Community Assessment using Evidence Networks

Folke Mitzlaff[1], Martin Atzmueller[1], Dominik Benz[1],
Andreas Hotho[2], and Gerd Stumme[1]

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

[2] University of Wuerzburg, Data Mining and Information Retrieval Group
Am Hubland, 97074 Wuerzburg, Germany

{mitzlaff, atzmueller, benz, stumme}@cs.uni-kassel.de, hotho@informatik.uni-wuerzburg.de

Abstract. Community mining is a prominent approach for identifying (user) communities in social and ubiquitous contexts. While there are a variety of methods for community mining and detection, the effective evaluation and validation of the mined communities is usually non-trivial. Often there is no evaluation data at hand in order to validate the discovered groups.

This paper proposes an approach for (relative) community assessment. We introduce a set of so-called *evidence networks* which are capturing typical interactions in social network applications. Thus, we are able to apply a rich set of implicit information for the evaluation of communities. The presented evaluation approach is based on the idea of reconstructing existing social structures for the assessment and evaluation of a given clustering. We analyze and compare the presented approach applying user data from the real-world social bookmarking application BibSonomy. The results indicate that the evidence networks reflect the relative rating of the explicit ones very well.

1 Introduction

With the rise of social applications, a wealth of data is stored, and finding relevant entries in the overwhelming user generated data repositories becomes more and more of a problem. Personalizing the access to such systems is a key approach for preventing users to get "lost in data". Promising approaches for such personalization are *user recommendation* or *community mining* techniques. Knowing a user's peer group is, for example, used to adjust search results according to his or her interests [17], or for showing the latest activities or most popular resources only for a set of relevant users.

Parallel to the rise of the Social Web, mobile phones became more and more powerful and are equipped with more and more sensors, giving rise to Mobile Web applications. Today, we observe the amalgamation of these two trends, leading to a Ubiquitous Web, whose applications will support us in many aspects of the daily life at any time and any place. Data are now available which were never accessible before. We expect therefore that the approach presented in this paper will be extendable to ubiquitous applications especially to sensor networks as well.

However, ultimately judging whether users are related or not is rather difficult since any given pair of users shares some properties. Hence, it is usually non-trivial to objectively assess the quality of a given community. As a consequence, Siersdorfer proposed an evaluation paradigm which is based on the notion of reconstructing *an existing social structure* [31], for example, considering friendship links in social applications. Transferring this paradigm to the evaluation of community mining techniques, we propose to assess a given community structure using *a set of existing social structures*. To this end, we introduce a set of so-called *evidence networks* which are capturing typical interactions in social network applications. These interactions can be regarded as proxies for social relations between users, i. e., as implicit connections. Thus, evidence networks provide *evidences* of social relations, but do not require explicit interaction and linking.

Our framework allows to compare the community structure that is computed with a given community mining algorithm with these evidence networks. It provides thus a method for evaluating and comparing different community mining approaches. For the assessment, we apply standard community quality functions, e. g., modularity [26] and conductance [10].

Existing "social structures" are often sparse (compared to the large number of evaluation objects present in clustering and community detection approaches) and usually not publicly available. In addition to explicit networks (e. g., by adding someone as a "friend"), this work also analyzes relations which are *implicitly* acquired in a typical "Web 2.0" application (e. g., by visiting a user's profile page). Our hypothesis in using these networks, which we call evidence networks, is the assumption, that the set of social interactions is drawn from a certain "social population", thus the interactions indicate connections in this distribution, and they manifest themselves with varying degree in different networks. By considering samples of such a "social constellation", we aim to collect evidences for the underlying user relatedness.

Considering implicit evidence networks for evaluation encompasses several advantages. In every application where users may interact, there are implicit evidence networks, even if no explicit user relationship is being implemented. Implicit networks may also be captured anonymously on a client network's proxy server. Typically implicit networks are also significantly larger than explicit networks. Similar interaction networks accrue in the context of ubiquitous applications (e. g., users which are using a given service at the same place and time). Unfortunately no dataset containing such interaction was available during the evaluation, but these interactions lead to implicit user relationships which naturally fit into the framework of evidence networks described in this section.

This paper proposes an approach for the evaluation of communities using implicit information formalized in evidence networks. Our context is given by social applications such as social networking, social bookmarking, and social resource sharing systems. The proposed evaluation paradigm is based on the notion of *reconstructing existing social structures*: This paradigm suggests to measure the quality of a given division of the users by assessing the corresponding community structure in an existing social structure: We basically project the different clusters according to the division of users on an existing network, and assess the created structures using measures for community evaluation. The contribution of this paper is *not* the presentation of a new algorithm for detecting communities. Instead, we rather focus on a better understanding of what a good user community is and how to assess and evaluate a given community allocation:

– We propose evidence networks for community assessment and evaluation. These evidence networks are thoroughly analyzed with respect to the contained community structure.
– We apply standard community evaluation measures using the set of evidence networks. It is shown, that there is a strong common community structure across different evidence networks.
– The results suggest a basis for a new evaluation framework of community detection methods in social applications.

The context of the presented analysis is given by social applications such as social networking, social bookmarking, and social resource sharing systems, considering our own system BibSonomy[3][5] as an example. But the presented analysis is not only relevant for the evaluation of community mining techniques, but also concerning their usage for building new community detection or user recommendation algorithms, among others.

The rest of the paper is structured as follows: Section 2 first describes basic notions of the presented approach. Then, it describes this task in detail and introduces our novel approach together with the concept of evidence networks and their characteristics. After that, we analyze and compare in Section 4 the features of the networks using data from the real-world BibSonomy system: We define different emerging networks of user relatedness, and analyze these in detail applying our proposed method. Finally, Section 5 concludes the paper with a summary and interesting directions for future work.

2 Evidence Networks for Community Evaluation

In the following, we briefly introduce basic notions, terms and measures used in this paper. For more details, we refer to standard literature, e. g., [13]. After that, we describe and define several explicit and implicit networks for the evaluation of communities. Finally, we discuss related work.

2.1 Preliminaries

This section summarizes basic notions and terms with respect to graphs, explicit and implicit relations, communities, and community measures.

A *graph* $G = (V, E)$ is an ordered pair, consisting of a finite set V which consists of the *vertices* or *nodes*, and a set E of *edges*, which are two element subsets of V. A *directed graph* is defined accordingly: E denotes a subset of $V \times V$. For simplicity, we write $(u, v) \in E$ in both cases for an edge belonging to E and freely use the term *network* as a synonym for a graph. The *degree* of a node in a network measures the number of connections it has to other nodes. The *adjacency matrix* $A_{ij}, i = 1 \dots n, j = 1 \dots n$ of a set of nodes S with $n = |S|$ contained in a graph measures the number of connections of node $i \in S$ to node $j \in S$. A *path* $v_0 \to_G v_n$ of *length* n in a graph G is

[3] http://www.bibsonomy.org/

a sequence $v_0, \ldots, v_n$ of nodes with $n \geq 1$ and $(v_i, v_{i+1}) \in E$ for $i = 0, \ldots, n-1$. A *shortest path* between nodes u and v is a path $u \rightarrow_G v$ of minimal length. The *transitive closure* of a graph $G = (V, E)$ is given by $G^* = (V, E^*)$ with $(u, v) \in E^*$ iff there exists a path $u \rightarrow_G v$. A *strongly connected component (scc)* of G is a subset $U \subseteq V$, such that $u \rightarrow_{G^*} v$ exists for every $u, v \in U$. A *(weakly) connected component (wcc)* is defined accordingly, ignoring the direction of edges $(u, v) \in E$.

