Descriptive Community Detection

Martin Atzmueller

Abstract Subgroup discovery and community detection are standard approaches for identifying (cohesive) subgroups. This paper presents an organized picture of recent research in descriptive community (and subgroup) detection. Here, it summarizes approaches for the identification of descriptive patterns targeting both static as well as dynamic (sequential) relations. We specifically focus on attributed graphs, i. e., complex relational graphs that are annotated with additional information. This relates to attribute information, for example, assigned to the nodes and/or edges of the graph. Combining subgroup discovery and community detection, we also summarize an efficient and effective approach for descriptive community detection.

1 Introduction

Subgroup discovery (Klösgen, 1996; Wrobel, 1997; Atzmueller, 2015b) aims at identifying interesting descriptive subgroups contained in a dataset - from a compositional network analysis view, aiming at a description given, e. g., by a set of attribute values. The subgroups are identified in such a way that they are interesting with respect to a certain target property. In the context of ubiquitous data and social media, interesting target concepts are given, e. g., by binary variables for obtaining characteristic descriptions of certain phenomena, densely connected graph structures (communities) or exceptional spatio-semantic distributions (Atzmueller, 2014, 2016b). This directly bridges the gap to community detection methods (Newman and Girvan, 2004; Fortunato, 2010; Xie et al, 2013) that focus on structural aspects of a network/graph, for finding densely connected subgroups of nodes.

Tilburg University, Tilburg Center for Cognition and Communication (TiCC),
University of Kassel, Research Center for Information System Design (ITeG)

e-mail: m.atzmuller@uvt.nl

Preprint of: Martin Atzmueller (2017) Descriptive Community Detection.
In: Rokia Missaoui, Sergei Obiedkov, Sergei Kuznetsov (eds.)
Formal Concept Analysis in Social Network Analysis. Springer, Berlin/Heidelberg (In Press)

This paper, an extended and significantly revised version of (Atzmueller, 2015a) presents an organized picture of recent research in subgroup discovery and community detection specifically focusing on attributed graphs. We start with the introduction of necessary background concepts in Section 2. After that, we provide a compact overview on prominent methods for community detection, and discuss the exceptional model mining approach. Next, Section 3 describes recent work on mining attributed graphs for description-oriented approaches. Then, Section 4 summarizes the COMODO algorithm combining both community detection and subgroup discovery in a description-oriented approach (Atzmueller and Mitzlaff, 2011; Atzmueller et al, 2016a), for which we also describe an extension for sequential pattern mining. Finally, we conclude with a summary and point out interesting future directions in Section 5.

2 Subgroup Discovery

In general, subgroup discovery can be applied for any standard dataset in tabular form in a straight-forward manner using available efficient algorithms, e. g., (Atzmueller, 2015b), as implemented in the VIKAMINE (Atzmueller and Puppe, 2005; Atzmueller and Lemmerich, 2012) platform. Also, for compositional analysis of social networks, i. e., where nodes have attached attribute information, we can directly apply subgroup discovery for identifying interesting subgroups of nodes according to a given quality measure. The description space is then given by all the compositional variables and their respective value domains. As we will see below, it is also possible to combine a structural with a compositional analysis of a network, resulting in description-oriented community detection using subgroup discovery.

2.1 Patterns and Subgroups

Basic concepts used in subgroup discovery (Klösgen, 1996; Wrobel, 1997; Atzmueller, 2015b) are patterns and subgroups. Intuitively, a *pattern* describes a *subgroup*, i. e., the subgroup consists of instances that are covered by the respective pattern. It is easy to see, that a pattern describes a fixed set of instances (subgroup), while a subgroup can also be described by different patterns, covering the subgroup' instances. Below, we define these concepts more formally.

A *database* $D = (I, A)$ is given by a set of individuals I and a set of attributes A. A *selector* or *basic pattern* $sel_{a_i=v_j}$ is a Boolean function $I \rightarrow \{0, 1\}$ that is true if the value of attribute $a_i \in A$ is equal to v_j for the respective individual. For a numeric attribute a_{num} whose range is divided into intervals $e_j = [min_j, max_j]$ selectors $sel_{a_{num} \in [min_j; max_j]}$ can be defined for each interval $[min_j; max_j]$ in the domain of a_{num}. The Boolean function is then set to true if the value of attribute a_{num} is within the respective interval. The set of all basic patterns is denoted by S.

Definition 1. A *subgroup description* or (complex) *pattern* sd is given by a set of basic patterns $sd = \{sel_1, \ldots, sel_l\}$, where $sel_i \in S$, which is interpreted as a conjunction, i.e., $sd(I) = sel_1 \wedge \ldots \wedge sel_l$, with $length(sd) = l$.

Without loss of generality, we focus on a conjunctive pattern language using nominal attribute–value pairs as defined above in this paper; internal disjunctions can also be generated by appropriate attribute–value construction methods, if necessary (Atzmueller and Puppe, 2006). We call a pattern p a *superpattern* (or *refinement*) of a *subpattern* p_s, iff $p_s \subset p$.

Definition 2. A *subgroup (extension)*

$$sg_{sd} := ext(sd) := \{i \in I | sd(i) = true\}$$

is the set of all individuals which are covered by the pattern sd.

As search space for subgroup discovery the set of all possible patterns 2^S is used, that is, all combinations of the basic patterns contained in S. Then, appropriate efficient algorithms, e. g., (Atzmueller, 2015b) can be applied.

2.2 Interestingness of a Pattern

A large number of quality functions has been proposed in the literature, see (Geng and Hamilton, 2006) for a comprehensive list, in order to estimate the interestingness of a pattern selected according to the analysis task.

