Discovery of Complex User Communities

Chapter · October 2015

DOI: 10.1007/978-3-319-23835-7_1

Symeon Papadopoulos
The Centre for Research and Technology, Hellas

Dimitrios Vogiatzis
National Center for Scientific Research Demokritos

Discovery of Complex User Communities

Georgios Paliouras[1], Symeon Papadopoulos[2], and Dimitrios Vogiatzis[1]

[1] NCSR "Demokritos"
[2] Centre for Research and Technology Hellas (CERTH)

Abstract. This chapter serves as an introduction to the book on *User Community Discovery*, setting the scene for the presentation in the rest of the book of various methods for the discovery of user communities in the social Web. In this context, the current chapter introduces the various types of user community, as they appeared in the early days of the Web, and how they converge to the common concept of *active user community* in the social Web. In this manner, the chapter aims to clarify the use of terminology in the various research areas that study user communities. Additionally, the main approaches to discovering user communities are briefly introduced and a number of new challenges for community discovery in the social Web are highlighted. In particular we emphasize the complexity of the networks that are constructed among users and other entities in the social Web. Social networks are typically *multi-modal*, i.e. containing different types of entity, *multi-relational*, i.e. comprising different relation types, and *dynamic*, i.e. changing over time. The complexity of the networks calls for new versatile and efficient methods for community discovery. Details about such methods are provided in the rest of the book.

1 Introduction

Nowadays, much of our social activity takes place online. Technological advances have led to the emergence of the *social Web* (also known as *Web 2.0*), which made Web users active participants, generating their own content and forming Online Social Networks (OSNs). Friendship and interaction among users lead naturally to the formation of communities either by explicit connections or by connections that can be inferred through the similarity of the users or the online trails of their interactions (implicit communities). The composition of communities and the underlying characteristics that their members share is valuable knowledge to companies that provide products and services online. As an example, taking into account the communities in which social network users participate can considerably improve the relevance of the recommended content and ultimately the engagement of users, i.e. a form of *social recommendation* [102].

The focus of this book is on the discovery of implicit communities of users[3], i.e. beyond the explicit friendship or other types of connection that can be found

[3] Although the term "community discovery" is more suitable to describe this process, throughout the text we adopt the term "community detection", which is the prevalent term used in the literature to refer to this problem.

in OSNs. This discovery process is data-driven, based on statistical and machine learning approaches, has made its appearance in user modeling research more than 20 years ago [74,20], and has been recently revived in the context of statistical physics [27] and graph mining research [94]. Such data mining methods have supported new powerful ways of personalization, such as collaborative filtering and recommendation [98,40,49,50]. However, the advent of the social Web has introduced a number of new challenges and opportunities for community discovery methods. Social media, generated primarily by the users themselves and exchanged over OSNs, make user data richer, larger in size and highly multidimensional [79]. In fact, user data is becoming truly social data, like the data one would collect by observing the interaction of the members of a society. Discovery methods need to be adapted to this new type of data, in order to be able to identify the complex multi-scale community structures that are formed. At the same time, the ease with which such rich social data can be collected increases the responsibility of researchers and service providers to respect the privacy of individuals and user groups.

This chapter serves as an introduction to the book, setting the scene for the presentation of various methods for discovering user communities in the rest of the book. Based on the categorization introduced in [75], the chapter starts (section 2) by presenting the different types of user community that have been studied in the literature and how these converge into the common concept of *active user community* in the social Web. Then, in section 3, it presents the basic characteristics of communities, as subgraphs of a larger graph of connected entities, e.g., a network of users. Section 4 presents briefly the main approaches to community discovery, while section 5 explains the challenges and opportunities for community discovery in the social Web. A more thorough treatment of community discovery methods in the social Web is provided in the rest of the book chapters. Section 6 highlights some example applications of community discovery in the social Web. Finally, section 7 summarizes the main concepts introduced in this chapter and discusses some of the main open issues.

2 Types of Web community

According to [75] the term "Web community" has been used with three different meanings in computer science literature:

User communities correspond to clusters of users of a Web site, who share common interests. Typical such communities are the customers of online shops. User communities are formed on the basis of user log data, as recorded on the Web servers of the site. These data are analyzed by statistical and machine learning methods [76] in order to identify groups of users with common interests or behavior, e.g., users who buy similar products.

Web communities correspond to subgraphs of the Web graph, i.e. clusters of Web pages, that are densely connected. Such dense subgraphs are identified by graph mining methods [26,52] and indicate Web sites that provide similar or related content, e.g., art museums in different countries.

Web-based communities are associated with systems that support the formation or strengthening of real-life communities. These were typically either local communities (e.g., within a University campus) or interest-driven ones (e.g., professional associations). Early work in this field focused on *community networks* [95] and *virtual communities* [90], which then evolved into *Web community portals* [100] and finally into *Web-based communities* [10]. What makes this type of community particularly interesting is the fact that users start producing content for the community. In other words, they move from being passive consumers of content to becoming active community participants. This idea, aided by corresponding technological advances, later developed in what we now call the social Web.

The social Web has been facilitated by technological advances in the interaction of the users with Web resources (a.k.a. Web 2.0 technologies) and has facilitated, in turn, two important socio-technical developments, often referred to as social media and Online Social Networks (OSNs). Social media and user-generated content represent the widespread participation of Web users in the generation and publication of digital content, which has now become more dynamic than ever before. It is this active participation of the users in content publishing that has led to the use of the term *active user* in the social Web. On the other hand, OSNs are typically Web applications that support the active networking of users, much in the spirit of Web-based communities. OSNs can be considered the natural descendants of *Web-based communities* and community networks. As such, they bear similarities to those earlier paradigms, for instance the goal of linking people with common interests or needs. However, OSNs also have significant differences from their predecessors [11], among which are:

- Their much larger user base.
- The diversity of their user base and their detachment from particular themes or geographic locations.
- The fact that people link to each other, but do not necessarily join predefined groups. A social network is a graph, in the sense of the Web itself, rather than a group of people.
- Their participatory nature turns passive consumers into active users, who provide content and information of many new and interesting types.