For a set V, we define a *relation* R as a subset $R \subseteq V \times V$. A relation R is naturally mapped to a corresponding graph $G_R := (V, R)$. We say that a relation R among individuals U is *explicit*, if $(u, v) \in R$ only holds, when at least one of u, v *deliberately* established a connection to the other (e. g., user u added user v as a friend in an online social network). We call R *implicit*, if $(u, v) \in R$ can be *derived* from other relations, e. g., it holds as a side effect of the actions taken by u and v in a social application. Explicit relations are thus given by explicit links, e. g., existing links between users. Implicit relations can be derived or constructed by analyzing secondary data.

The concept of a *community* is vague and can be intuitively defined as a group $\mathcal{C}$ of individuals out of a population $\mathcal{U}$ such that members of $\mathcal{C}$ are densely "related" one to each other but sparsely "related" to individuals in $\mathcal{U} \setminus \mathcal{C}$. In the following, a *community allocation* of a population $\mathcal{U}$ refers to a set of communities $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ with $\bigcup_{1 \leq i \leq n} \mathcal{C}_i \subseteq \mathcal{U}$ and $\mathcal{C}_i \neq \emptyset$ for $1 \leq i \leq n$. Note that this also allows *overlapping communities*, i. e., $\mathcal{C}_i \cap \mathcal{C}_j \neq \emptyset$ may hold for some $i, j \in \{1, \ldots, n\}$.

This concept maps to vertex sets $C \subseteq V$ in graphs $G = (V, E)$ where nodes in C are densely connected but sparsely connected to nodes in $V \setminus C$. Though defined in terms of graph theory, the community concept remains vague. For a given graph $G = (V, E)$ and a community $C \subseteq V$ we set $n := |V|$, $m := |E|$, $n_C := |C|$, $m_C := |\{(u, v) \mid u, v \in C\}|$, $\overline{m}_C := |\{(u, v) \mid u \in C, v \notin C\}|$ and for a node $u \in V$ its degree is denoted by $d(u)$. Several approaches for formalizing communities in graphs exist and corresponding community structures were observed and analyzed in a variety of different networks [27, 26, 21, 22].

In the context of evaluation measures for evidence networks we consider two measures: *Conductance* [22] and *Modularity* [26]. These consider the evaluation from two different perspectives. Modularity mainly focuses on the links *within* communities, while the conductance also takes the links between communities into account.

Conductance can be defined as the ratio between the number of edges within the community and the number of edges leaving the community. Thus, the conductance $C(S)$ of a set of nodes S is given by $C(S) = c_S/(2m_S + c_S)$ where c_S denotes the size of the edge boundary, $c_S := |\{(u, v) : u \in S, v \notin S\}|$ and m_S denotes the number of edges within S, $m_S := |\{(u, v) \in E : u, v \in S\}|$. More community-like partitions exhibit a low conductance, cf. [22]. The conductance of a set of clusters is then given by the average of the conductance of the single clusters.

The modularity function is based on comparing the number of edges within a community with the expected such number given a null-model (i. e., a randomized model). Thus, the modularity of a community clustering is defined to be the fraction of the edges that fall within the given clusters minus the expected such fraction if edges were distributed at random. This can be formalized as follows: The modularity $M(S)$ of a set of nodes S in graph G with its assigned adjacency matrix $A \in \mathbb{N}^{n \times n}$ is given by

$$M(A) = \frac{1}{2m} \sum_{i,j} \big(A_{i,j} - \frac{k_i k_j}{2m}\big) \delta(c_i, c_j)\,,$$

where c_i is the cluster to which node i belongs, m denotes the number of edges in G and c_j is the cluster to which node j belongs; k_i and k_j denote i and j's degrees respectively; $\delta(c_i, c_j)$ is the *Kronecker delta* symbol that equals 1 iff $c_i = c_j$, and 0 otherwise. For *directed networks* the modularity becomes

$$M(A) = \frac{1}{m} \sum_{i,j} \big(A_{i,j} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{m}\big) \delta(c_i, c_j)\,,$$

where k_i^{in} and k_j^{out} are i and j's in- and out- degree respectively [20].

The (Pearson) *correlation coefficient* r is used for measuring linear dependence between two random variables X and Y. We apply the sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}\,.$$

were $\bar{X}$ and $\bar{Y}$ denote the sample mean of X and Y respectively [1].

2.2 Evidence Networks

Social networks and social resource sharing systems like BibSonomy usually capture links between users explicitly, e. g., in a *friend-network* or a *follower-network*. However, besides these explicit relations, there are a number of other *implicit* evidences of user relationships in typical social resource sharing systems. These are given by, e. g., clicklogs or page visit information. In some systems, it is also possible to copy content from other users. Then, the logging information can be transformed into a corresponding user-graph structure which we call *evidence network*, following [25].

In the following sections, we define typical explicit and implicit networks in the context of social bookmarking applications. All of these are implemented in the social resource sharing system BibSonomy, but are also found in other resource sharing and social applications. Even more implicit user interaction occurs in the context of ubiquitous web applications. Examples are users which are using a given service at the same place and time, or communication relationships based on proximity sensors [32], among many others. During our evaluation period we did not have access to such sensor data, but these interactions lead to implicit user relationships which naturally fit into the framework of evidence networks described in this section.

Explicit Relation Networks In the context of the BibSonomy system, we distinguish the following explicit networks: The follower-graph, the friend-graph, and the group graph that are all established using explicit links between users. Formally, these graphs can be defined as follows:

- The *Follower-Graph* $G_1 = (V_1, E_1)$ is a directed graph with $(u, v) \in E_1$ iff user u follows the posts of user v, i. e., user u monitors the posts and is able to keep track of new posts of user v.
- The *Friend-Graph* $G_2 = (V_2, E_2)$ is a directed graph with $(u, v) \in E_2$ iff user u has added user v as a friend. In the BibSonomy system, the only purpose of the friend graph so far is to restrict access to selected posts so that only users classified as "friends" can observe them.
- The *Group-Graph* $G_3 = (V_3, E_3)$ is an undirected graph with $\{u, v\} \in E_3$ iff user u and v share a common group, e. g., defined by a special interest group.