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

Many quality functions for a single target concept (e. g., binary (Klösgen, 1996; Atzmueller, 2015b) or numerical (Atzmueller, 2015b; Lemmerich et al, 2016)), trade off the size $n = |ext(sd)|$ of a subgroup for the deviation $t_{sd} - t_0$, where t_{sd} is the average value of a given target concept in the subgroup identified by the pattern sd and t_0 the average value of the target concept in the general population. In the binary case, the averages relate to the *share* of the target concept. Thus, typical quality functions are of the form

$$q_a(sd) = n^a \cdot (t_{sd} - t_0),\ a \in [0; 1]\,. \tag{1}$$

For binary target concepts, this includes, for example, the *weighted relative accuracy* for the size parameter $a = 1$ or a simplified binomial function, for $a = 0.5$. *Multi-target concepts*, e. g., (Klösgen, 2002a,b; Atzmueller et al, 2015; Atzmueller, 2015b) that define a target concept captured by a set of variables can be defined similarly, e. g., by extending an univariate statistical test to the multivariate case,

e. g., (Atzmueller et al, 2015): Then, the multivariate distributions of a subgroup and the general population are compared in order to identify interesting patterns.

While a quality function provides a *ranking* of the discovered subgroup patterns, often also a statistical assessment of the patterns is useful in data exploration. Quality functions that directly apply a statistical test, for example, the Chi-square quality function, e. g., (Atzmueller, 2015b) provide a p-value for simple interpretation. However, the Chi-square quality function estimates deviations in two directions. An alternative, which can also be directly mapped to a p-value is given by the *adjusted residual* quality function q_r, since the values of q_r follow a large standard normal distribution (Agresti, 2007):

$$q_r = n(t_{sd} - t_0) \cdot \frac{1}{\sqrt{nt_0(1-t_0)(1-\frac{n}{N})}} \tag{2}$$

The result of top-k subgroup discovery is the set of the k patterns $sd_1, \ldots, sd_k$, where $sd_i \in 2^S$, with the highest interestingness according to the applied quality function. A subgroup discovery task can now be specified by the 5-tuple: (D, c, S, q, k) , where c indicates the target concept; the search space 2^S is defined by the set of basic patterns S.

For several quality functions *optimistic estimates* (Grosskreutz et al, 2008; Atzmueller, 2015b) can be applied for determining upper quality bounds: Consider the search for the k best subgroups: If it can be proven that no subset of the currently investigated hypothesis is interesting enough to be included in the result set of k subgroups, then we can skip the evaluation of any subsets of this hypothesis, but can still guarantee the optimality of the result. More formally, an optimistic estimate $\mathrm{oe}(q)$ of a quality function q is a function such that $p \subseteq p' \rightarrow (\mathrm{oe}(q))(p) \geq q(p')$, i. e., such that no refinement p' of the pattern p can exceed the quality obtained by $(\mathrm{oe}(q))(p)$.

2.3 Community Detection

Communities and cohesive subgroups have been extensively studied in social sciences, e. g., using social network analysis methods (Wasserman and Faust, 1994). Community detection methods can be classified according to several dimensions, e. g., disjoint vs. overlapping communities. Here, actors in a network can only belong to exactly one community, or to multiple communities at the same time. Furthermore, we distinguish between methods that work on extended (attributed) graphs, i. e., including descriptive information about the nodes. Below, we provide an overview on representative methods, including several basic methods working on simple graphs. After that, we elaborate on methods for detecting overlapping communities, before we focus on descriptive methods.

2.3.1 Basics of Community Detection

Wasserman and Faust (1994) discuss social network analysis in depth and provide an overview on the analysis of subgroups/communities in graphs, including clique-based, degree-based and matrix-perturbation-based methods. Furthermore, several algorithms for community detection have been proposed, formalizing the notions of interesting community structures, and introducing the modularity quality measure (Newman, 2004; Newman and Girvan, 2004; Newman, 2006). Fortunato (2010) presents a thorough survey on the state of the art community detection algorithms in graphs, focussing on detecting *disjoint* communities.

For assessing the quality of a community, usually not only the density of the community is assessed but the connection density of the community is compared to the density of the rest of the network (Newman and Girvan, 2004). For the modularity measure the number of connections within the community is compared to the statistically "expected" number based on all available connections in the network. Besides modularity, prominent examples of community quality measures include for example, the segregation index (Freeman, 1978) and the inverted average out-degree fraction (Yang and Leskovec, 2012).

2.3.2 Detecting Overlapping Communities

Overlapping communities allow an extended modeling of actor–actor relations in social networks: Nodes of a corresponding graph can then participate in multiple communities. This is also typically observed in real-world networks regarding different complementary facets of social interactions (Palla et al, 2005). A general overview on algorithms for overlapping community detection is provided by Xie et al. Xie et al (2013). For example, clique percolation methods proposed in (Palla et al, 2005, 2007) detect k-cliques and then merge them into overlapping communities. Xie and Szymanski (2013) present methods that extend the idea of label propagation (Raghavan et al, 2007). (Lancichinetti et al, 2009) describe an approach for overlapping and hierarchical community structure using a local community metric. The presented metric itself is computed locally but still assesses a global clustering. Further statistical and local optimization algorithms include the COPRA (Gregory, 2010) algorithm by Gregory using label-propagation of neighboring nodes until a consensus is reached, and the MOSES (McDaid and Hurley, 2010) algorithm by McDaid and Hurley using statistical model-based techniques. Concerning quality measures, extensions of the modularity metric for handling overlapping communities are described in (Muff et al, 2005; Nicosia et al, 2009; Lin et al, 2009).