Using the terminology of [75], we call communities of users in OSNs *active user communities*. These are naturally related to Web-based communities, but they are also related to the other two types of community mentioned above.

Due to the fact that OSNs are naturally mapped to graphs, active user communities are subgraphs of the larger graph, similar to *Web communities*. In contrast to the Web though, the nodes of the social graph are typically users and thus its subgraphs form *user communities*. In this manner, the three different types of community seem to converge to the common and much richer structure of the active user community in the social Web.

Beyond the fact that they bring together the three different notions of community, active user communities introduce a number of novel and interesting possibilities. In the social Web, users provide much richer information about their preferences and needs, than what the logs of a Web server could reveal. They choose their neighbors in the network, they publish their own content, they rate and tag content that other people have provided and participate in a number of online activities. Due to the variety of entities involved in the social Web, one may choose to create graphs other than those connecting users, e.g., by relating content items posted by the users, or even multi-partite graphs that relate different types of entity (see chapter "Community Discovery in Multi-Mode Networks" of this book). Additionally, these graphs may contain different edge types, e.g., edges that represent friendship and others that represent communication among the users. This multi-relational nature of the social Web (see chapter "Discovering Communities in Multi-relational Networks" of this book) makes the task of active community discovery particularly interesting and challenging.

3 Representation of Communities

3.1 Communities as graphs

The common representation of a community is that of a graph $C = (V, E)$, where V and E represent the nodes and edges respectively, that connect closely-related users or other entities. Usually, this graph is a subgraph of a larger one $C \subseteq G$, e.g., all the users of a social network, and is assumed to have a dense structure, representing the close relation among its members. However, there is considerable discussion about the best choice for representing such a dense structure.

A large body of literature assumes that G should be partitioned into subsets (the communities), whereby each node of the graph belongs to exactly one subgraph. However, in reality, especially in the context of the social Web, community membership is more complicated. A user may well belong to more than one community and may belong to different communities in different degrees (Figure 1(a)). The need to allow overlaps between communities was realized even in early research in community discovery (e.g., [74]). Being the most dense graph structure, cliques (i.e. fully connected subgraphs: $\forall v \in V(C) : degree_C(v) = |V| - 1$) were among the first options to be considered. Thus, in the early days of community detection research, community graphs G were required to be maximal cliques of C, that are not subsumed by other cliques of C. The strictness of this definition often leads to unwanted effects, such as very low participation of nodes to communities.

In order to overcome this problem, alternative, less strict definitions were used, such as cores or k-shells (introduced in [97]). A k-shell is a subgraph C of a graph $G = (V, E)$ iff $\forall v \in V(C) : degree_G(v) \geq k$ and C is the maximum subraph with this property. An equivalent definition is that k-shell is a maximal subset of nodes such that each is reachable from each of the others by at least k node independent paths. Two paths are defined as node independent if they share none of the same nodes, with the exception of the start and the end nodes [69, Sect.7.8.2]. Among their other benefits, k-shells can be computed efficiently in large graphs, such as social networks. This concept of core has been expanded in [92,9], and core-periphery structures have been studied in the context of weighted networks [29], hypergraphs [87], as well as in temporal networks [63].

Additionally, a number of other graph structures have been used in the literature to address the issue of overlapping communities, some based on cliques, e.g. (Palla et al. [77]), but many others as well (Clauset [18], Luo et al. [58], Gregory [37], Chen et al. [17]). An extensive treatment of the problem of overlapping community detection is presented in [108].

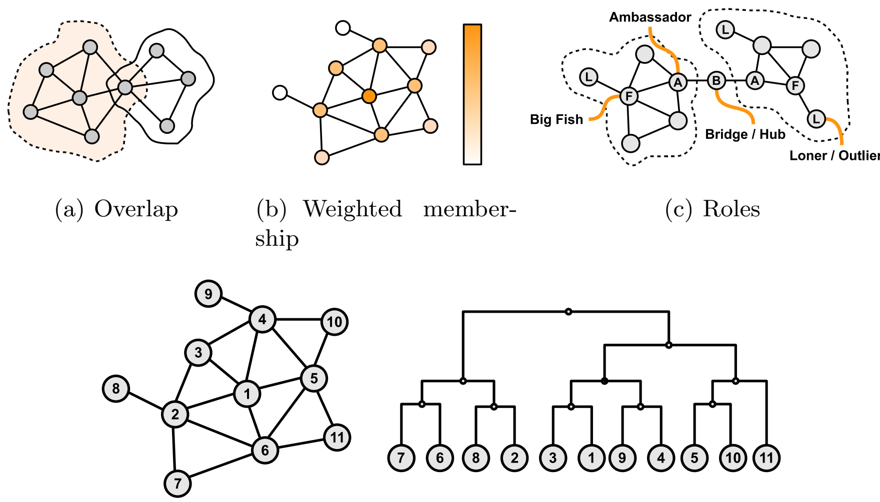
3.2 Community attributes

The assumption that communities either partition the graph or are allowed to overlap is an important differentiating attribute of community representation and discovery approaches. In addition, there are other attributes that nodes of a network may have in relation to communities. For instance, different nodes may participate with varying degrees in a community depending on their centrality[4] within it (Figure 1(b)). Moreover, nodes may have discrete roles: for example, Xu et al. [109] define two roles (*hubs* and *outliers*) for nodes that are not assigned to any community. Hubs are connected to multiple communities and act as liaisons, thus enabling interactions among communities. Outliers are connected to a single community through a single link, therefore they are usually considered as noise. Community-based node roles are also discussed by Scripps et al. [96]. Specifically, the roles of “loners”, “big fish”, “bridges” and “ambassadors” are defined (Figure 1(c)).