Implicit Relation Networks Concerning implicit relationships, we propose the following networks: The click-graph, the copy graph, and the visit graph that are built by analyzing the actions of users, i. e., clicking on links, copying resources, and visiting pages of other users, respectively. Formally, the graphs are defined as follows:

- The *Click-Graph* $G_4 = (V_4, E_4)$ is a directed graph with $(u, v) \in E_4$ iff user u has clicked on a link on the user page of user v.
- The *Copy-Graph* $G_5 = (V_5, E_5)$ is a directed graph with $(u, v) \in E_5$ iff user u has copied a resource, i. e., an publication reference from user v.
- The *Visit-Graph* $G_6 = (V_6, E_6)$ is a directed graph with $(u, v) \in E_6$ iff user u has navigated to the user page of user v.

Each implicit graph G_i, $i = 4, \ldots, 6$ is given a weighting function $c_i \colon E_i \to \mathbb{N}$ that counts the number of corresponding events (e. g., $c_5(u, v)$ counts the number of posts which user u has copied from v).

2.3 Evaluation Paradigm

Several approaches exist for assessing the quality of a given set of communities. Considering users as points in appropriate feature spaces, objective functions based on the resulting distribution of data points can be applied (e. g., overlaps of the user's tag clouds, [16]). Modeling inter-user relations in terms of graphs, various graph indices defined for measuring the quality of graph clusterings can be applied (see, e. g., [15] for a survey). These indices capture the intuition of internally densely connected clusters with sparse connections between the different clusters. Furthermore, the modularity measure (see Section 2.1) is based on the observation, that communities within social networks are internally more densely connected than one would expect in a corresponding null model, i. e., in a random graph.

Accordingly, most methods for community detection try to optimize the produced community division with respect to a given quality measure. However, care must be taken, since different measures might exhibit certain biases, i. e., they tend to reward communities with certain properties which might lead to respectively skewed community structures [22]. Given the diversity of user interests, no single quality measure can potentially reflect all reasons for two users being contained within the same or different communities (or even both). Ultimately, a user study can quantify, how well a given community structure coincides with the actual reception of the users.

Dealing with the related task of *user recommendations*, Siersdorfer [31] proposed an evaluation paradigm, which is based on the *reconstruction of existing social structures*. Applied to the community detection setting in the context of a social bookmarking system as BibSonomy, this paradigm suggests to measure the quality of a given division of the users by assessing the corresponding community structure in an existing social structure. For our evaluation paradigm we therefore transform this principle to evaluating community structures using evidence networks (see Section 2.2): Our input is given by an arbitrary community clustering of a given set of users – independent of any community detection method. This clustering is then assessed using the implicit evidence networks. We show in the evaluation setting that this procedure is consistent with applying explicit networks that contain explicit user links but are rather sparse compared to the evidence networks.

Concerning our application setting, BibSonomy incorporates three relations among users, all of which potentially can serve as a basis for such an evaluation, namely the *Friend-Graph*, the *Follower-Graph* and the *Group-Graph*. Before such a network can be utilized as a reference for quality assessments, it has to be thoroughly analyzed, since different structural properties may influence the resulting assessment, cf. [25]. But more importantly, one has to cope with the sparsity of the explicit user relations: The Friend-Graph of BibSonomy, for example, only spans around 1000 edges among 700 users of all 5600 considered users and all possible 30 million edges. Thus, feature spaces for users, for example, using tags or resources as describing elements potentially capture a richer set of relations than those modeled in the graphs. In the following, we therefore consider the much more dense *implicit evidence networks* as discussed in [25], which can be typically observed in a running resource sharing system. In our analysis, we investigate whether they are consistent with the existing explicit networks in BibSonomy as a reference for evaluating community detection methods.

3 Related Work

Despite the absence of well-established gold-standards, the growing need for automated user community assessment is reflected in a considerable number of proposed paradigms. Evaluation approaches of generated links between users can broadly be divided in content-based and structure-based methods (relying on given links between users). In the following, we discuss related work concerning community mining in general, community detection methods, evaluation measures, metrics and evaluation paradigms.

Fortunato [14] discusses various aspects connected to the concept of community structure in graphs. Basic definitions as well as existing and new methods for community detection are presented. This work is a good entry point for the topic of community mining. In [19], Lancichinetti presents a thorough comparison of many different state of the art community detection algorithms in graph. The performance of algorithms are compared relative to a class of adequately generated artificial benchmark graphs.

Karamolegkos et al. [16] introduced metrics for assessing user relatedness and community structure by means of the normalized size of user profile overlaps. They evaluate their metrics in a live setting, focussing on the optimization of the given metrics.

An LDA [7] based community detection method for folksonomies is presented in [17] which is evaluated indirectly by measuring the improvement of search results achieved by incorporating the mined community information. Using a metric which is purely based on the structure of graphs, Newman presents algorithms for finding communities and assessing community structure in graphs [28]. A thorough empirical analysis of the impact of different community mining algorithms and their corresponding objective function on the resulting community structures is presented in [22], which is based on the size resolved analysis of community structure in graphs [21].

Recently Siersdorfer et al. [31] proposed an evaluation technique for recommendation tasks in folksonomies which is based on the reconstruction of existing links (e. g., friendship lists). The performance of a given system is assessed by applying quality measures which are derived from established measures used in information retrieval. Schifanella et al. [30] investigated the relationship of topological closeness (in terms of the length of shortest paths) with respect to the semantic similarity between the users. Crandall et al. [12] discuss similarity and social influence in online communities, providing the general idea that friends interact similarly concerning activities of users. Their results indicate, that there are clear feedback effects between similarity between actors and future interactions.

Another aspect of our work is the analysis of implicit link structures which can be obtained in a running Web 2.0 system and how they relate to other existing link structures. Baeza-Yates et al. [4] propose to present query-logs as an implicit folksonomy where queries can be seen as tags associated to documents clicked by people making those queries. Based on this representation, the authors extracted semantic relations between queries from a query-click bipartite graph where nodes are queries and an edge between nodes exists when at least one equal URL has been clicked after submitting the query. Krause et al. [18] analyzed term-co-occurrence-networks in the logfiles of internet search systems. They showed that the exposed structure is similar to a folksonomy.

Analyzing Web 2.0 data by applying complex network theory goes back to the analysis of (samples from) the web graph [8]. Mislove et al. [24] applied methods from social network analysis as well as complex network theory and analyzed large scale crawls from prominent social networking sites. Some properties common to all considered social networks are worked out and contrasted to properties of the web graph. Newman analyzed many real life networks, summing up characteristics of social networks [29].

Concerning the approaches mentioned above, this work unifies topics of community mining, community evaluation, and social structures: We provide an approach for relative community assessment using the link structure of different (implicit) networks capturing user interactions. These so-called evidence networks are thoroughly analyzed with respect to the contained community structure. It is shown, that there is a strong common community structure across different evidence networks using standard community evaluation functions. The results suggest a basis for a new evaluation framework of community detection methods in social applications.