2.4 Exceptional Model Mining

A general framework for multi-target quality functions in subgroup discovery is given by *exceptional model mining* (Leman et al, 2008; Atzmueller, 2015b): It tries to identify interesting patterns with respect to a local model derived from *a set* of attributes. The interestingness can be defined, e.g., by a significant deviation from a model that is derived from the total population or the respective complement set of instances within the population.

In general, a model consists of a specific *model class* and *model parameters* which depend on the values of the model attributes in the instances of the respective pattern cover. The quality measure q then determines the interestingness of a pattern according to its model parameters. Following (Lemmerich et al, 2012), we outline some simple examples below, focusing on relations between pairs (correlation) and sets of variables (logistic regression):

- A relatively simple example for an exceptionality measure considers the task of identifying subgroups in which the correlation between two numeric attributes is especially strong, e. g., as measured by the Pearson correlation coefficient. This *correlation model class* has exactly one parameter, i.e., the correlation coefficient.
- Furthermore, using a *simple linear regression model*, we can compare the slopes of the regression lines of the subgroup to the general population or the subgroups' complement. This *simple linear regression model* shows the dependency between two numeric variables x and y: It is built by fitting a straight line in the two dimensional space by minimizing the squared residuals e_j of the model:

  $$y_i = a + b \cdot x_i + e_j$$

  The slope

  $$b = \frac{cov(x, y)}{var(x)}$$

  computed given the covariance $cov(x, y)$ of x and y, and the variance $var(x)$ of x can then be used for identifying interesting patterns (Leman et al, 2008).
- The *logistic regression model* is used for the classification of a binary target attribute $y \in T$ from a set of independent binary attributes $x_j \in T \setminus y, j = 1, \ldots, |T| - 1$. The model is given by:

  $$y = \frac{1}{1 + e^{-z}},\ z = b_0 + \sum_j b_j x_j \,.$$

  Interesting patterns are then those, for example, for which the model parameters b_j differ significantly from those derived from the total population.

Considering network structures, we can also adapt exceptional model mining to that setting. Essentially, it can be regarded as a description-oriented approach for assessing network structures, if the patterns are used to induce graphs or subgraphs. As we will discuss below, we can then also apply exceptional model mining for descriptive community detection, in essence combining subgroup discovery and community detection into a unified approach.

Below, we first outline a quality function for comparing graph structures that correspond to individual patterns (QAP). After that, we discuss quality functions used in community detection in order to assess subgraphs that are induced by some criterion, e. g., by a descriptive pattern.

For some notation, we follow the notions presented in (Atzmueller et al, 2016a): As outlined above, the concept of a *community* intuitively describes a group C of individuals out of a population such that members of C are strongly "related" among each other but weakly "related" to individuals outside of C. By intuition, this relates, for example, to strongly connected groups of actors in social networks. This idea translates to communities as vertex sets $C \subseteq V$ of a graph $G = (V, E)$. To determine the amount of relatedness (or connectedness, and thus, the community quality of such a subset) several measures have been proposed.

For further concepts regarding our terminology and also the standard community quality functions outlined below, we follow the notation introduced in (Atzmueller et al, 2016a): For a given undirected graph $G = (V, E)$ and a community $C \subseteq V$: $n := |V|$, let $m := |E|$, $n_C := |C|$, $m_C := |\{\{u, v\} \in E : u, v \in C\}|$ – the number of *intra-edges* of C, and $\bar{m}_C := |\{\{u, v\} \in E : |\{u, v\} \cap C| = 1\}|$ – the number of *inter-edges* of C. Here, it is also convenient to introduce an *inter-degree* for a node $u \in C$ (that depends on the choice of C) by $\bar{\mathrm{d}}_C(u) := |\{\{u, v\} \in E : v \notin C\}|$, counting the number of edges between u and nodes outside of C, and $\mathrm{d}(u) =:= |\{\{u, v\} \in E\}|$ is the degree of node u.

There is a wide range of different community evaluation functions $2^V \rightarrow \mathbb{R}$ for estimating the community quality. In the context of this paper, we focus on *maximizing* local quality functions for single communities (which are induced by specific patterns). Therefore, we consider the inverse of a quality measure in those cases, where the measure itself indicates higher quality by lower values.

- Concerning network structures, we can compare adjacency matrices induced by a specific pattern, see (Atzmueller, 2016a). For the assessment we can apply, for example, the quadratic assignment procedure (Krackhardt, 1987) (QAP): it 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 (Krackhardt, 1987).
  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 some reference model (see (Atzmueller et al, 2016d; Atzmueller, 2016a)), and 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

For an in-depth description of QAP, we refer to (Krackhardt, 1987). Furthermore, for the transition matrix, we refer to (Atzmueller et al, 2016c,d) for more details on the matrix construction step.

- Regarding the quality of a subgraph induced by a pattern, we can adapt the well known modularity measure to the idea of assessing the induced subgraph captured by a local pattern, i. e., a community pattern (with an associated subgroup description).

