In: Proceedings of the International Workshop on Social Network Analysis using Formal Concept Analysis (SNAFCA 2015), CEUR Workshop Proceedings, CEUR-WS.org. 2015

Subgroup and Community Analytics on Attributed Graphs

Martin Atzmueller

University of Kassel
Knowledge and Data Engineering Group
Wilhelmshöher Allee 73, 34121 Kassel, Germany

atzmueller@cs.uni-kassel.de

Abstract. Subgroup discovery and community detection are two approaches having been studied in different research areas like data mining and social network analysis. In this context, these techniques are especially helpful in order to provide for analytical and explorative data mining approaches. We present an organized picture of recent research in subgroup discovery and community detection specifically focusing on attributed graphs. That is, we include complex relational graphs that are annotated with additional information, e.g., attribute information on the nodes and/or edges of the graph. In addition, we especially summarize a method combining both community detection and subgroup discovery resulting in a description-oriented approach for community analytics.

1 Introduction

Subgroup discovery [5,23,49] and community detection [15,37,51] are especially helpful in order to provide for analytical and explorative data mining approaches.

Subgroup discovery aims at identifying interesting descriptive subgroups contained in a dataset - from a compositional network analysis view, aimining 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 [4], interesting target concepts are given, e. g., by binary variables for obtaining characteristic descriptions of certain phenomena [10], densely connected graph structures (communities) [7] or exceptional spatio-semantic distributions [12]. This directly bridges the gap to community detection methods that focus on structural aspects of a network/graph, for finding densely connected subgroups of nodes.

This paper 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, Section 3 provides a compact overview on prominent methods for community detection, also including recent work on mining attributed graphs. In addition, we specifically summarize the COMODO algorithm combining both community detection and subgroup discovery in a description-oriented approach [7, 11]. Finally, we conclude with a summary and point out interesting future directions in Section 5.

2 Subgroup Discovery and Analytics

Below, we first introduce some basic notation. After that, we provide a brief summary of fundamental concepts with respect to subgroup discovery. We discuss basic interestingness measures and also show extensions to more complicated target concepts using exceptional model mining.

2.1 Basic Notation

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. The set of all basic patterns is denoted by S.

For a numeric attribute a_{num} selectors $sel_{a_{num} \in [min_j; max_j]}$ can be defined analogously 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 range.

2.2 Patterns and Subgroups

Basic elements used in subgroup discovery [3, 5, 23, 49] 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, if there are different options for covering the subgroup' instances. In the following, we define these concepts more formally.

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. 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., [8, 13, 30] can be applied.

2.3 Interestingness of a Pattern

A large number of quality functions has been proposed in literature, cf.. [18] for estimating the interestingness of a pattern – selected according to the analysis task.

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

Many quality functions for a single target concept (e. g., binary [5, 23] or numerical [5, 28]), trade-off the size $n = |ext(sd)|$ of a subgroup and 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., [5, 12, 24, 50] 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., [12]: Then, the multivariate distributions of a subgroup and the general population are compared in order to identify interesting (and exceptional) 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., [5] 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, cf. [2]:

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

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 set of basic patterns S.

For several quality functions *optimistic estimates* [5, 8, 21, 28] 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 $p \subseteq p' \to \mathrm{oe}(q(p)) \geq q(p')$, i. e., such that no refinement p' of the pattern p can exceed the quality obtained by $\mathrm{oe}(q(p))$.

2.4 Exceptional Model Mining

A general framework for multi-target quality functions in subgroup discovery is given by *exceptional model mining* [5,27]: 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 [29], we outline some examples below.

- A 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 coveriance $cov(x, y)$ of x and y, and the variance $var(x)$ of x can then be used for identifying interesting patterns, cf. [27].
- The *logistic regression model* is used for the classification of a binary target attribute $y \in T$ from a set of independent binary attributes $x_j \in T \backslash y, j = 1, \ldots, |T| - 1$. 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.

2.5 Subgroup Discovery in Social Network Analysis

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., [8, 13, 30], as implemented in the VIKAMINE [9] system. 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, i. e., combining structural and compositional aspects into a quality function, resulting in description-oriented community detection using subgroup discovery.