4 Evaluation

In the following, we first describe the data used for the evaluation of the evidence networks. After that, we describe the characteristics of the applied evidence networks, and discuss relations between the networks. After that, we present the conducted experiments. We conclude with a detailed discussion of the experimental results.

4.1 Evaluation Data and Setting

Our primary resource is an anonymized dump of all public bookmark and publication posts until January 27, 2010, from which we extracted *explicit* and *implicit* relations. It consists of 175,521 tags, 5,579 users, 467,291 resources and 2,120,322 tag assignments. The dump also contains friendship relations modeled in BibSonomy concerning 700 users. Additionally, it contains the *follower* relation, which is explicitly established between user u and v, if u is interested in v's posts and wants to stay informed about new posts, as discussed above. Furthermore, we utilized the "*click log*" of BibSonomy, consisting of entries which are generated whenever a logged-in user clicked on a link in BibSonomy. A log entry contains the URL of the currently visited page together with the corresponding link target, the date and the user name[4]. For our experiments we considered all click log entries until January 25, 2010. Starting in October 9, 2008, this dataset consists of 1,788,867 click events. We finally considered all available apache web server log files, ranging from October 14, 2007 to January 25, 2010. The file consists of around 16 GB compressed log entries. We used all log entries available, ignoring the different time periods, as this is a typical scenario for real-world applications.

Table 1. High level statistics for all relations where U denotes the set of all users in BibSonomy.

	Copy	Visit	Click	Follower	Friend	Group
$\|V_i\|$	1427	3381	1151	183	700	550
$\|E_i\|$	4144	8214	1718	171	1012	6693
$\|V_i\|/\|U\|$	0.25	0.58	0.20	0.03	0.12	0.10
#scc	1108	2599	963	175	515	90
largest scc	309	717	150	5	17	228
#wcc	37	11	55	37	140	89
largest wcc	1339	3359	1022	83	283	228

4.2 Characteristics of the Networks

In the following, we briefly summarize the link symmetry characteristics and degree distribution of the extracted networks and discuss its power-law distribution. The analysis is restricted to the large (weakly) connected components of the network.

[4] Note: For privacy reasons a user may deactivate this feature.

Link symmetry: Mislove et al. [24] showed for Flickr, LiveJournal and YouTube that 60-80% of the direct friendship links between users are symmetric. Among others, one reason for this is that refusing a friendship request is considered impolite. However, the friendship relation of BibSonomy differs significantly. Only 43% of the friendship links between users are reciprocal.

When more features are available exclusively along friendship links (e. g., sending posts), the friendship graph's structure will probably change and links will get more and more reciprocal. But concerning the implicit networks we will see, that link asymmetry is determined by a structure common to all our implicit networks.

Degree distribution: One of the most crucial network properties is the probability distribution ruling the likelihood $p(k)$, that a node v has in- or out-degree k respectively. In most real life networks, the so called *degree distribution* follows a power law [11], that is $p(k) \sim k^{-\alpha}$ where $\alpha > 1$ is the exponent of the distribution. Online social networks [24], collaborative tagging systems [9], scientific collaboration networks [2] among others are shown to expose power law distributions.

For comparability, we calculated a best fitting power law model for each distribution using a maximum likelihood estimator [11] and noted the corresponding Kolmogorov-Smirnov goodness-of-fit metrics in Table 2 for reference. All in- and out-degree distributions except those from the groups graph show a power law like behavior, though there are significant deviations.

Table 2. Power law parameters

	Copy	Visit	Click	Follower	Friend	Group
α_{in}	2.48	2.9	2.86	2.48	3.47	3.5
α_{out}	1.75	2.2	2.7	2.78	2.24	3.5
D_{in}	0.0603	0.0227	0.023	0.0278	0.0617	0.1503
D_{out}	0.0571	0.0364	0.0394	0.0919	0.0939	0.1503

4.3 Relations between Networks

Another basic evaluation considered associational properties between the networks, i. e., whether links present in one or more network are associated with links in another network. For the evaluation, we considered all possible links between users (user pairs) and assessed whether a link in a network existed, or not. Table 3 shows the associations for the friend graph. The associations are given in the form of subgroup rules (e. g., [3]), or association rules with the single element *friend*, denoting link membership in the friend graph, in the rule head. The parameters given in the table denote the support of the rule (i. e., the relative number of covered links), its confidence (or precision), the recall, and the F1-Measure, as the harmonic mean of precision and recall, e. g., [3]. Additionally, the table includes the *lift* of the rule denoting the relative increase in confidence/precision considering the default rule (with an empty precondition), i. e., is obtained by dividing the rule's confidence/precision by the default precision.

The first (baseline) rule therefore shows the default associations for the friend graph, i. e., the default confidence/precision for link membership is 5.5%. The second rule can be read as follows: IF users are connected in the *click* AND *copy* AND *visit* graphs, THEN the probability of being connected in the *friend* graph is 23.9% (confidence/precision), with a support of 1.7%, and an F1-Measure of 11.2%.

Table 3. Associations between evidence networks and the friend graph. The table shows the support (SUP), confidence/precision (CONF), the recall (REC), the F1-Measure (F1M), and the Lift (LIFT) for the different associations, measured in percentages.

#	RULE	SUP	CONF	REC	F1M	LIFT
1	friend ←	100	5.5	100	10.4	1.00
2	friend ← click copy visit	1.7	23.9	7.3	11.2	4.35
3	friend ← click copy	2	20.7	7.4	10.9	3.76
4	friend ← copy visit	4.6	17.8	14.9	16.2	3.24
5	friend ← click visit	7.7	13.5	19	15.8	2.45
6	friend ← click	9.5	11.3	19.7	14.4	2.05
7	friend ← visit	45.1	6.9	56.8	12.3	1.25
8	friend ← copy	24.4	3.8	16.8	6.2	0.69

While the table shows significant increases of the rule confidences compared to the baseline measured by the lift parameter, the individual confidence values are rather low. This can be explained by the different sizes of the networks (the friend graph is the smallest network contained in the table). Additionally, the table shows interesting results comparing the individual networks: While the single *copy* network is only a limited predictor for friend-relationships (line 8) the combination with the *click* and/or *visit* networks significantly increases the associations quality, compared to considering the isolated networks (ll. 2, 3, 4, 6 and 7).

4.4 Applied Clustering Method

Starting our experiments we faced a vicious circle: For assessing the quality of a community structure, we need a preferably good method for obtaining such a structure in the beginning. However, since we do not want to examine a particular clustering algorithm and prove its performance, we use a rather simple approach which is on the one hand easy to understand, on the other hand, it can be broadly parameterized and allows the construction of a randomized variety of initial clusterings.

First experiments were conducted using the well known k*-means algorithm* [23]. For that, each user u is represented by a vector $(u_1, \ldots, u_T) \in \mathbb{R}^T$ where T is the total number of tags and u_i is the total number of times user u assigned the tag i to resources in BibSonomy ($i = 1, \ldots, T$). The resulting clusters had poor quality, assigning most users to a single cluster. Due to the sparsity of the considered high dimensional vector space representation (there are more than $170,000$ tags), the underlying search for nearest neighbors fails (cf., e. g., [6] for a discussion).