It is also possible to impose hierarchical (Figure 1(d)) or multi-scale structure on communities [1]. Community organization may be defined at different scales with respect to real-world systems. For instance, a set of users of a social Web application may be organised in a community focused on a very specific topic (e.g., fans of a particular indie-rock band) and at the same time they may be considered as members of a broader community (rock music). In many cases, however, such hierarchies are not flexible enough to model the complexities of multi-level organization in social Web systems, as for instance in the case of folksonomies [15].

An orthogonal dimension concerning the characterization of communities is time. This is a significant aspect of community detection that is worth further attention especially due to the volatile and highly dynamic nature of social Web data and interactions. According to a recent survey on the topic [35], a three-layered stream of graph snapshots can be used to capture the evolution of social interactions. Graph snapshots are used at the lowest level to capture the state at specific points in time. At a higher level, these are grouped into segments, 3D tensor structures, encompassing short-term evolutions. At the top level the complete graph stream captures the history of interactions among the graph nodes. An alternative representation of graph evolution relies on an initial (base) graph

[4] Centrality quantifies how often nodes belong to the paths connecting other nodes.



(a) Overlap (b) Weighted membership (c) Roles

(d) Hierarchy

Fig. 1. Several attributes that may characterise community structure: (a) overlap, (b) weighted membership, (c) node (vertex) roles within/across communities, (d) hierarchical organization. Image originally presented in [79].

that corresponds to the original state, and a stream of changes (e.g., node additions/removals, edge additions/removals). Accordingly, time-awareness can be incorporated in the underlying community structure of time-evolving graphs, either by considering a series of community structures defined at the corresponding graph snapshots and a set of pairwise community structure associations across snapshots, or by considering an initial (base) community structure and a stream of changes on this structure, or by inherently integrating time-awareness into the community detection process [32]. A set of basic changes (or transformations) that community structures may undergo are described in the seminal work by Palla et al. [77]: Essentially, there are three types of transformation: (a) one-to-one, which involves community growth or contraction, (b) one-to-many, which involves one community splitting to many or many communities merging to one, and (c) one-to-zero (and zero-to-one), which involves the emergence or the extinction of a community.

4 Community Detection on Simple Graphs

Depending on the underlying methodological principles, five broad classes of community detection and graph clustering methods were defined in [79] : (a) cohesive subgraph discovery, (b) vertex (or node) clustering, (c) community quality optimization, (d) divisive, and (e) model-based. Note that these classes are not mutually exclusive. For instance, spectral community detection methods [72] can be considered both to perform quality optimization and be divisive.

Cohesive subgraph discovery. The methods of this class presume a specification of the structural properties that a subgraph of the network should satisfy to be considered a community. Once such a subgraph structure is chosen, methods involve the enumeration of subgraphs in the network under study. Local community definitions, such as cliques, n-cliques, k-cores, LS sets and lambda sets, are examples of such cohesive structures and therefore algorithmic schemes for enumerating such structures, such as the Bron-Kerbosch algorithm [13] and the efficient k-core decomposition algorithm of Batagelj and Zaversnik [4], belong to this class of community detection methods. In addition, methods such as the Clique Percolation Method (Palla et al. [78]) and the SCAN algorithm (Xu et al. [109]), which lead to the discovery of subgraph structures with well-specified properties, fall under the same class of methods.

Vertex (node) clustering. Such techniques originate from traditional data clustering research. A typical means of casting a graph clustering problem to one that can be solved by conventional data clustering methods (such as k-means and hierarchical agglomerative clustering) is by embedding graph nodes in a vector space, where pairwise distances between nodes can be calculated. Another popular method is to use the spectrum of the graph for mapping graph nodes to points in a low-dimensional space, where the cluster structure is more profound (Donetti and Munoz [23], Von Luxburg [59]). Other node similarity measures such as structural equivalence (Breiger et al. [12]) and neighborhood overlap have been used to compute similarities between graph nodes (Wasserman and Faust [107]). Finally, a noteworthy method, called Walktrap (Pons and Latapy [81]), makes use of a random-walk based similarity between nodes and between communities and uses modularity in a hierarchical agglomerative clustering scheme to derive an optimal node clustering structure.

Community quality optimization. There is a large number of methods that are founded on the basis of optimizing some graph-based measure of community quality. Subgraph density and cut-based measures, such as normalized cut (Shi and Malik [99]) and conductance (Kannan et al. [47]), were among the first to be used for quantifying the quality of some network division into clusters. A whole new wave of research was stimulated by the measure of modularity. Approximate modularity maximization schemes abound in the literature. Apart from the seminal greedy optimization technique of Newman [70], and speeded up versions of it, such as max-heap based agglomeration (Clauset et al. [19]) and iterative heuristic schemes (Blondel et al. [7]), more sophisticated optimization methods have been devised, for instance, extremal optimization (Duch et al. [24]), speeded simulated annealing (Massen and Doye [61]) and spectral optimization (Newman [71]). Methods aiming at the optimization of local measures of community quality, such as local and subgraph modularity (Clauset [18], Luo et al. [58]), also belong to this category. Finally, this category includes methods that exploit the “hills” and “valleys” in the distribution of network-based node or edge functions, e.g., the ModuLand framework proposed by Kovács et al. [51] and the "reachability" measure by Chen et al. [17], and the highly popular OSLOM method [53], which performs local optimization of a fitness function expressing the statistical significance of clusters.