  In general, the *modularity* MOD (Newman, 2004; Newman and Girvan, 2004; Newman, 2006) of a graph clustering with k communities $C_1, \ldots, C_k \subseteq V$ focuses on the number of edges *within* a community and compares that with the *expected* such number given a null-model (i.e., a corresponding random graph where the node degrees of G are preserved). It is given by

$$\text{MOD} = \frac{1}{2m} \sum_{u,v \in V} \left(A_{u,v} - \frac{\text{d}(u)\,\text{d}(v)}{2m} \right) \delta(C(u), C(v)) , \tag{3}$$

  where $C(i)$ denotes for $i \in V$ the community to which node i belongs. $\delta(C(u), C(v))$ is the *Kronecker delta* symbol that equals 1 if $C(u) = C(v)$, and 0 otherwise. So, the *modularity* assesses the community quality of a graph partitioning, but can also be adapted to overlapping communities, e. g., (Muff et al, 2005; Nicosia et al, 2009; Lin et al, 2009) for considering the complete graph structure.

  For exceptional model mining, however, we need to consider individual patterns. In order to focus on a subgraph induced by a pattern, the *modularity contribution* of a single community C in a *local context* (subgraph induced by the nodes contained in the community C) can then be computed (Newman, 2006; Nicosia et al, 2009) as:

$$\text{MODL}(C) = \frac{1}{2m} \sum_{u,v \in C} \left(A_{u,v} - \frac{\text{d}(u)\,\text{d}(v)}{2m} \right) ,$$

  yielding

$$\text{MODL}(C) = \frac{2m_C}{2m} - \sum_{u,v \in C} \frac{\text{d}(u)\,\text{d}(v)}{4m^2} = \frac{m_C}{m} - \sum_{u,v \in C} \frac{\text{d}(u)\,\text{d}(v)}{4m^2} .$$

- The *segregation index* SIDX (Freeman, 1978) is another prominent measure from community detection. It focuses on the local contribution of the pattern, and compares the number of expected inter-edges to the number of observed inter-edges, normalized by the expectation:

$$\mathrm{SIDX}(C) = \frac{E(\bar{m}_C) - \bar{m}_C}{E(\bar{m}_C)} = 1 - \frac{\bar{m}_C n(n-1)}{2mn_C(n-n_C)} \tag{4}$$

- Finally, the *Inverse Average-ODF (out-degree fraction)* IAODF (Yang and Leskovec, 2012) captures the basic intuition of a community regarding the contained vs. the outgoing edges discussed above. As another local measure, IAODF compares the number of *inter-edges* to the number of all edges of a community C, and averages this for the whole community by considering the fraction for each individual node:

$$\mathrm{IAODF}(C) := 1 - \frac{1}{n_C} \sum_{u \in C} \frac{\bar{\mathrm{d}}_{\mathrm{C}}(u)}{\mathrm{d}(u)} \tag{5}$$

3 Community Detection and Description

While the community detection methods described above only focus on the graph structure, richer graph representations, i. e., *attributed graphs*, enable approaches that specifically exploit the descriptive information of the labels assigned to nodes and/or edges of the graph. Nodes of a network representing users, for example, can be labeled with tags that the respective users utilized in social bookmarking systems, or nodes (denoting actors) can be labeled with properties of the latter. Then, *explicit descriptions* for the characterization of a community can be provided.

Concerning methods that focus on such descriptions in general, an approach for community detection using features identified by frequent pattern mining is presented in (Adnan et al, 2009); closed frequent patterns are derived and are then used for creating a social network model based on an entropy analysis. However, the network structure itself is not exploited. Similarly, Sese et al (2010) extracts subgraphs with common itemsets. Given a labeled graph, itemset-sharing subgraphs can then be enumerated. However, this approach also does not consider the density of graphs, nor any community measures.

Focusing on methods for generating *explicit descriptions connected with the graph structure*, we distinguish between two types of approaches: first, methods that mainly work on the graph structure but apply descriptive information for restricting the possible sets of communities; second, methods that mine descriptive patterns for obtaining community candidates evaluated using the graph structure. As a representative of the first type, the concepts of dense subgraphs and subspace clusters for mining cohesive patterns are combined in (Moser et al, 2009).

Starting with quasi-cliques, these are expanded until constraints regarding the description or the graph structure are violated. Similarly, Günnemann et al (2013) combines subspace clustering and dense subgraph mining, also interleaving quasi-clique and subspace construction. As an example for the second type outlined above, Galbrun et al (2014) proposes an approach for the problem of finding overlapping communities in graphs and social networks that aims at detecting the top-k communities such that the total edge density over all k communities is maximized. The three algorithmic variants proposed in Galbrun et al (2014) apply a greedy strategy for detecting dense subgroups, and restrict the result set of communities, such that each edge can belong to at most one community. This partitioning involves a global approach on the community quality. Furthermore, Silva et al (2012) study the correlation between attribute sets and the occurrence of dense subgraphs in large attributed graphs. The proposed method considers frequent attribute sets using an adapted frequent item mining technique, and identifies the top-k dense subgraphs induced by a particular attribute set, called structural correlation patterns. The DCM method presented in (Pool et al, 2014) includes a two-step process of community detection and community description. A heuristic approach is applied for discovering the top-k communities. Pool et al. utilize a special interestingness function which is based on counting outgoing edges of a community similar to the IAODF measure; for that, they also demonstrate the trend of a correlation with the modularity function.

Furthermore, the COMODO algorithm (Atzmueller et al, 2016a) that we summarize in the next section combines community detection and subgroup discovery resulting in a description-oriented approach. By specifying a standard quality function the quality of the communities to discover can be estimated. Then, this quality function can be specifically selected according to the analysis task.

4 Community Detection using Exceptional Model Mining

For providing both structurally valid and interpretable communities we utilize the graph structure as well as additional descriptive features of the nodes. Hence, we identify communities as sets of nodes together with a *description* composed of the nodes' features. Such a *community pattern* then provides an intuitive description of the community, e. g., by an easily interpretable conjunction of attribute-value pairs. Basically, we aim at identifying communities according to standard community quality measures. Below, we first provide an algorithmic overview on the approach and summarize exemplary evaluation results. After that, we sketch the application of the algorithm for community detection on dynamic networks, i. e., for identifying exceptional sequential patterns.

