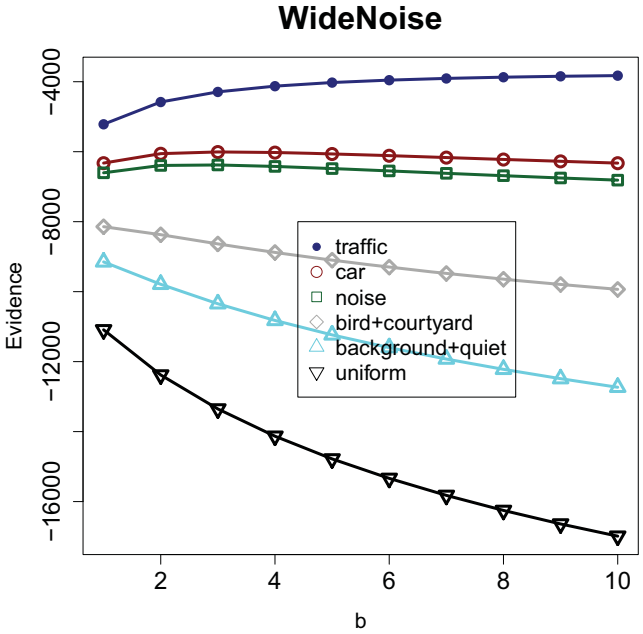

Figure 2: Detailed view on the quality of the community patterns in Table 2. We show the individual quality values obtained using q_E for increasing b (degree of belief in the pattern).

4.3. Discussion

Considering Tables 1-2 we observe that the ranking between q_C and q_Q is consistent for the *conforming* and the *deviating* patterns. Furthermore, both for the conforming as well as the deviating patterns the ranking is always clearly distinguishable using q_E (also being consistent with q_Q). As can be observed especially for the conforming patterns, the evidence plot enables a detailed view on the behavior, as also supported by using Bayes factors analysis [42]. These trends also confirm ranking consistency and validity.

Alltogether, our results demonstrate the potential of the proposed approach and the presented quality measures. Specifically, the quality measure q_C provides a comprehensive view on the patterns, and induces a consistent ranking of the patterns compared to our baseline q_Q utilizing the QAP test. Furthermore, it allows a detailed inspection using the values of q_E for comparing patterns in detail. When the q_Q values are relatively close, for example, q_C still enables a convincing decision on the ranking by visual inspection using Bayes factor analysis [42]. In particular, we can then conveniently apply the evidence plot q_E and compare the individual patterns and their induced transition matrices, respectively, to the uniform transition matrix.

From a qualitative point of view, the patterns shown in the Tables 1-2 are intuitive to interpret and also tend to conform to our expectations concerning the reference behavior of both datasets: For the patterns in Table 1 we observe conformance for labels that are quite general for LNM 2013, while there are deviations for specialized ones. In addition, for the *WideNoise Plus* dataset we observe similar trends concerning noisy and relatively quiet environments.

5. Conclusions

In this paper, we provided a novel problem formalization for detecting descriptive community patterns capturing exceptional sequential trails. We presented a framework for comparing such patterns to a specific reference model, and for identifying the top-k patterns, and proposed suitable quality measures. We demonstrated the applicability of our proposed framework using a case study on two real-world social network datasets. In our experiments, we observed the strengths of the Bayesian inference approach (captured by the novel quality measures q_C and q_E) compared to a baseline (q_Q), also regarding visual inspection. Alltogether, the presented results showed the applicability and benefit of the proposed approach for detecting the top-k patterns and for obtaining a comprehensive view on those. Complemented by flexible visualizations, e. g., [2], [47], the patterns and their ranking can also be inspected for a detailed assessment.

For future work, we aim to extend the analysis using more (diverse) data, e. g., further behavioral [50], industrial, and social media data, also enabled by information extraction methods, e. g., [51]. Furthermore, we aim to include more background knowledge for refining the reference models, e. g., by considering causal relations, e. g., [17], and social distributional approaches, e. g., [29]. Supporting interactive visualization, introspection, and explanation methods (e. g., [52]) are further interesting directions for future work.

Acknowledgments

This work has been supported by the BMBF project FEE under grant number 01IS14006E.

References

[1] 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*, vol. 9, no. 7, 2014.

[2] M. Atzmueller, A. Schmidt, and M. Kibanov, "DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails," in *Proc. WWW 2016 (Companion)*. IW3C2 / ACM, 2016.

[3] M. Atzmueller, S. Doerfel, and F. Mitzlaff, "Description-Oriented Community Detection using Exhaustive Subgroup Discovery," *Information Sciences*, vol. 329, pp. 965–984, 2016.

[4] M. Atzmueller, "Subgroup Discovery," *WIREs: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015.

[5] W. Duivesteijn, A. J. Feelders, and A. Knobbe, "Exceptional Model Mining," *Data Min. Knowl. Disc.*, vol. 30, no. 1, pp. 47–98, 2016.