Divisive. These methods rely on the identification of network elements (edges and nodes) that are positioned between communities. For instance, the seminal algorithm by Girvan and Newman [36] progressively removes the edges of a network, based on some edge betweeness measure until communities emerge as disconnected components of the graph. Several measures of edge betweeness have been devised, for instance, edge, random-walk, and current-flow betweeness (Newman and Girvan [73]), as well as information centrality (Fortunato et al. [28]) and the edge clustering coefficient (Radicchi et al. [85]). A similar principle is adopted by node removal methods (Vragović and Louis [105]); such methods remove nodes in order to reveal the underlying community structure. Finally, min-cut/max-flow methods (Flake et al. [26], Ino et al. [44]) adopt a different divisive perspective: they try to identify graph cuts (i.e. sets of edges that separate the graph into pieces) of minimum size.

Model-based. This is a broad and more recent category of methods that either consider a dynamic process taking place on the network, which reveals its communities, or they consider an underlying model of statistical nature that can generate the division of the network into communities. Examples of dynamic processes are label propagation (Raghavan et al. [86], Leung et al. [54], Gregory [37]), synchronization of Kuramoto oscillators (Arenas et al. [3]), diffusion flow, better known as Markov Cluster Algorithm (Van Dongen [104]), and the popular spin model by Reichardt and Bornholdt [89]. In addition, community detection can be cast as a modelling problem, such as the well-known stochastic block model [41] and its extensions [2], or a statistical inference problem (Hastings [39]), assuming some underlying probabilistic model, such as the planted partition model, that generates the community structure and estimating the parameters of this model. Other model-based approaches rely on the principle that a good clustering is determined by a low encoding cost and thus they perform community detection by finding the cluster structure that results in the lowest possible cluster encoding cost (Chakrabarti [16], Rosvall and Bergstrom [93]).

Several of the aforementioned and other methods and are discussed in detail in the survey articles by Danon et al. [21], Fortunato [27], Porter et al. [82], and Schaeffer [94]. Also, a useful listing of a large number of community detection methods appears in the supplementary material of Kovács et al. [51]. The majority of the aforementioned methods have been designed for use with undirected graphs. A thorough treatment of the community detection problem in the context of directed networks is presented in [60].

It is worth stressing that methods such as the ones listed above, have been proposed and applied in the context of User and Web communities (as defined in section 2, typically giving rise to graphs with a single type of node and edge. In recent years, however, there has been a tendency for increasing complexity in the structure of real-world networks with the emergence of active user communities, as discussed above. This naturally leads to more sophisticated data models, for instance networks with attributed nodes [111], networks with content-associated edges [83], as well as multi-mode and multi-relational networks. Consequently, the active user communities call for new representations and new detection methods. For instance, modern OSNs, such as Twitter and Facebook, can be modelled by means of different entities, such as users, resources, and tags, and relations, such as likes, comments, affiliations, etc. In the next section, we examine how this additional complexity has brought new developments to the research field of community detection.

5 Communities in the Social Web

As previously discussed, OSNs are complex systems comprising multiple entity types associated with multiple relations. Thus, there may be users connected to other users, but also users creating posts, or replying to, sharing, commenting or rating posts. Moreover, the posts could contain references to external resources such as Web pages, be associated with a theme or a geographic location and carry a timestamp.

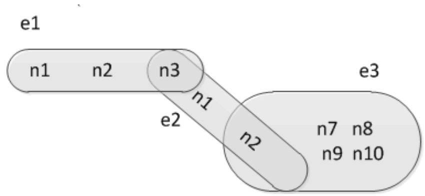
A complete representation of an OSN would require a graph to represent users, posts, and other entities in the form of nodes, whereas replies, comments, participation in groups would be represented as relations. Moreover, the relations between entities could be binary, ternary or of higher order. In the context of this chapter we use the term *multi-dimensional graph* to refer to graphs that contain edges that comprise more than two nodes, or to graphs that contain multiple types of edge. An overview of such graphs can be found in [48]. We divide multi-dimensional graphs into two broad categories, *hypergraphs* and *multi-relational graphs.*

Hypergraphs. Multi-dimensional graphs that contain a single type of edge will be referred to as hypergraphs, formally defined as $G = (V, \mathbf{E})$, where V is the set of nodes and $\mathbf{E} \subseteq V \times V \ldots \times V$, is the set of hyperedges. A further distinction between hypergraphs could be made between single partite and multi-partite hypergraphs. In the latter case there are nodes of more than one type, i.e. $\mathbf{E} \subseteq V_1 \times V_2 \ldots \times V_k$, given k types of node (see also Figure 2). Despite their recent popularity in social network analysis, hypergraphs are not a novel concept, as they were already proposed by Berge in 1973 [5].

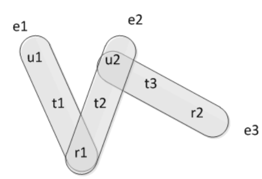
In hypergraphs, the definitions of graph properties such as node degree, path, cycle, clustering coefficient and clique have been extended to accommodate the new properties of the hyperedges. For instance, hypergraph path definitions can be found in [106], centrality in [8] and cliques in [22].

Tri-partite graphs in particular, have been studied in the context of social tagging systems. They arose from the need to represent triadic relations. For instance, a seller, a buyer and a broker participating in a business transaction; or a person seeing a movie, and annotating it with tags. Thus, one partite could stand for users, another for tags and a third for movies. Hyperedges could be interpreted as tag assignments by a user to a resource. Moreover, it is often assumed that there are no connections between nodes of the same partite and each edge contains one node from each partite. In this case, i.e. if each hyperedge has a node from each partite, the result is a k-partite uniform hypergraph. Three commonly used datasets for the study of hypergraphs are excerpts from Delicious, MovieLens and LastFM[5].