4.1 COMODO*: Description-Oriented Community Detection*

Below, we summarize the COMODO algorithm presented in (Atzmueller et al, 2016a): It focuses on *description-oriented community detection* using subgroup discovery, and aims at discovering the top-k communities (described by community patterns). The method is based on an adapted subgroup discovery approach (Atzmueller and Mitzlaff, 2011; Lemmerich et al, 2012), and also tackles typical problems that are not addressed by standard approaches for community detection such as pathological cases like small community sizes. COMODO utilizes optimistic estimates (Grosskreutz et al, 2008; Wrobel, 1997), which are efficient to compute, in order to prune the search space significantly. For that, a number of standard community evaluation functions have been applied using optimistic estimates for an efficient approach.

4.1.1 Algorithmic Overview

COMODO utilizes both the graph structure, as well as descriptive information of the attributed graph. This information is contained in two data structures: The graph structure is encoded in graph G while the attribute information is contained in database D describing the respective attribute values of each node. In a preprocessing step, we merge these data sources. Since the communities considered in our approach do not contain isolated nodes, we can describe them as sets of edges. We transform the data (of the given graph G and the database D containing the nodes' descriptive information) into a new data set focusing on the edges of the graph G: Each data record in the new data set represents an edge between two nodes. The attribute values of each such data record are the common attributes of the edge's two nodes. For a more detailed description, we refer to (Atzmueller et al, 2016a).

COMODO utilizes an extended FP-tree (frequent pattern tree) structure inspired by the FP-growth algorithm, which compiles the data in a convenient prefix pattern tree structure for mining frequent item sets, see (Agrawal and Srikant, 1994) for a detailed description. Our adapted tree structure is called the *community pattern tree* (CP-tree) that allows to efficiently traverse the solution space. The tree is built in two scans of the graph data set and is then mined in a recursive divide-and-conquer manner, see (Atzmueller and Lemmerich, 2009; Lemmerich et al, 2012) for more details. In the main algorithmic procedure of COMODO, patterns containing only one basic pattern are mined first. Then, patterns conditioned on the occurrence of a (prefixed) complex pattern (as a set of basic patterns, chosen in the previous recursion step) are considered recursively. For more algorithmic details, we refer to (Atzmueller et al, 2016a). As described there, we can apply standard quality functions efficiently using optimistic estimates, e. g., for the *modularity* or the *segregation index*, see (Atzmueller et al, 2016a) for more details.

4.1.2 Illustrative Evaluation Results

Below, we present illustrative evaluation results (Atzmueller et al, 2016a) considering the efficiency of the applied optimistic estimates, and the validity of the obtained patterns. For that, we compared the total number of search steps, that is community allocations that are considered by the COMODO algorithm, with no optimistic estimate pruning to optimistic estimate pruning using different commmunity quality measures. Additionally, we measured the impact of using different minimal community size thresholds. Some results are shown in Figure 1 for the BibSonomy click graph for $k = 10, 20, 50$ and minimal size thresholds $\tau_n = 10, 20$. We consider a number of standard community quality functions, that is, the *segregation index*, the *Inverse Average-ODF*, and the *modularity*.

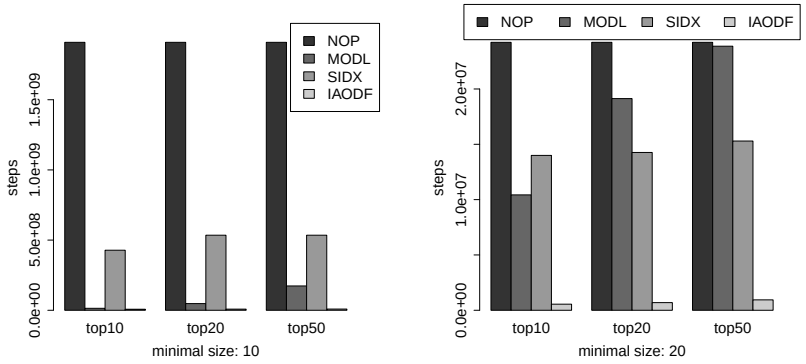

Fig. 1 Runtime performance of COMODO on the BibSonomy click graph, see (Atzmueller et al, 2016a) for more details: Search steps with no optimistic estimate pruning (*NOP*) vs. community quality functions with optimistic estimate pruning: MODL (Local modularity), SIDX (Segregation Index) and IAODF (Inverse Average-ODF), for minimal size thresholds $\tau_n = 10, 20$.

The large, exponential search space can be exemplified, e. g., for the click graph with a total of about $2 \cdot 10^{10}$ search steps for a minimal community size threshold $\tau_n = 10$. The results demonstrate the effectiveness of the proposed descriptive mining approach applying the presented optimistic estimates. The implemented pruning scheme makes the approach scalable for larger data sets, especially when the local modularity quality function is chosen to assess the communities' quality. Concerning the validity of the patterns, we focused on structural properties of the patterns and the subgraphs induced by the respective comunity patterns. We applied the significance test described in (Koyuturk et al, 2007) for testing the statistical significance of the density of a discovered subgraph. Furthermore, we compared COMODO to three baseline community detection algorithms (McDaid and Hurley, 2010; Gregory, 2010; Pool et al, 2014), where COMODO consistently shows a significantly better performance concerning validity and description length; for more details, we refer to (Atzmueller et al, 2016a).