3 A Brief Overview on Community Detection

Communities and cohesive subgroups have been extensively studied in social sciences, e. g., using social network analysis methods [48]. 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.

3.1 Basics of Community Detection

Wasserman and Faust [48] 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, Newman et al. [37–39] propose several algorithms for community detection, formalizing the notions of interesting community structures, and introducing the modularity quality measure. Fortunato [15] 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 community's density is assessed but the connection density of the community is compared to the density of the rest of the network [37]. The core idea of the evaluation function is to apply an objective evaluation criterion, for example, for the modularity measure the number of connections within the community 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 [16] and the inverted average out-degree fraction [53].

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 [34, 41]. A general overview on algorithms for overlapping community detection is provided by Xie et al. [51]. For example, clique percolation methods proposed by Palla et al. [41, 42] detect k-cliques and then merge them into overlapping communities. Xie and Szymanski [52] present methods extending the idea of label propagation [44]. Lancichinetti et al. [26] 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 [20] algorithm by Gregory using label-propagation of neighboring nodes until a consensus is reached, and the MOSES [33] 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 [32, 36, 40].

3.3 Community Detection and Description

While the methods described above only focus on the graph structure for mining communities, 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. Then, *explicit descriptions* for the characterization of a community can be provided. Concerning methods that focus on such descriptions in general, Adnan et al. [1] present an approach for community detection using features identified by frequent pattern mining; 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. [46] extract 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, Moser et al. [35] combine the concepts of dense subgraphs and subspace clusters for mining cohesive patterns. Starting with quasi-cliques, these are expanded until constraints regarding the description or the graph structure are violated. Similarly, Günnemann et al. [19] combine 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. [17] propose an approach for the problem of finding overlapping communities in graphs and social networks that aims to detect the top-k communities such that the total edge density over all k communities is maximized. The three algorithmic variants proposed by Galbrun et al. apply a greedy strategy for detecting dense subgroups, and restrict the result set of communities, such that each edge can belong to at most community. This partitioning involves a global approach on the community quality. Silva et al. [47] 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 by Pool et al. [43] 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 [7] that we summarize in the next section combines community detection and subgroup discovery resulting in a description-oriented approach. It allows the specification of a standard quality function for estimating the quality of the communities to discover. This quality function can be selected (or also be specifically modeled) according to the analysis task.

4 Combining Community Detection and Subgroup Discovery

The COMODO algorithm presented in [7] focuses on *description-oriented community detection* using subgroup discovery. 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. Basically, we aim at identifying communities according to standard community quality measures, while providing characteristic descriptions at the same time.

4.1 Algorithmic Overview

The COMODO algorithm for description-oriented community detection aims at discovering the top-k communities (described by community patterns) with respect to a number of standard community evaluation functions. The method is based on an adapted subgroup discovery approach [11, 29], and also tackles typical problems that are not addressed by standard approaches for community detection such as pathological cases like small community sizes. COMODO is a fast branch-and-bound algorithm utilizing optimistic estimates [21, 49] which are efficient to compute. This allows COMODO to prune the search space significantly. As discussed above, COMODO utilizes both the graph structure, as well as descriptive information of the attributed graph, i. e., the label information of the nodes. 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 [7].

The FP-growth algorithm (cf. [22]) for mining association rules, and the SD-Map* algorithm for fast exhaustive subgroup discovery [8] form the basis of COMODO. COMODO utilizes an extended FP-tree structure, called the *community pattern tree* (CP-tree) 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, cf. [8, 29]. In the main algorithmic procedure of COMODO, first patterns containing only one basic pattern are mined. Then recursively, patterns conditioned on the occurrence of a (prefixed) complex pattern (as a set of basic patterns, chosen in the previous recursion step) are considered. For more algorithmic details, we refer to [7].

As outlined in [7] we can compute standard quality functions efficiently, e. g., for the *Modularity* [37–39] or the *Segregation Index* [16], using according optimistic estimates.