The usage of a uniform k-partite network could be extended to represent other social networks that were not originally conceived to be such; for instance in Twitter, we may consider three partites: users, named entities (such as people, places and organizations) that are referred to in the tweet text, and references to external resources such as URLs.



(a) Hyper graph



(b) Multi-partite graph

Fig. 2. Hypergraphs and hyperedges: (a) general hypergraph, (b) tripartite graph; e1, e2, e3 denote hyperedges. The nodes in (a) are of single type, whereas the nodes in (b) are of three types: u, t, r.

Multi-relational graphs. Multi-relational graphs comprise more that one relations between their nodes. These can be represented as $G = (V, \mathbf{E})$, where V is the set of nodes and $\mathbf{E} = E_1, E_2, \ldots E_l$ is a set of sets of edges, l is the number of relations, and $E_i \subseteq V \times V$. Each edge type carries certain semantics, for instance relations could denote colleagues, friends, etc. Multi-relational graphs appeared in the field of artificial intelligence as semantic networks [84], in the field of machine translation as a representational language [91] and ultimately they originate from predicate logic. Multi-relational networks are a natural way to represent the various forms of relation in OSN, but also to represent information across OSNs that could be used to unify relevant information about users. For instance, in Figure 3, various users are present in both the Twitter and the LinkedIn OSNs. In the aforementioned definition of multi-relational networks we implied that the relations are binary, but this definition can be generalized to cover multi-relational hyper-graphs.

5.1 Definitions of communities in the social Web

Next, we provide definitions of the community types for the social Web that are relevant in the current presentation. We will be making references to the

[5] http://grouplens.org/datasets/hetrec-2011/

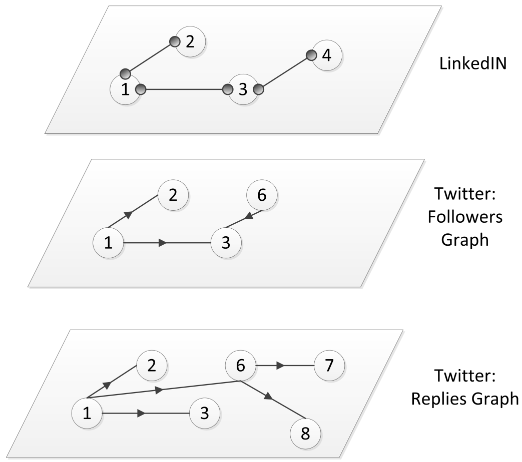


Fig. 3. Multi-relational graph representing three types of relation over two networks. Contacts in LinkedIn; and Followers and Replies in Twitter. The node labels denote the user, for instance user 1 participates in all relations, across all networks.

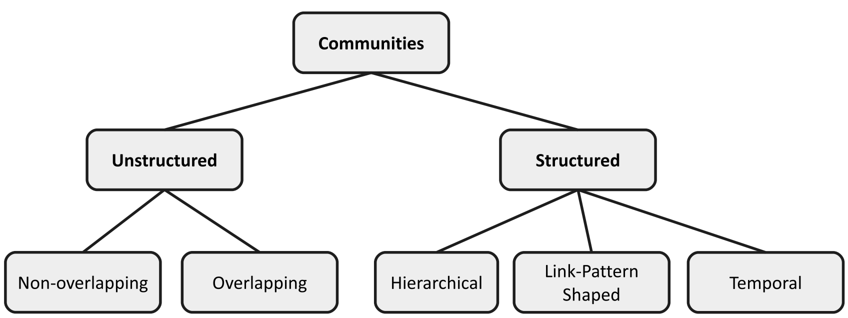


Fig. 4. A taxonomy of communities

diagram of Figure 4. Let $G(V, E)$ denote a social network of $|V|$ nodes and $|E|$ edges. A set of non-overlapping communities is represented as $\{C_1, \ldots C_k\}$, where $C_i = \{v_{i(1)}, v_{i(2)} \ldots\}$, $v_{i(k)} \in V$, and $\forall\ i, j\ C_i \bigcap C_j = \emptyset$. In the case of overlapping communities $\exists\ i, j,\ C_i \bigcap C_j \neq \emptyset$.

Communities can also be structured, as in the case of hierarchical communities, where $\exists\ i, j,\ C_i \subset C_j$. Additionally, In multi-partite networks we can discover a complex community structure that relates communities of a single type of node. Relations between communities are shaped by the link (or edge) pattern. That is, $\forall i\ P_i \subset V$, where P_i are the nodes from a single partite, and $C_{P_{i,j}}$ denotes a community j from partite i. The relation of communities from different partites can be expressed as: $\{C_{P_{i1,j(1)}}, C_{P_{i1,j(2)}}, C_{P_{i2,j(1)}}, C_{P_{i2,j(2)}}, C_{P_{i3,j(3)}}, \ldots\}$, where the first index denotes the partite, and j is an index to the community. For instance in Figure 5, the complex community structure can be expressed as: $\{C_{P_{1,1}}, C_{P_{2,1}}, C_{P_{3,1}}\}$, and $\{C_{P_{2,1}}, C_{P_{2,2}}, C_{P_{3,1}}\}$.