4.2 Sequential Pattern Analysis: Detecting Exceptional Link Trails

In addition to static community detection, we can also consider temporal aspects, i. e., focusing on sequences of states or events which can be applied for a variety of analysis ranging from the analysis of human behavior (Atzmueller et al, 2016c) to industrial applications (Atzmueller et al, 2016d). In an extended modeling approach, we can map transitions between states to a weighted network, according to a first order Markov chain model. Below, we outline an approach for detecting exceptional sequential link trails captured by community patterns, see (Atzmueller, 2016a) for a detailed description.

As before, 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 transitions). As a simple example, we can 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 and their sequential link trails. Essentially, we apply descriptive community detection (e. g., using COMODO) on the attributed graph, where the edges indicate transitions between states according to a first-order Markov chain modeling approach (Lempel and Moran, 2000; Singer et al, 2014).

4.2.1 Modeling

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. In particular, we assess a pattern capturing a set of nodes that model the state space of the respective transitions.

For constructing a reference model, we construct transition matrices corresponding to the *observed data*. For those observed sequences we can simply construct transition matrices counting the transitions between the individual states. We construct an according matrix M^N with $m^N_{ij} = |suc(i,j)|$, where $suc(i,j)$ denotes the successive sequences from state i to state j contained in the sequence.

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 transi-

tions. 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. For identifying exceptional models (M_P induced by P) we can then apply, e. g., the QAP quality function $q_Q(P) = QAP(M_N, M_P)$ introduced above.

4.2.2 Results

For some illustrative results (see (Atzmueller, 2016a) for more details), we utilized data from the EveryAware[1] project, e. g., (Atzmueller et al, 2014). Specifically, we focused on collectively organized noise measurements collected using the *WideNoise Plus* application between December 14, 2011 and June 6, 2014, see Atzmueller et al (2015) for more details. *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.

Table 1 Illustrative 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_Q	Size	Description
1	0.94	5078	*traffic*
2	0.89	3990	*car*
3	0.76	3326	*noise*
4	0.43	707	*bird* $\wedge$ *courtyard*
5	0.24	600	*background* $\wedge$ *quiet*

In total, 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. Table 1 shows exemplary exceptional conforming and deviating patterns using q_Q as quality measure. In addition, it shows the sizes of the covered subsets. From a qualitative point of view, the patterns shown in the table are intuitive to interpret and also tend to conform to our expectations concerning the reference behavior of the dataset, where we can clearly identify deviations concerning noisy and relatively quiet environments.

[1] `http://www.everyaware.eu`

5 Conclusions

In this paper, we have presented an organized view on descriptive community detection. Specifically, we described subgroup discovery for compositional network analysis concerning properties of the actors, with extensions to the analysis of complex target concepts like correlations between a set of variables, or dense subgraphs – captured by exceptional model mining approaches. Then, this directly extends to community detection on attributed graphs. In particular, we summarized the COMODO algorithm that combines community detection and exceptional model mining, resulting in a description-oriented approach for community analytics. We furthermore sketched an extension to dynamic data, considering sequential patterns capturing exceptional sequential link trails. This adds one further dimension to the descriptive approaches, by considering by static as well as dynamic phenomena, and enables the modeling and investigation of complex analysis tasks.

For future work, we aim to extend the analysis towards further time-oriented representations, e. g., considering sequences of graphs, and the evolution of communities, e. g., (Kibanov et al, 2014, 2015). Also, we aim to integrate and exploit methods for generating descriptions and the respective relations in link analytics, e. g., in link prediction (Scholz et al, 2012, 2013a,b) on multiplex networks. Then, besides the detection of communities, also their analysis and assessment in the form of descriptive patterns is highly relevant, e. g., (Atzmueller et al, 2005, 2006; Atzmueller and Puppe, 2008; Atzmueller and Lemmerich, 2013) also concerning their semantic grounding (Mitzlaff et al, 2013, 2014), and integration into explanation-aware approaches (Clancey, 1983; Roth-Berghofer and Cassens, 2005; Atzmueller and Roth-Berghofer, 2010). Furthermore, developing scalable methods for enabling such approaches for large and complex datasets, e. g., (Lemmerich et al, 2012; Atzmueller et al, 2016b) are another interesting direction for future work.

References

Adnan M, Alhajj R, Rokne J (2009) Identifying Social Communities by Frequent Pattern Mining. In: Proc. 13th Intl. Conf. Information Visualisation, IEEE Computer Society, Washington, DC, USA, pp 413–418

Agrawal R, Srikant R (1994) Fast Algorithms for Mining Association Rules. In: Bocca JB, Jarke M, Zaniolo C (eds) Proc. 20th Int. Conf. Very Large Data Bases, (VLDB), Morgan Kaufmann, pp 487–499

Agresti A (2007) An Introduction to Categorical Data Analysis. Wiley-Blackwell

Atzmueller M (2014) Data Mining on Social Interaction Networks. Journal of Data Mining and Digital Humanities 1

Atzmueller M (2015a) Subgroup and Community Analytics on Attributed Graphs. In: Kuznetsov SO, Missaoui R, Obiedkov S (eds) Proc. International Workshop on Social Network Analysis using Formal Concept Analysis (SNAFCA-2015), CEUR-WS, vol 1534

Atzmueller M (2015b) Subgroup Discovery – Advanced Review. WIREs: Data Mining and Knowledge Discovery 5(1):35–49