4.2 Exemplary Evaluation Results

The evaluation of COMODO considers two aspects: The efficiency of the applied optimistic estimates, and the validity of the obtained community patterns. In order to evaluate the efficiency, we count the number of search steps, i. e., community allocations that are considered by the COMODO algorithm. We compared the total number of search steps (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. Exemplary 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: The *segregation index* [16], the *inverse average ODF (out degree fraction)* [31], and the *modularity* [37].

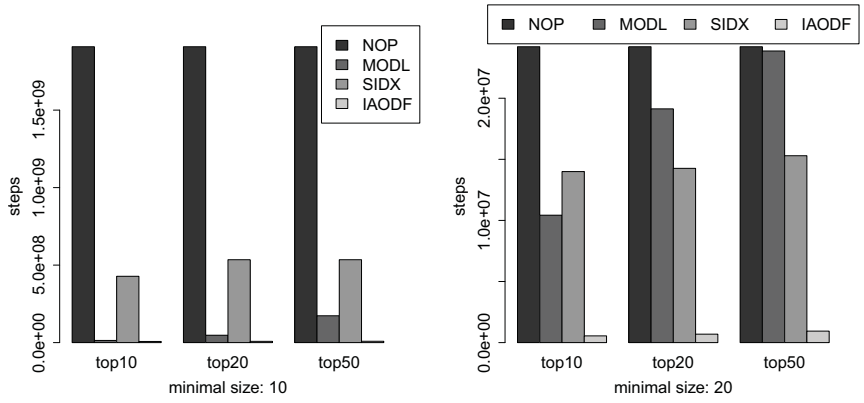

Fig. 1. Runtime performance of COMODO on the BibSonomy click graph [7]: Search steps with no optimistic estimate pruning (*NOP*) vs. community quality functions with optimistic estimate pruning: MODL (Local Modularity), SIDX (Segregation Index) and IAODF (Inverse Average-ODF), for minimal size thresholds $\tau_n = 10, 20$.

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 [25] for testing the statistical significance of the density of a discovered subgraph. Furthermore, we compared COMODO to three baseline community detection algorithms [20,33,43], where COMODO consistently shows a significantly better performance concerning validity and description length (for more details, we refer to [7]).

5 Conclusions and Outlook

In this paper, we have presented an organized view on subgroup and community analytics on attributed graphs. 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. Then, this directly extends to community analytics on attributed graphs. Here, we started with an introduction of basic methods for community detection, continuing on methods for mining overlapping communities, to approaches that target descriptions leveraging structural and compositional attribute information. In particular, we summarized the COMODO algorithm that combines subgroup discovery and community detection, resulting in a description-oriented approach for community analytics.

For future work, we aim to extend the analysis towards time-oriented representations, e. g., considering sequences of graphs. Also, we aim to integrate and exploit methods for generating descriptions and the respective relations in link analytics, e. g., in link prediction [45] on multidimensional networks. Further interesting directions for future work are given by methods support integrated visual exploration and analytics, also including semi-automatic approaches for assessment of the results, e. g., [6, 14].