[6] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, 1st ed., ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1994, no. 8.

[7] S. Fortunato, "Community Detection in Graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[8] M. E. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 69, no. 2, pp. 1–15, 2004.

[9] S. Fortunato and C. Castellano, *Encyclopedia of Complexity and System Science*. Heidelberg: Springer, 2007, ch. Community Structure in Graphs.

[10] G. Palla, I. J. Farkas, P. Pollner, I. Derenyi, and T. Vicsek, "Directed Network Modules," *New J. Phys.*, vol. 9, no. 6, p. 186, 2007.

[11] A. Lancichinetti, S. Fortunato, and J. Kertsz, "Detecting the Overlapping and Hierarchical Community Structure in Complex Networks," *New J. Phys.*, vol. 11, no. 3, 2009.

[12] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 43:1–43:35, Aug. 2013.

[13] C. Bothorel, J. D. Cruz, M. Magnani, and B. Micenkova, "Clustering Attributed Graphs: Models, Measures and Methods," *Network Science*, vol. 3, no. 03, pp. 408–444, 2015.

[14] A. Silva, W. Meira Jr, and M. J. Zaki, "Mining Attribute-Structure Correlated Patterns in Large Attributed Graphs," *Proc. VLDB Endowment*, vol. 5, no. 5, pp. 466–477, 2012.

[15] S. Günnemann, I. Färber, B. Boden, and T. Seidl, "GAMer: A Synthesis of Subspace Clustering and Dense Subgraph Mining," in *Knowledge and Information Systems (KAIS)*. Springer, 2013.

[16] S. Pool, F. Bonchi, and M. van Leeuwen, "Description-driven Community Detection," *TIST*, vol. 5, no. 2, 2014.

[17] M. Atzmueller, *Knowledge-Intensive Subgroup Mining – Techniques for Automatic and Interactive Discovery*, ser. Dissertations in Artificial Intelligence-Infix (Diski). IOS Press, March 2007, vol. 307.

[18] L. Akoglu, H. Tong, and D. Koutra, "Graph Based Anomaly Detection and Description," *Data Min. Knowl. Disc.*, vol. 29, no. 3, 2015.

[19] S. Ranshous, S. Shen, D. Koutra, S. Harenberg, C. Faloutsos, and N. F. Samatova, "Anomaly Detection in Dynamic Networks: A Survey," *WIREs: Computational Stat.*, vol. 7, no. 3, pp. 223–247, 2015.

[20] W. Klösgen, "Explora: A Multipattern and Multistrategy Discovery Assistant," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press, 1996, pp. 249–271.

[21] S. Wrobel, "An Algorithm for Multi-Relational Discovery of Subgroups," in *Proc. 1st European Symposium on PKDD*. Heidelberg, Germany: Springer, 1997, pp. 78–87.

[22] M. Atzmueller and F. Lemmerich, "Fast Subgroup Discovery for Continuous Target Concepts," in *Proc. ISMIS*, ser. LNCS, vol. 5722. Heidelberg, Germany: Springer, 2009, pp. 1–15.

[23] B.-E. Macek, C. Scholz, M. Atzmueller, and G. Stumme, "Anatomy of a Conference," in *Proc. ACM Hypertext*. New York, NY, USA: ACM Press, pp. 245–254.

[24] C. Scholz, M. Atzmueller, and G. Stumme, "On the Predictability of Human Contacts: Influence Factors and the Strength of Stronger Ties," in *Proc. SocialCom*. Boston, MA, USA: IEEE, 2012.

[25] M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme, "Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles," in *Modeling and Mining Ubiquitous Social Media*, ser. LNAI. Heidelberg, Germany: Springer, 2012, vol. 7472.

[26] M. Atzmueller and F. Mitzlaff, "Efficient Descriptive Community Mining," in *Proc. FLAIRS*. AAAI Press, 2011, pp. 459 – 464.

[27] M. Atzmueller and F. Lemmerich, "Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information," *International Journal of Web Science*, vol. 2, no. 1/2, 2013.

[28] M. Atzmueller, "Data Mining on Social Interaction Networks," *Journal of Data Mining and Digital Humanities*, vol. 1, June 2014.

[29] F. Mitzlaff, M. Atzmueller, A. Hotho, and G. Stumme, "The Social Distributional Hypothesis," *SNAM*, vol. 4, no. 216, 2014.

[30] M. Kibanov, M. Atzmueller, C. Scholz, and G. Stumme, "Temporal Evolution of Contacts and Communities in Networks of Face-to-Face Human Interactions," *Sci. Chi. Inf. Sci.*, vol. 57, no. 3, pp. 1–17, 2014.

[31] P. L. Pirolli and J. E. Pitkow, "Distributions of Surfers' Paths Through the World Wide Web: Empirical Characterizations," *World Wide Web*, vol. 2, no. 1-2, 1999.