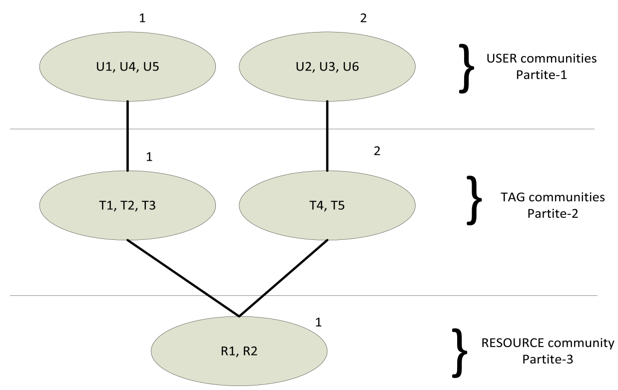


Fig. 5. Complex structure in multi-partite communities. Each user community is related to one tag community, and all are related to a single resource community. The numbers on the ovals correspond to community identifiers.

Finally, in the case of temporal communities there is an ordering for each community as well as a labeling of the evolutionary phenomena that the communities undergo from time frame to time frame, i.e. $C_{t_1(i)} \prec C_{t_2(i)} \ldots$, and $e_k = f(C_{t_1(i)}, C_{t_2(i)})$, where e_k is the evolutionary phenomenon that occurred to $C_{t_1(i)}$ and transformed it into $C_{t_2(i)}$. Typically the evolutionary phenomenon can be growth, shrinking, continuation, dissolution, etc. An extensive discussion of temporal communities is included in [42] and [43].

5.2 Community detection on social graphs

We briefly review some community detection methods in hypergraphs and multi-relational graphs, aiming to provide an overview of some widely established methods. A detailed analysis of hypergraph and multi-relational methods is exposed in Chapters "Community Discovery in Multi-Mode Networks" and "Discovering Communities in Multi-relational Networks" of the current volume respectively. Moreover, there is also a relevant review in [48].

Community detection has been applied in the context of hypergraphs broadly following two main approaches: a) mapping the hypergraph to a simpler structure to discover communities with any algorithm for simple graphs (such as the ones discussed in section 4), and b) discovering communities directly on the hypergraph. Following the first approach, multi-partite graphs can be mapped to graphs, where the nodes are typically users. In this case the connections between the nodes are meant to represent the node similarity or proximity in the original hypergraph. However, in the case of bi-partite graphs it has been shown that the discovered communities in the simple graph are not always a faithful representation of the original communities under some structural measures [38].

Following the second approach, established methods of community detection in single-partite graphs have been extended to multi-partite graphs. For instance, modularity maximization, which is a widely used method in single-partite graphs has been formulated for bi-partite graphs [65]. Then it was suggested that tri-partite uniform graphs can be projected onto bi-partite graphs and the modularity of the tri-partite graph can be computed as the aggregated modularity of bi-partite graphs [68]. However, this method results in a proliferation of bi-partite structures, which makes it challenging to scale beyond more than three partites. Also, modularity maximization was subsequently formulated for tri-partite graphs [66,67]. Moreover, spectral clustering, which can be used to embed a simple graph into the Euclidean space and then to perform clustering in that space, has been extended to multi-partite graphs [114].

Another major category of algorithms attempt to cluster links instead of nodes to detect communities; this concept has also been extended to tri-partite graphs [34]. In particular, the first step is to detect pairwise similarities between hyperedges, which is done by considering the neighbourhood of each node that is incident to a hyperedge. The result is a pairwise matrix of hyperedge similarities, on which any community detection method on simple graphs can be applied. Eventually, communities of hyperedges are obtained, and consequently the node communities can be overlapping.

In multi-relational networks there are several ways to deal with the different edge types when performing community detection. First, one could simply ignore the semantics of the edges by integrating all different edges connecting two nodes in a single one. This approach is problematic, for instance in the case that some edges denote friendship, whereas others denote animosity (as in the case of signed networks). Thus it is important to have a specific perspective on the network because this will substantiate the relations and allow their integration. A general approach is to consider each relation and the relevant nodes as a separate network, and then to proceed with some sort of integration [103]. Another approach is to focus is on the discovery of relations that bind a given set of users. This results in a network with a weight matrix that is a linear combination of the various relations [14], The resulting network is single-relational and hence communities can be discovered with standard methods.

The methods mentioned above essentially try to reduce a multi-relational network into a single-relational one. There are however approaches that follow a different path: they try to detect groups of consistent interactions (i.e. relations) between the entities of an OSN, and these are considered to form a community. There are two basic approaches in this line of research. The first is based on tensor factorization. The multiple relations between the entities of an OSN can be represented as the modes of a tensor. For instance, relations among users together with relations between users and visited items form a three-mode tensor. Tensor decomposition can be used for extracting latent features that are later used to build communities [57]. The second approach in community detection directly in multi-relational networks is to extend the widely used statistical measure of modularity maximization [45].

Temporal networks can also be considered as multi-relational by considering discrete time steps. Given also the fact that time is an ordinal attribute, there can be an ordering of time relationships. Thus, an email exchange network, or a co-authorship network can be subject to an analysis that discovers communities per time frame and then associates them across time frames to study their evolution [101]. This problem has also been addressed with matrix factorization [31]. Issues such as the longevity of email communities, or the forms of their evolution are important for their characterization.

6 Applications of Community Detection

An important application of community detection is in the domain of user-contributed multimedia mining. The majority of social Web applications offer facilities for users to upload and annotate multimedia content. Typical and popular examples of social Web multimedia sharing applications are Flickr and Instagram for images, SoundCloud for audio and YouTube, Vimeo and DailyMotion for video. A common application associated with multimedia sharing on the social Web is the mining of multimedia communities, i.e. communities comprising media items. Having such structure at one's disposal, one can derive user communities by translating item communities to user communities through the respective authorship/ownership (e.g., users belong to the same communities that their respective items belong to). Alternatively, to derive multimedia communities, one may take into account the rich social context (in the form of interactions or affiliations) that is associated with the respective users. Given the huge increase in the amounts of user-contributed content in social Web multimedia sharing applications, being able to perform clustering on them can help their users navigate larger parts of the content more efficiently (i.e. by looking at one representative item per cluster instead of all cluster members).