References

1. Adnan, M., Alhajj, R., Rokne, J.: Identifying Social Communities by Frequent Pattern Mining. In: Proc. 13th Intl. Conf. Information Visualisation. pp. 413–418. IEEE Computer Society, Washington, DC, USA (2009)
2. Agresti, A.: An Introduction to Categorical Data Analysis. Wiley-Blackwell (2007)
3. Atzmueller, M.: Knowledge-Intensive Subgroup Mining – Techniques for Automatic and Interactive Discovery, DISKI, vol. 307. IOS Press (March 2007)
4. Atzmueller, M.: Mining Social Media: Key Players, Sentiments, and Communities. WIREs: Data Mining and Knowledge Discovery 1069 (2012)
5. Atzmueller, M.: Subgroup Discovery – Advanced Review. WIREs: Data Mining and Knowledge Discovery 5(1), 35–49 (2015)
6. Atzmueller, M., Baumeister, J., Hemsing, A., Richter, E.J., Puppe, F.: Subgroup Mining for Interactive Knowledge Refinement. In: Proc. 10th Conference on Artificial Intelligence in Medicine (AIME 05). pp. 453–462. LNAI 3581, Springer, Heidelberg, Germany (2005)
7. Atzmueller, M., Doerfel, S., Mitzlaff, F.: Description-Oriented Community Detection using Exhaustive Subgroup Discovery. Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.05.008
8. Atzmueller, M., Lemmerich, F.: Fast Subgroup Discovery for Continuous Target Concepts. In: Proc. International Symposium on Methodologies for Intelligent Systems. LNCS, vol. 5722, pp. 1–15. Springer, Heidelberg, Germany (2009)
9. Atzmueller, M., Lemmerich, F.: 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 (2012)
10. Atzmueller, M., Lemmerich, F., Krause, B., Hotho, A.: Who are the Spammers? Understandable Local Patterns for Concept Description. In: Proc. 7th Conference on Computer Methods and Systems. Oprogramowanie Nauko-Techniczne, Krakow, Poland (2009)
11. Atzmueller, M., Mitzlaff, F.: Efficient Descriptive Community Mining. In: Proc. 24th International FLAIRS Conference. pp. 459 – 464. AAAI Press, Palo Alto, CA, USA (2011)

12. Atzmueller, M., Mueller, J., Becker, M.: Mining, Modeling and Recommending 'Things' in Social Media, chap. Exploratory Subgroup Analytics on Ubiquitous Data. No. 8940 in LNAI, Springer, Heidelberg, Germany (2015)
13. Atzmueller, M., Puppe, F.: SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In: Proc. European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). pp. 6–17. Springer, Heidelberg, Germany (2006)
14. Atzmueller, M., Puppe, F.: A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. Journal of Applied Intelligence 28(3), 210–221 (2008)
15. Fortunato, S.: Community Detection in Graphs. Physics Reports 486(3-5), 75 – 174 (2010)
16. Freeman, L.: Segregation In Social Networks. Sociological Methods & Research 6(4), 411 (1978)
17. Galbrun, E., Gionis, A., Tatti, N.: Overlapping Community Detection in Labeled Graphs. Data Min. Knowl. Discov. 28(5-6), 1586–1610 (Sep 2014)
18. Geng, L., Hamilton, H.J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys 38(3) (2006)
19. Günnemann, S., Färber, I., Boden, B., Seidl, T.: GAMer: A Synthesis of Subspace Clustering and Dense Subgraph Mining. In: Knowledge and Information Systems. Springer (2013)
20. Gregory, S.: Finding Overlapping Communities in Networks by Label Propagation . New J. Phys. (12) (2010)
21. Grosskreutz, H., Rüping, S., Wrobel, S.: Tight Optimistic Estimates for Fast Subgroup Discovery. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. LNCS, vol. 5211, pp. 440–456. Springer, Heidelberg, Germany (2008)
22. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns Without Candidate Generation. In: Chen, W., Naughton, J., Bernstein, P.A. (eds.) Proc. SIGMOD. pp. 1–12. ACM Press (05 2000)
23. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI Press (1996)
24. Klösgen, W.: Handbook of Data Mining and Knowledge Discovery, chap. 16.3: Subgroup Discovery. Oxford University Press, New York (2002)
25. Koyuturk, M., Szpankowski, W., Grama, A.: Assessing Significance of Connectivity and Conservation in Protein Interaction Networks. Journal of Computational Biology 14(6), 747–764 (2007)
26. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New Journal of Physics 11(3) (2009)
27. Leman, D., Feelders, A., Knobbe, A.: Exceptional Model Mining. In: Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol. 5212, pp. 1–16. Springer (2008)
28. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast Exhaustive Subgroup Discovery with Numerical Target Concepts. Data Mining and Knowledge Discovery (2015 (accepted))
29. Lemmerich, F., Becker, M., Atzmueller, M.: 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 (2012)
30. Lemmerich, F., Rohlfs, M., Atzmueller, M.: Fast Discovery of Relevant Subgroup Patterns. In: Proc. Intl. FLAIRS Conference. pp. 428–433. AAAI Press, Palo Alto, CA, USA (2010)
31. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. CoRR abs/0810.1355 (2008)
32. Lin, Y.R., Chi, Y., Zhu, S., Sundaram, H., Tseng, B.L.: Analyzing Communities and Their Evolutions in Dynamic Social Networks. ACM Trans. Knowl. Discov. Data 3, 8:1–8:31 (April 2009)