[32] G. K. Orman, V. Labatut, M. Plantevit, and J.-F. Boulicaut, "A Method for Characterizing Communities in Dynamic Attributed Complex Networks," in *Proc. IEEE/ACM ASONAM*, 2014, pp. 481–484.

[33] F. Mitzlaff, M. Atzmueller, D. Benz, A. Hotho, and G. Stumme, "Community Assessment using Evidence Networks," in *Analysis of Social Media and Ubiquitous Data*, ser. LNAI, vol. 6904, 2011.

[34] F. Mitzlaff, M. Atzmueller, G. Stumme, and A. Hotho, "Semantics of User Interaction in Social Media," in *Complex Networks IV*, ser. SCI. Heidelberg, Germany: Springer, 2013, vol. 476.

[35] P. Singer, D. Helic, A. Hotho, and M. Strohmaier, "Hyptrails: A Bayesian Approach for Comparing Hypotheses about Human Trails," in *Proc. WWW*, ACM. New York, NY, USA: ACM Press, 2015.

[36] M. Becker, P. Singer, F. Lemmerich, A. Hotho, D. Helic, and M. Strohmaier, "VizTrails: An Information Visualization Tool for Exploring Geographic Movement Trajectories," in *Proc. ACM Hypertext*. New York, NY, USA: ACM, 2015, pp. 319–320.

[37] R. Lempel and S. Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect," *Computer Networks*, vol. 33, no. 1, pp. 387–401, 2000.

[38] C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler, "Inferring Markov Chains: Bayesian Estimation, Model Comparison, Entropy Rate, and Out-of-Class Modeling," *Phys. Rev. E*, vol. 76, no. 1, p. 011106, 2007.

[39] R. Bayardo, R. Agrawal, and D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases," *Data Mining and Knowledge Discovery*, vol. 4, pp. 217–240, 2000.

[40] D. Krackhardt, "QAP Partialling as a Test of Spuriousness," *Social Networks*, vol. 9, pp. 171–186, 1987.

[41] C. X. Ling, J. Huang, and H. Zhang, "AUC: A Statistically Consistent and More Discriminating Measure than Accuracy," in *Proc. IJCAI*, vol. 3, 2003, pp. 519–524.

[42] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[43] R. Schaller, M. Harvey, and D. Elsweiler, "Relating User Interaction to Experience During Festivals," in *Proc. Information Interaction in Context Conference*. Palo Alto, CA, USA: ACM, 2014, pp. 38–47.

[44] M. Atzmueller, T. Hanika, G. Stumme, R. Schaller, and B. Ludwig, "Social Event Network Analysis: Structure, Preferences, and Reality," in *Proc. IEEE/ACM ASONAM*. Boston, MA, USA: IEEE Press, 2016.

[45] A. Schmidt, M. Atzmueller, and M. Hollender, "Data Preparation for Big Data Analytics: Methods & Experiences," in *Enterprise Big Data Engineering, Analytics, and Management*. IGI Global, 2016.

[46] M. Atzmueller, M. Becker, M. Kibanov, C. Scholz, S. Doerfel, A. Hotho, B.-E. Macek, F. Mitzlaff, J. Mueller, and G. Stumme, "Ubicon and its Applications for Ubiquitous Social Computing," *NRHM*, vol. 20, no. 1, pp. 53–77, 2014.

[47] M. Atzmueller, J. Mueller, and M. Becker, *Mining, Modeling and Recommending 'Things' in Social Media*, ser. LNAI. Heidelberg, Germany: Springer, 2015, no. 8940, ch. Exploratory Subgroup Analytics on Ubiquitous Data.

[48] M. Atzmueller, M. Becker, S. Doerfel, M. Kibanov, A. Hotho, B.-E. Macek, F. Mitzlaff, J. Mueller, C. Scholz, and G. Stumme, "Ubicon: Observing Social and Physical Activities," in *Proc. IEEE CPSCom*. Washington, DC, USA: IEEE Computer Society, 2012, pp. 317–324.

[49] F. Lemmerich, M. Becker, and M. Atzmueller, "Generic Pattern Trees for Exhaustive Exceptional Model Mining," in *Proc. ECML/PKDD*. Heidelberg, Germany: Springer, 2012.

[50] M. Atzmueller and K. Hilgenberg, "Towards Capturing Social Interactions with SDCF: An Extensible Framework for Mobile Sensing and Ubiquitous Data Collection," in *Proc. MSM 2013, Hypertext 2013*. New York, NY, USA: ACM Press, 2013.

[51] M. Atzmueller, P. Kluegl, and F. Puppe, "Rule-Based Information Extraction for Structured Data Acquisition using TextMarker," in *Proc. LWA 2008*. University of Wuerzburg, 2008.

[52] M. Atzmueller and T. Roth-Berghofer, "The Mining and Analysis Continuum of Explaining Uncovered," in *Proc. 30th SGAI International Conference on Artificial Intelligence (AI-2010)*, 2010.