Atzmueller M (2016a) Detecting Community Patterns Capturing Exceptional Link Trails. In: Proc. IEEE/ACM ASONAM, IEEE Press, Boston, MA, USA

Atzmueller M (2016b) Local Exceptionality Detection on Social Interaction Networks. In: Proc. ECML-PKDD 2016: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer Verlag, Berlin

Atzmueller M, Lemmerich F (2009) Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. International Symposium on Methodologies for Intelligent Systems, Springer, Heidelberg, Germany, LNCS, vol 5722, pp 1–15

Atzmueller M, Lemmerich F (2012) VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, Heidelberg, Germany

Atzmueller M, Lemmerich F (2013) Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. International Journal of Web Science 2(1/2)

Atzmueller M, Mitzlaff F (2011) Efficient Descriptive Community Mining. In: Proc. 24th International FLAIRS Conference, AAAI Press, Palo Alto, CA, USA, pp 459 – 464

Atzmueller M, Puppe F (2005) Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science 11(11):1752–1765

Atzmueller M, Puppe F (2006) SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In: Proc. European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Springer, Heidelberg, Germany, pp 6–17

Atzmueller M, Puppe F (2008) A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. Journal of Applied Intelligence 28(3):210–221

Atzmueller M, Roth-Berghofer T (2010) The Mining and Analysis Continuum of Explaining Uncovered. In: Proc. 30th SGAI International Conference on Artificial Intelligence (AI-2010)

Atzmueller M, Baumeister J, Hemsing A, Richter EJ, Puppe F (2005) Subgroup Mining for Interactive Knowledge Refinement. In: Proc. 10th Conference on Artificial Intelligence in Medicine (AIME 05), Springer, Heidelberg, Germany, LNAI 3581, pp 453–462

Atzmueller M, Baumeister J, Puppe F (2006) Introspective Subgroup Analysis for Interactive Knowledge Refinement. In: Proc. 19th International Florida Artificial Intelligence Research Society Conference 2006 (FLAIRS-2006), AAAI Press, Palo Alto, CA, USA, pp 402–407

Atzmueller M, Becker M, Kibanov M, Scholz C, Doerfel S, Hotho A, Macek BE, Mitzlaff F, Mueller J, Stumme G (2014) Ubicon and its Applications for Ubiquitous Social Computing. New Review of Hypermedia and Multimedia 20(1):53–77

Atzmueller M, Mueller J, Becker M (2015) Mining, Modeling and Recommending 'Things' in Social Media, Springer, Heidelberg, Germany, chap Exploratory Subgroup Analytics on Ubiquitous Data. No. 8940 in LNAI

Atzmueller M, Doerfel S, Mitzlaff F (2016a) Description-Oriented Community Detection using Exhaustive Subgroup Discovery. Information Sciences 329:965–984

Atzmueller M, Mollenhauer D, Schmidt A (2016b) Big Data Analytics Using Local Exceptionality Detection. In: Enterprise Big Data Engineering, Analytics, and Management, IGI Global, Hershey, PA, USA

Atzmueller M, Schmidt A, Kibanov M (2016c) DASHTrails: An Approach for Modeling and Analysis of Distribution-Adapted Sequential Hypotheses and Trails. In: Proc. WWW 2016 (Companion), IW3C2 / ACM

Atzmueller M, Schmidt A, Kloepper B, Arnu D (2016d) HypGraphs: An Approach for Modeling and Comparing Graph-Based and Sequential Hypotheses. In: Proc. ECML-PKDD Workshop on New Frontiers in Mining Complex Patterns (NFMCP), Riva del Garda, Italy

Clancey WJ (1983) The Epistemology of a Rule-Based Expert System: A Framework for Explanation. Artificial Intelligence 20:215–251

Fortunato S (2010) Community Detection in Graphs. Physics Reports 486(3-5):75 – 174

Freeman L (1978) Segregation In Social Networks. Sociological Methods & Research 6(4):411

Galbrun E, Gionis A, Tatti N (2014) Overlapping Community Detection in Labeled Graphs. Data Min Knowl Discov 28(5-6):1586–1610

Geng L, Hamilton HJ (2006) Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys 38(3)

Gregory S (2010) Finding Overlapping Communities in Networks by Label Propagation . New J Phys (12)

Grosskreutz H, Rüping S, Wrobel S (2008) Tight Optimistic Estimates for Fast Subgroup Discovery. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, Heidelberg, Germany, LNCS, vol 5211, pp 440–456

Günnemann S, Färber I, Boden B, Seidl T (2013) GAMer: A Synthesis of Subspace Clustering and Dense Subgraph Mining. In: Knowledge and Information Systems, Springer

Kibanov M, Atzmueller M, Scholz C, Stumme G (2014) Temporal Evolution of Contacts and Communities in Networks of Face-to-Face Human Interactions. Science China 57

Kibanov M, Atzmueller M, Illig J, Scholz C, Barrat A, Cattuto C, Stumme G (2015) Is Web Content a Good Proxy for Real-Life Interaction? A Case Study Considering Online and Offline Interactions of Computer Scientists. In: Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE Press, Boston, MA, USA

Klösgen W (1996) Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in Knowledge Discovery and Data Mining, AAAI Press, pp 249–271

Klösgen W (2002a) Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, chap 16.3: Subgroup Discovery

Klösgen W (2002b) Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, chap 5.2: Subgroup Patterns

Koyuturk M, Szpankowski W, Grama A (2007) Assessing Significance of Connectivity and Conservation in Protein Interaction Networks. Journal of Computational Biology 14(6):747–764

Krackhardt D (1987) QAP Partialling as a Test of Spuriousness. Social Networks 9:171–186