A common analysis approach for mining communities of user-generated multimedia is to first construct a similarity graph that captures the pairwise similarities between media items and then to apply a community detection approach with the goal of extracting clusters of similar media items. Moëllic et al. [64] pursue photo clustering by use of a shared nearest neighbors approach on two graphs of photos where edges between photos are computed either by use of shared tags (tag-based graph) or based on visual similarity (visual graph). The employed clustering technique is shown to achieve improved clustering performance compared with conventional clustering algorithms (k-means and one of its speeded-up variants). Also, comparing the results of their methods with the clusters available from Flickr (Groups explicitly defined by users of the application), the authors noted similar clustering quality.

A more sophisticated application of graph clustering is presented by Li et al. [55]. Their goal is to collect different representative (*iconic*) photos for popular landmarks and use the massive visual content that is associated with them in order to create 3D landmark models. They devised a multi-stage photo processing framework, in which an important task was to group iconic photos together in order to reduce the amount of photos that are processed by the computationally intensive 3D reconstruction step. They achieve photo grouping by creating a photo graph where photos are connected by edges when they are visually similar and by applying the N-cut graph clustering algorithm by Shi and Malik [99] on this graph. In that way, they managed to reconstruct the major views of three famous landmarks (Statue of Liberty, Notre Dame and San Marco).

Papadopoulos et al. [80] identified real-world landmarks and events in large tagged photo collections by use of photo cluster classification. They applied the SCAN algorithm (Xu et al. [109]) on a hybrid photo similarity graph that encodes both visual and tag similarity between photos. Subsequently, the derived photo and tag communities are classified as landmarks or events based on cluster features such as the cluster duration and number of unique users with photos in the cluster. In their analysis, manual inspection of the results reveals that most of the clusters correspond either to famous landmarks of the city or to real events (e.g., music concerts). Furthermore, the automatically selected cluster tags provide meaningful descriptions for them.

Gargi et al. [30] used community detection to perform clustering of YouTube videos. As a first step, they formed a video similarity graph using co-watch statistics (i.e., an edge between two videos is inserted if many users watch the two videos in the same session). Then, they applied a multi-step community detection approach, consisting of a local seed community detection step based on the concept of density and conductance, and a cluster refinement step, where the text similarity between videos is taken into account to ensure topical coherence between the videos. The proposed approach was designed with scalability in mind, due to the very large size of the underlying video similarity graph, and was shown to lead to meaningful and coherent clusters.

The association of content together with structure in social Web multimedia sharing applications has recently motivated the development of community detection approaches that take into account such content [83]. Edge content provides unique insights into communities because it characterizes the nature of the interactions between participants more effectively. This is because the use of purely structural information cannot easily characterize the nature of the interactions between participants effectively. Similarly, the information which is available only at the nodes may not be able to easily distinguish the different interactions of nodes that belong to multiple communities. Correspondingly, the use of edge content enables richer insights which can be used for more effective community detection.

Content recommendation is another important application of community detection in OSNs. As mentioned before, users contribute content and they may also comment, tag or vote on the content produced by them or by others. This forms the context, and indeed the context can influence users' experience. Digg[6] is a social network that allows news sharing among users. In particular users submit news stories related to a topic, and they may vote, comment or reply to stories submitted by other users. Apart from users, there are other entities such as stories, and comments. Relations among the entities can be binary such as (user, story) or ternary such as (user, story, comment), and they vary with time. A method based on online tensor factorization has been used to detect communities that comprise users voting or commenting with respect to news items [56], thus capturing the social news context. Based on this, the next step is to predict future votes of a user on stories as well as to recommend stories.

Beyond content clustering and recommendation, another application of community detection that has been increasingly attracting interest in view of the ubiquitous use of OSNs is the automatic organization of users' online connections into meaningful groups, also referred to as "social circles", e.g. family, colleagues, etc. This is particularly valuable as a user empowerment mechanism, since it enables OSN users to share different content with different groups of their connections (e.g., personal content with close friends, professional content with colleagues, etc.). Jones and O'Neill were among the first to apply community detection on this problem [46] and found out that using the SCAN algorithm [109] led to the discovery of social circles that matched well with the perceptions of users (probed with the use of a carefully designed user study). Mcauley and Leskovec defined and formulated this problem as a multi-membership node clustering problem on a user's ego network and proposed a hybrid clustering approach relying both on the network structure of the ego network (using community detection) and on the attribute similarity between the user and their connections [62].

Finally, discovering the communities in which users participate across multiple networks may be used for user profiling. For instance, discovering a user's communities in LinkedIn could reveal his/her interests, the opinion of other users on him/her, and thus it will provide useful information for a relatively recent profile or activity of the same user in Twitter. This process may require the matching of user profile information across different networks, where a variety of methods may be used [112].

[6] http://digg.com/

7 Conclusions and Open Issues

This chapter laid out the basis and context for this book, which is the discovery of communities in the social Web. The emphasis was on the challenges one faces in adapting community discovery methods to the complexity of the social graph. The sources of this complexity are the multitude of entities involved in the social Web, the variety of relationships that are formed among them and the sheer volume of data that need to be analyzed. The main approaches to the problem were highlighted in this chapter, but more details about the corresponding methods are provided in the rest of the book.