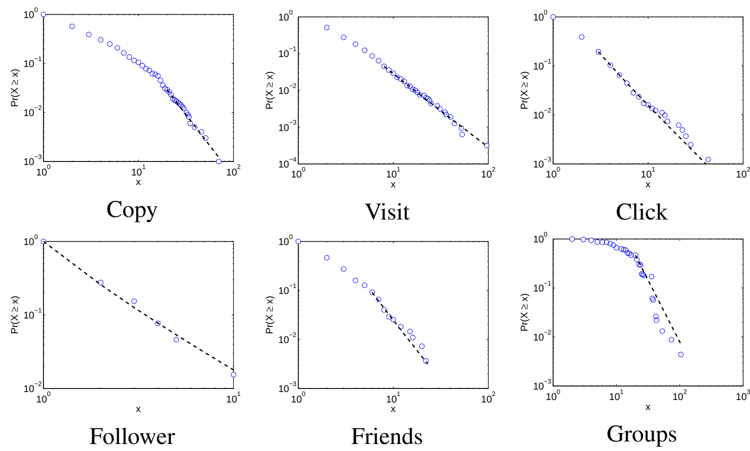


Fig. 1. In-degree distribution of the different evidence networks

To bypass this problem, we reduced the number of dimensions. There are a variety of approaches for dimensionality reduction. We chose to cluster the tags for building "*topics*", consisting of associated sets of tags. A user u is thus represented as a vector $\boldsymbol{u} \in \mathbb{R}^{T'}$ in the topic vector space, where $T' \ll T$ is the number of topics.

For our experiments, we used a *latent dirichlet allocation* [7] method for building topics, which efficiently build interpretable tag clusters and has been successfully applied in similar contexts to tagging systems (cf. [31]). In the following, our models are denoted with "*LDA-n-kMeans-k*", where n denotes the number of topics and k the number of clusters. In total, we obtained 40 different basic clusterings.

4.5 Experiments and Results

Our experiments aim at examining whether the *implicit evidence networks* described in Section 2.2 are admissible complements for the sparse *explicit networks*. This would justify using, e. g., the Visit-Graph and thus allow to assess more than 53% of the active users (in contrast to only 12% covered by the Friend-Graph) applying the evaluation paradigm "*reconstruction of existing social structures*" described in Section 2.3.

The most fundamental property of a sound measure is the relative discrimination of "better" and "worse" community structures, allowing algorithms to approximate optimal structures stepwise by applying local heuristics. For analyzing how quality assessment by applying the different evidence networks is sensitive to small disturbances, we conducted a series of randomized experiments.

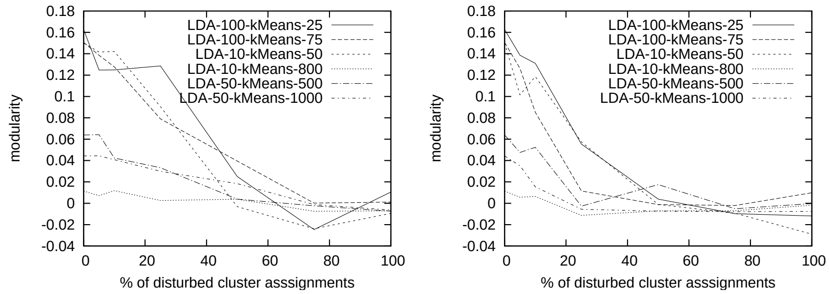
We started with community structures constructed by the basic feature clustering described above, using 10, 50, 100, and 500 topics, and constructing clusterings ranging from 10 to 1,000 clusters in total. Any clustering or community detection method could be used here (e. g., we also conducted the same series of experiments applying a graph clustering algorithm). We focussed on the applied method as it is easy to understand and can be broadly parameterized; it allows for a simple generation of a variety of (randomized) initial clusterings. We gradually added noise to these initial structures and at each step assessed the resulting community structure by calculating the quality measures described in Section 2.1 for the different evidence networks: Two different approaches for adding noise to a given division into communities were applied. The first approach (from now on called “Random” for short) randomly chooses a node u belonging to some community c_u. This node is than assigned to another randomly chosen community $c' \neq c_u$. Note that this kind of disturbance leads to a different distribution of cluster sizes. The second approach (from now on called “Shuffle”) randomly swaps the community allocation of randomly chosen nodes belonging to different communities, which leads to community structures with the same community size distribution.

Figures 2 and 3 show the corresponding results of calculating the modularity for each evidence network at every level of disturbance in the underlying community structure (higher modularity values indicate stronger community structure). Similarly, Figures 4 and 5 show the results of calculating the conductance. For the ease of presentation, we selected from all considered clusterings a subset which represents a broad range of assessed community qualities. We emphasize that this experiment does not aim at selecting a “best” community structure, rather than examining the relative rating of slightly worse structures when applying the different evidence networks (based on the assumption, that randomly disturbing communities decreases their quality).

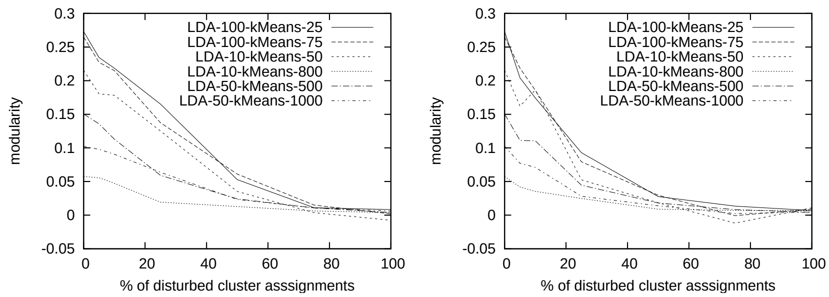
We see that the modularity on every evidence network is consistent with the level of disturbance, that is, the modularity value monotonically decreases with increasing percentage of disturbed nodes. Slight deviations (e. g., looking at the alternating gradients of the Follower-Graph) are most likely statistical effects due to the limited size of the corresponding evidence network. These results are supported by the figures showing the corresponding plots for the conductance values, since lower conductance values indicate stronger clustering.

Note that conductance and modularity give precedence to different community structures. In particular, structures with many small communities are preferred according to their conductance ($k = 500, 800, 1000$), whereas smaller numbers of clusters are preferred according to their modularity (Figure 6 exemplary shows two corresponding cluster size distributions). This behavior is consistent with the corresponding bias of the applied measures as discussed in [22].