33. McDaid, A., Hurley, N.: Detecting Highly Overlapping Communities with Model-Based Overlapping Seed Expansion. In: Proc. International Conference on Advances in Social Networks Analysis and Mining. pp. 112–119. ASONAM, IEEE Computer Society, Washington, DC, USA (2010)
34. Mitzlaff, F., Atzmueller, M., Hotho, A., Stumme, G.: The Social Distributional Hypothesis. Journal of Social Network Analysis and Mining 4(216) (2014)
35. Moser, F., Colak, R., Rafiey, A., Ester, M.: Mining Cohesive Patterns from Graphs with Feature Vectors. In: SDM. vol. 9, pp. 593–604. SIAM (2009)
36. Muff, S., Rao, F., Caflisch, A.: Local Modularity Measure for Network Clusterizations. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 72(5), 056107 (2005)
37. Newman, M.E., Girvan, M.: Finding and Evaluating Community Structure in Networks. Phys Rev E Stat Nonlin Soft Matter Phys 69(2), 1–15 (2004)
38. Newman, M.E.J.: Detecting Community Structure in Networks. Europ Physical J 38 (2004)
39. Newman, M.E.J.: Modularity and Community Structure in Networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
40. Nicosia, V., Mangioni, G., Carchiolo, V., Malgeri, M.: Extending the Definition of Modularity to Directed Graphs with Overlapping Communities. J. Stat. Mech. p. 03024 (2009)
41. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature 435(7043), 814–818 (June 2005)
42. Palla, G., Farkas, I.J., Pollner, P., Derenyi, I., Vicsek, T.: Directed Network Modules. New Journal of Physics 9(6), 186 (2007)
43. Pool, S., Bonchi, F., van Leeuwen, M.: Description-driven Community Detection. Transactions on Intelligent Systems and Technology 5(2) (2014)
44. Raghavan, U., R., A., Kumara, S.: Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. Phys Rev E 76:036106 (2007)
45. Scholz, C., Atzmueller, M., Barrat, A., Cattuto, C., Stumme, G.: New Insights and Methods For Predicting Face-To-Face Contacts. In: Kiciman, E., Ellison, N.B., Hogan, B., Resnick, P., Soboroff, I. (eds.) Proc. International AAAI Conference on Weblogs and Social Media. AAAI Press, Palo Alto, CA, USA (2013)
46. Sese, J., Seki, M., Fukuzaki, M.: Mining Networks with Shared Items. In: Proc. 19th ACM International Conference on Information and Knowledge Management. pp. 1681–1684. ACM, New York, NY, USA (2010)
47. Silva, A., Meira Jr, W., Zaki, M.J.: Mining Attribute-Structure Correlated Patterns in Large Attributed Graphs. Proc. VLDB Endowment 5(5), 466–477 (2012)
48. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. No. 8 in Structural Analysis in the Social Sciences, Cambridge University Press, 1 edn. (1994)
49. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery. pp. 78–87. Springer, Heidelberg, Germany (1997)
50. Wrobel, S., Morik, K., Joachims, T.: Maschinelles Lernen und Data Mining. Handbuch der Künstlichen Intelligenz 3, 517–597 (2000)
51. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study. ACM Comput. Surv. 45(4), 43:1–43:35 (Aug 2013)
52. Xie, J., Szymanski, B.K.: LabelRank: A Stabilized Label Propagation Algorithm for Community Detection in Networks. In: Proc. IEEE Network Science Workshop. West Point, NY (April 2013)
53. Yang, J., Leskovec, J.: Defining and Evaluating Network Communities Based on Ground-truth. In: Proc. ACM SIGKDD Workshop on Mining Data Semantics. pp. 3:1–3:8. MDS '12, ACM, New York, NY, USA (2012)