Lancichinetti A, Fortunato S, Kert J (2009) Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New Journal of Physics 11(3)

Leman D, Feelders A, Knobbe A (2008) Exceptional Model Mining. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, Lecture Notes in Computer Science, vol 5212, pp 1–16

Lemmerich F, Becker M, Atzmueller M (2012) Generic Pattern Trees for Exhaustive Exceptional Model Mining. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer, Heidelberg, Germany

Lemmerich F, Atzmueller M, Puppe F (2016) Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. Data Mining and Knowledge Discovery 30:711–762

Lempel R, Moran S (2000) The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. Computer Networks 33(1):387–401

Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. CoRR abs/0810.1355

Lin YR, Chi Y, Zhu S, Sundaram H, Tseng BL (2009) Analyzing Communities and Their Evolutions in Dynamic Social Networks. ACM Trans Knowl Discov Data 3:8:1–8:31

McDaid A, Hurley N (2010) Detecting Highly Overlapping Communities with Model-Based Overlapping Seed Expansion. In: Proc. International Conference on Advances in Social Networks Analysis and Mining, IEEE Computer Society, Washington, DC, USA, ASONAM, pp 112–119

Mitzlaff F, Atzmueller M, Stumme G, Hotho A (2013) Semantics of User Interaction in Social Media. In: Ghoshal G, Poncela-Casasnovas J, Tolksdorf R (eds) Complex Networks IV, Studies in Computational Intelligence, vol 476, Springer, Heidelberg, Germany

Mitzlaff F, Atzmueller M, Hotho A, Stumme G (2014) The Social Distributional Hypothesis. Journal of Social Network Analysis and Mining 4(216)

Moser F, Colak R, Rafiey A, Ester M (2009) Mining Cohesive Patterns from Graphs with Feature Vectors. In: SDM, SIAM, vol 9, pp 593–604

Muff S, Rao F, Caflisch A (2005) Local Modularity Measure for Network Clusterizations. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 72(5):056,107

Newman ME, Girvan M (2004) Finding and Evaluating Community Structure in Networks. Phys Rev E Stat Nonlin Soft Matter Phys 69(2):1–15

Newman MEJ (2004) Detecting Community Structure in Networks. Europ Physical J 38

Newman MEJ (2006) Modularity and Community Structure in Networks. Proceedings of the National Academy of Sciences 103(23):8577–8582

Nicosia V, Mangioni G, Carchiolo V, Malgeri M (2009) Extending the Definition of Modularity to Directed Graphs with Overlapping Communities. J Stat Mech p 03024

Palla G, Deri I, Farkas I, Vicsek T (2005) Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature 435(7043):814–818

Palla G, Farkas IJ, Pollner P, Derenyi I, Vicsek T (2007) Directed Network Modules. New Journal of Physics 9(6):186

Pool S, Bonchi F, van Leeuwen M (2014) Description-driven Community Detection. Transactions on Intelligent Systems and Technology 5(2)

Raghavan U, R A, Kumara S (2007) Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. Phys Rev E 76:036106

Roth-Berghofer TR, Cassens J (2005) Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In: Muñoz-Avila H, Ricci F (eds) Case-Based Reasoning Research and Development, 6th International Conference on Case-Based Reasoning, ICCBR 2005, Chicago, IL, USA, August 2005, Proceedings, Springer Verlag, Heidelberg, no. 3620 in Lecture Notes in Artificial Intelligence LNAI, pp 451–464

Scholz C, Atzmueller M, Stumme G (2012) On the Predictability of Human Contacts: Influence Factors and the Strength of Stronger Ties. In: Proc. 4th ASE/IEEE International Conference on Social Computing (SocialCom), IEEE Computer Society, Boston, MA, USA

Scholz C, Atzmueller M, Barrat A, Cattuto C, Stumme G (2013a) New Insights and Methods For Predicting Face-To-Face Contacts. In: Kiciman E, Ellison NB, Hogan B, Resnick P, Soboroff I (eds) Proc. International AAAI Conference on Weblogs and Social Media, AAAI Press, Palo Alto, CA, USA

Scholz C, Atzmueller M, Kibanov M, Stumme G (2013b) How Do People Link? Analysis of Contact Structures in Human Face-to-Face Proximity Networks. In: Proc. ASONAM 2013, ACM Press, New York, NY, USA

Sese J, Seki M, Fukuzaki M (2010) Mining Networks with Shared Items. In: Proc. 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, pp 1681–1684

Silva A, Meira Jr W, Zaki MJ (2012) Mining Attribute-Structure Correlated Patterns in Large Attributed Graphs. Proc VLDB Endowment 5(5):466–477

Singer P, Helic D, Taraghi B, Strohmaier M (2014) Detecting Memory and Structure in Human Navigation Patterns Using Markov Chain Models of Varying Order. PLOS ONE 9(7)

Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications, 1st edn. No. 8 in Structural Analysis in the Social Sciences, Cambridge University Press

Wrobel S (1997) An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery, Springer, Heidelberg, Germany, pp 78–87

Xie J, Szymanski BK (2013) LabelRank: A Stabilized Label Propagation Algorithm for Community Detection in Networks. In: Proc. IEEE Network Science Workshop, West Point, NY

Xie J, Kelley S, Szymanski BK (2013) Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study. ACM Comput Surv 45(4):43:1–43:35

Yang J, Leskovec J (2012) Defining and Evaluating Network Communities Based on Ground-truth. In: Proc. ACM SIGKDD Workshop on Mining Data Semantics, ACM, New York, NY, USA, MDS '12, pp 3:1–3:8