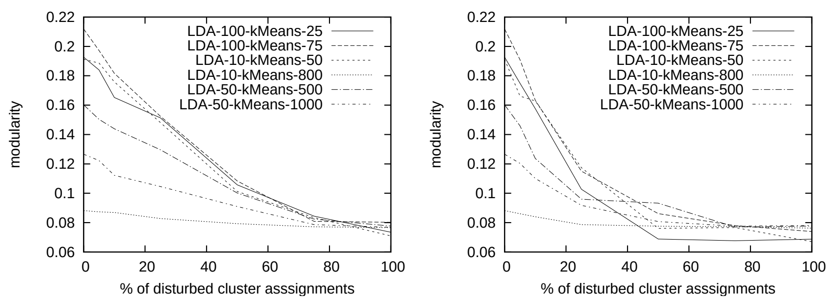
The preceding results consider the different evidence networks independently. However, we ultimately want to use the implicit networks as supplement for the sparse explicit social structures (in particular the Friend-Graph). We therefore expect the assessment of community structures applying the implicit networks to be consistent with the application of the explicit networks. This motivates the following experiment: We calculated the Pearson correlation coefficient for each of the implicit networks and one of the explicit networks. Following the paradigm of reconstructing existing social structures, the *explicit* networks yield sensible community scores (in terms of modularity and conductance). The following experimental setup aims at examining whether the corresponding community scores as induced by *implicit* networks are consistent (i. e., correlated) with the scores induced by the *explicit* networks. Table 4 shows the corresponding correlation coefficients (see Section 2.1) for modularity and conductance scores in the Friend-Graph and each of the graphs in Figures 2-5 (averaged per measure



(a) Follower-Graph (randomized left, shuffled right).



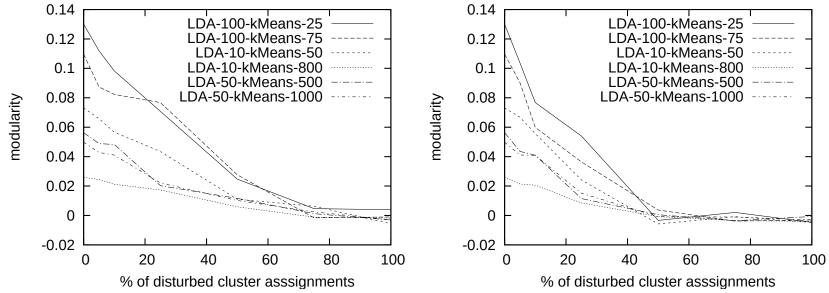
(b) Friend-Graph (randomized left, shuffled right).



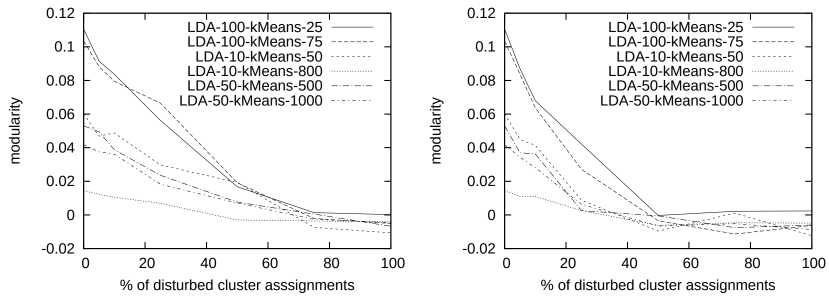
(c) Group-Graph (randomized left, shuffled right).

Fig. 2. Modularity calculated on different clusterings at varying levels of disturbed cluster assignments relative to explicit evidence networks

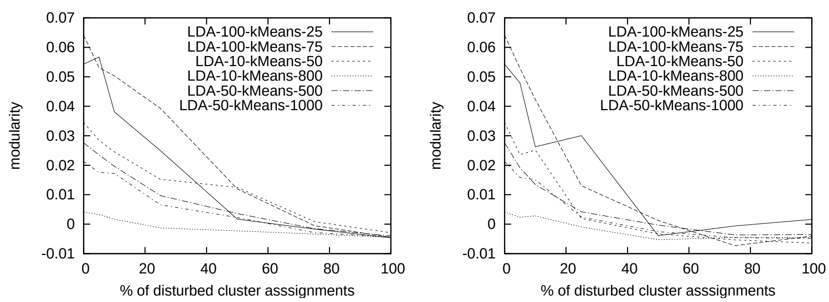
and randomization type). The averaged correlation coefficients suggest a surprisingly high correlation between the measures calculated on the implicit networks and those calculated on the friend graph. Especially the conductance graphs show high correlation coefficients with low standard deviations. In comparison, repeating the same experiment with the group graph as the most dense existing social structure shows lower correlation coefficients with higher standard deviation, cf. Table 5.



(a) Copy-Graph (randomized left, shuffled right).



(b) Click-Graph (randomized left, shuffled right).

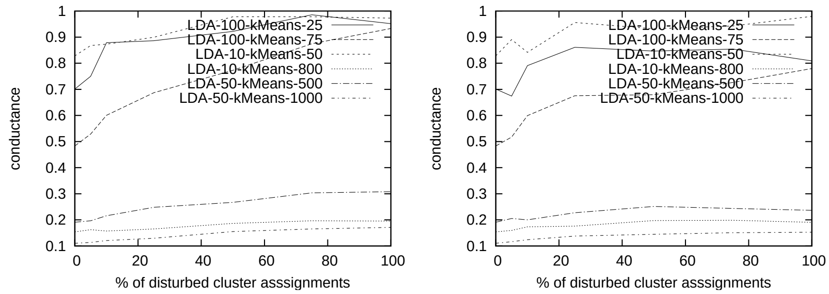


(c) Visit-Graph (randomized left, shuffled right).

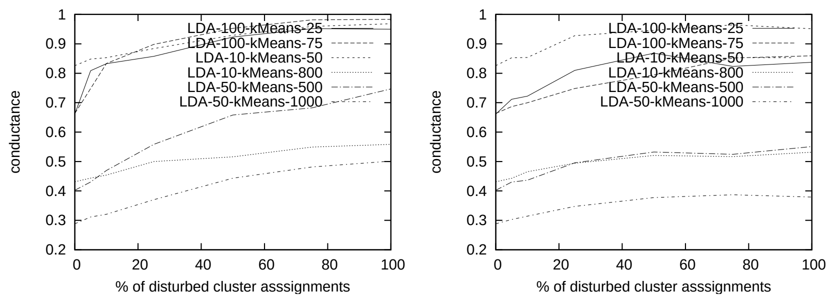
Fig. 3. Modularity calculated on different clusterings at varying levels of disturbed cluster assignments relative to implicit evidence networks

4.6 Discussion

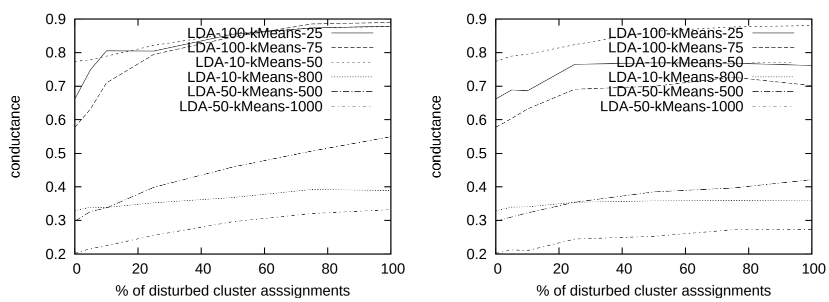
The experimental results presented in the previous section indicate that implicit evidence networks used for assessing the quality of a community structure are surprisingly consistent with the expected behavior as formalized by the existing explicit social structures, in particular concerning the Friend-Graph. In our experiments (considering 40 models per experiment) we observed a high correlation between the quality measures calculated on the implicit and explicit networks supporting this hypothesis.



(a) Follower-Graph (randomized left, shuffled right).



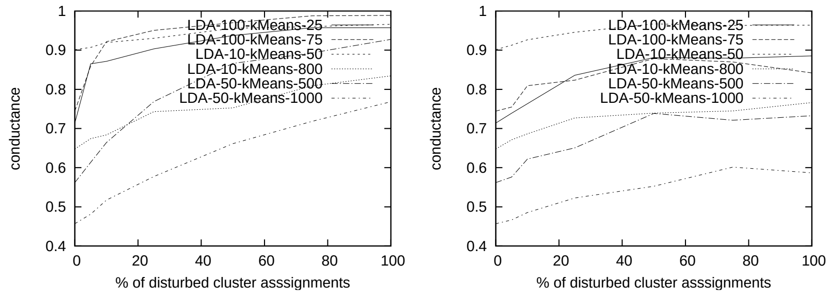
(b) Friend-Graph (randomized left, shuffled right).



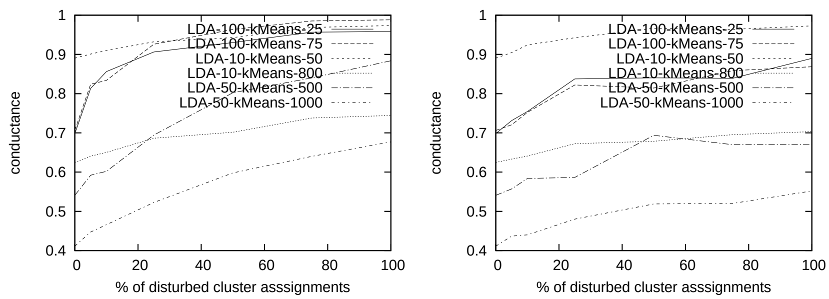
(c) Group-Graph (randomized left, shuffled right).

Fig. 4. Conductance calculated on different clusterings at varying levels of disturbed cluster assignments relative to explicit evidence networks

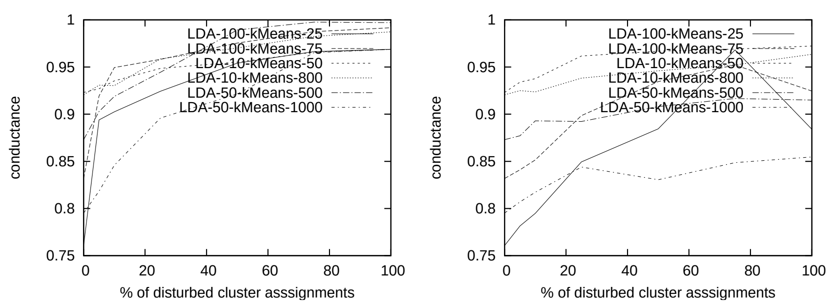
The implicit networks show a lower correlation with the group graph. At the first glance, this looks like a disappointing result. But the analysis of the group graph shows, that its properties significantly differ from typical social networks as discussed in [25, 24]. Most strikingly, its degree distribution follows not a power law and its distribution of strongly connected components differs. Therefore, we obtain a ranking of the explicit graphs: It is thus more desirable to model the friend graph's behavior more closely than the group graph's.



(a) Copy-Graph (randomized left, shuffled right).



(b) Click-Graph (randomized left, shuffled right).



(c) Visit-Graph (randomized left, shuffled right).

Fig. 5. Conductance calculated on different clusterings at varying levels of disturbed cluster assignments relative to implicit evidence networks

5 Conclusions

In this paper, we have presented evidence networks for the evaluation of communities. Since explicit graph data is often sparse and does not cover the whole instance space well, evidence networks provide a viable alternative and complement to explicit networks, if available. We have discussed several possible evidence networks, and their features. The presented evaluation paradigm is based on the idea of reconstructing existing social structures for the assessment and evaluation of a given clustering. Thus, the

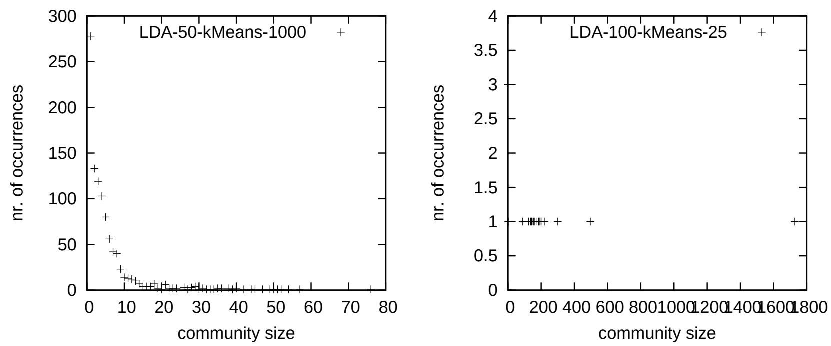


Fig. 6. Two opposed community size distributions as preferred by conductance (left) and modularity (right).

Table 4. Averaged Pearson correlation coefficient ρ_{G_i,G_2} together with it's empirical standard deviation for each of the experiments "Shuffle" (S) and "Randomize" together with the considered objective functions modularity (M) and conductance (C) on the different implicit evidence networks G_i and the friend graph G_2.

Evidence Network	R/M	S/M	R/C	S/C
Follower-Graph	0.86 ± 0.17	0.90 ± 0.12	0.89 ± 0.28	0.83 ± 0.41
Group-Graph	0.91 ± 0.13	0.95 ± 0.08	1.00 ± 0.01	0.96 ± 0.17
Copy-Graph	0.82 ± 0.17	0.87 ± 0.12	0.99 ± 0.03	0.98 ± 0.09
Click-Graph	0.80 ± 0.17	0.86 ± 0.13	0.99 ± 0.04	0.98 ± 0.07
Visit-Graph	0.72 ± 0.25	0.80 ± 0.18	0.97 ± 0.06	0.98 ± 0.08

Table 5. Averaged Pearson correlation coefficient ρ_{G_i,G_3} together with it's empirical standard deviation for each of the experiments "Shuffle" (S) and "Randomize" together with the considered objective functions modularity (M) and conductance (C) on the different implicit evidence networks G_i and the group graph G_3.

Evidence Network	R/M	S/M	R/C	S/C
Friend-Graph	0.91 ± 0.13	0.95 ± 0.08	1.00 ± 0.01	0.96 ± 0.17
Follower-Graph	0.72 ± 0.30	0.83 ± 0.20	0.89 ± 0.27	0.82 ± 0.40
Copy-Graph	0.67 ± 0.35	0.80 ± 0.23	0.98 ± 0.05	0.93 ± 0.29
Click-Graph	0.68 ± 0.35	0.80 ± 0.23	0.98 ± 0.04	0.94 ± 0.29
Visit-Graph	0.60 ± 0.42	0.73 ± 0.28	0.96 ± 0.07	0.93 ± 0.27

contribution of this paper considers a new kind of data for assessing user communities in social applications is introduced and formalized as so-called evidence networks: These are thoroughly analyzed with respect to the contained community structure (cf. Section 4). It is shown that there is a strong common community structure across different networks. Our conducted experiments furthermore suggest that the different networks are not contradictory but complementary.

The context of the presented analysis is given by social applications such as social networking, social bookmarking, and social resource sharing systems, considering our own system BibSonomy[5][5] as an example. The evaluation of the presented approach using real-world data from the social resource sharing tool BibSonomy indicated the soundness of the approach considering the consistency of community structures and the applied measures. But the presented analysis is not only relevant for the evaluation of community mining techniques, but also for implementing new community detection or user recommendation algorithms, among others.

For future work, we aim to investigate, how the different evidence networks can be suitably combined into a single network. For this, we need to further analyze the individual structure of the networks, and the possible interactions. Another interesting options for further research is the improvement of the clustering algorithms on the user – tag data. An improved preprocessing of the tagging data seems a promising direction for further improving the general approach. Furthermore, we plan to extend our experiments for a larger count of networks and clusterings in order to generalize the obtained results to a broader bases.

We also plan to compare the proposed ratings of community allocations with results of different user studies, partly integrated in a live setting in the running system BibSonomy. As another direction of research, we are considering to incorporate evidence networks in the community detection process (e. g., in terms of constraints).

References

1. Agresti, A.: An Introduction to Categorical Data Analysis. Wiley-Blackwell (2007)
2. Almendral, J.A., Oliveira, J., López, L., Mendes, J., Sanjuán, M.A.: The Network of Scientific Collaborations within the European Framework Programme. Physica A: Stat. Mech. and its Applic. 384(2), 675 – 683 (2007)
3. 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 (2009)
4. Baeza-Yates, R., Tiberi, A.: Extracting Semantic Relations from Query Logs. In: Proc. 13th ACM SIGKDD Conference. p. 85. ACM (2007)
5. Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The Social Bookmark and Publication Management System Bibsonomy. The VLDB Journal 19(6), 849–875 (2010)
6. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When Is ”Nearest Neighbor” Meaningful? In: Beeri, C., Buneman, P. (eds.) ICDT. Lecture Notes in Computer Science, vol. 1540, pp. 217–235. Springer (1999)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
8. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph Structure in the Web. Computer Networks 33(1-6), 309–320 (2000)
9. Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V., Loreto, V., Hotho, A., Grahl, M., Stumme, G.: Network Properties of Folksonomies. AI Communications 20(4), 245–262 (2007)
10. Chung, F.R.K.: Spectral Graph Theory. CBMS Reg. Conf. Series in Mathematics 92 (1997)
11. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-Law Distributions in Empirical Data. SIAM Review 51(4) (2009)

[5] http://www.bibsonomy.org/

12. Crandall, D.J., Cosley, D., Huttenlocher, D.P., Kleinberg, J.M., Suri, S.: Feedback Effects between Similarity and Social Influence in Online Communities. In: Proc. 14th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. pp. 160–168. ACM (2008)
13. Diestel, R.: Graph Theory. Springer, Berlin (2006)
14. Fortunato, S., Castellano, C.: Community Structure in Graphs (2007), arxiv:0712.2716 Chapter of Springer's Encyclopedia of Complexity and System Science
15. Gaertler, M.: Clustering. In: Brandes, U., Erlebach, T. (eds.) Network Analysis. LNCS, vol. 3418, pp. 178–215. Springer (2004)
16. Karamolegkos, P.N., Patrikakis, C.Z., Doulamis, N.D., Vlacheas, P.T., Nikolakopoulos, I.G.: An Evaluation Study of Clustering Algorithms in the Scope of User Communities Assessment. Computers & Mathematics with Applications 58(8), 1498–1519 (2009)
17. Kashoob, S., Caverlee, J., Kamath, K.: Community-Based Ranking of the Social Web. In: Proceedings of the 21st ACM Conference on Hypertext and hypermedia (2010)
18. Krause, B., Jäschke, R., Hotho, A., Stumme, G.: Logsonomy - Social Information Retrieval with Logdata. In: Proc. 19th Conf. on Hypertext and Hypermedia. pp. 157–166. ACM (2008)
19. Lancichinetti, A., Fortunato, S.: Community Detection Algorithms: A Comparative Analysis (2009), arxiv:0908.1062
20. Leicht, E.A., Newman, M.E.J.: Community Structure in Directed Networks. Phys. Rev. Lett. 100(11) (Mar 2008)
21. 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 (2008), arxiv:0810.1355
22. Leskovec, J., Lang, K.J., Mahoney, M.W.: Empirical Comparison of Algorithms for Network Community Detection (2010), cite arxiv:1004.3539
23. MacQueen, J.B.: Some Methods for Classification and Analysis of MultiVariate Observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press (1967)
24. Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. In: 7th ACM SIGCOMM Conf. on Internet Measurement. p. 42. ACM (2007)
25. Mitzlaff, F., Benz, D., Stumme, G., Hotho, A.: Visit me, Click me, be my Friend: An Analysis of Evidence Networks of User Relationships in BibSonomy. In: HT '10: Proc. 21st ACM Conference on Hypertext and Hypermedia. pp. 265–270. ACM, New York, NY, USA (2010)
26. 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)
27. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
28. Newman, M.E.J.: Detecting Community Structure in Networks. Europ Physical J 38 (2004)
29. Newman, M., Park, J.: Why Social Networks are different from Other Types of Networks. Physical Review E 68(3), 36122 (2003)
30. Schifanella, R., Barrat, A., Cattuto, C., Markines, B., Menczer, F.: Folks in Folksonomies: Social Link Prediction from Shared Metadata. In: Proc. 3rd ACM Int'l Conf. on Web search and data mining. pp. 271–280. ACM, New York, NY, USA (2010)
31. Siersdorfer, S., Sizov, S.: Social Recommender Systems for Web 2.0 Folksonomies. In: HT09: Proc. 20th ACM Conf. on Hypertext and Hypermedia. pp. 261–270. ACM, New York, NY, USA (2009)
32. Szomszor, M., Cattuto, C., Van den Broeck, W., Barrat, A., Alani, H.: Semantics, Sensors, and the Social Web: The Live Social Semantics Experiments. The Semantic Web: Research and Applications pp. 196–210 (2010)