Most of the work on community discovery in the social Web so far involves graph mining methods that have been used in the past for discovering Web communities and Web user communities. These methods are either used on simplified versions of the social graph, e.g., the user graph of a social network, or they are extended to deal with multiple entity types or multiple relations. The latter approach leads to a number of interesting new methods, which however are commonly limited by the assumptions made by the original methods. Therefore, the need arises for novel methods that inherently tackle the complexity of communities in the social Web. Such methods require an appropriate multi-dimensional and multi-relational model of the social graph, as well as algorithms for extracting higher-order relational patterns from this graph.

The area of statistical relational learning [33] lends itself naturally to the problem of community detection in multi-relational networks. Statistical relational methods combine the expressive power of relational logic with the statistical capabilities of probabilistic graphical models. This combination facilitates probabilistic inference and statistical analysis on top of graphical representations of relational knowledge. Statistical relational learning has started being used for community discovery in the social Web (e.g., in [110], [25]) and matrix factorization approaches are often considered to belong to this area (e.g., [56], [88]). However, these methods also need considerable improvement in order to be applied to the scale of the social Web, particularly due to their computational cost. Therefore, the development of statistical relational community discovery methods is both a promising and an open area for research.

Last but not least, an increasingly important research aspect in the context of social Web communities pertains to privacy risks and issues arising from the collective nature and function of user communities. In particular, the possibility of analyzing communities of mixed public-private user profiles for conducting inferences about the attributes of the private profiles poses new research and ethical questions with respect to information sharing and data mining in the context of social networks [113]. Community membership (even when it is not explicit) and online relations and interactions can be actually considered as a latent feature that could lead to the discovery of user interests and attributes [6]. Coupled with the fact that there is an abundance of weak annotations in the social Web in the form of e.g., hashtags, membership in groups/lists, etc., one may conclude that community detection and analysis approaches can be increasingly considered as a powerful tool for mining user profiles, thus raising important considerations and risks with respect to online privacy.

References

1. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764, 2010.
2. A.A. Amini, A. Chen, P.J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 08 2013.
3. A. Arenas, A. Díaz-Guilera, and C.J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.*, 96:114102, Mar 2006.
4. V. Batagelj and M. Zaversnik. An o(m) algorithm for cores decomposition of networks. *CoRR*, cs.DS/0310049, 2003.
5. C. Berge and E. Minieka. *Graphs and hypergraphs*, volume 7. North-Holland publishing company Amsterdam, 1973.
6. P. Bhattacharya, M.B. Zafar, N. Ganguly, S. Ghosh, and K.P. Gummadi. Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 357–360, New York, NY, USA, 2014. ACM.
7. V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, page P10008, 2008.
8. P. Bonanich, A.C. Holdren, and M. Johnston. Hyper-edges and multidimensional centrality. *Social Networks*, 26:189–203, 2004.
9. S. P. Borgatti and M. G. Everett. Models of core / periphery structures. *Social Networks*, 21:375–395, 1999.
10. C. Bouras, V. Igglesis, V. Kapoulas, and T. Tsiatsos. A web-based virtual community. *IJWBC*, 1(2):127–139, 2005.
11. D.M. Boyd and N.B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2007.
12. R.L. Breiger, S.A. Boorman, and P. Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3), 1975.
13. C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, September 1973.
14. D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. In *Knowledge Discovery in Databases: PKDD 2005*, pages 445–452. Springer, 2005.
15. C. Cattuto, A. Baldassarri, V.D.P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. *Advances in Complex Systems*, 11(4):597–608, 2008.
16. D. Chakrabarti. Autopart: Parameter-free graph partitioning and outlier detection. In Jean-Franois Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *PKDD*, volume 3202 of *Lecture Notes in Computer Science*, pages 112–124. Springer, 2004.
17. J. Chen, O. Zaïane, and R. Goebel. A visual data mining approach to find overlapping communities in networks. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ASONAM '09, pages 338–343, Washington, DC, USA, 2009. IEEE Computer Society.

18. A. Clauset. Finding local community structure in networks. *Phys. Rev. E*, 72:026132, August 2005.
19. A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, pages 1–6, 2004.
20. R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the Nineth International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 558–567, Newport Beach, CA, USA, 1997.
21. L. Danon, A.D. Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(9):P09008–09008, September 2005.
22. M. Dawande, P. Keskinocak, J.M. Swaminathan, and S. Tayur. On bipartite and multipartite clique problems. *Journal of Algorithms*, 41(2):388–403, 2001.
23. L. Donetti and M. A. Munoz. Detecting network communities: a new systematic and efficient algorithm. *J. Stat. Mech.*, P10012, 2004.
24. J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.
25. F. Esposito, S. Ferilli, T. Basile, and N. Di Mauro. Social networks and statistical relational learning: a survey. *International Journal of Social Network Mining*, 1(2):185–208, 2012.
26. G. Flake, S. Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 150–160, New York, NY, USA, 2000. ACM.
27. S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.
28. S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. *Physical Review E*, 70(5):056104, 2004.
29. A. Garas, F. Schweitzer, and S. Havlin. A k-shell decomposition method for weighted networks. *New Journal of Physics*, 14(8):083030, 2012.
30. U. Gargi, W. Lu, V.S. Mirrokni, and S. Yoon. Large-scale community detection on youtube for topic discovery and exploration. In L.A. Adamic, R.A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.
31. L. Gauvin, A. Panisson, and C. Cattuto. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PloS one*, 9(1):e86028, 2014.
32. Laetitia Gauvin, André Panisson, and Ciro Cattuto. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *CoRR*, abs/1308.0723, 2013.
33. L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.
34. S. Ghosh, P. Kane, and N. Ganguly. Identifying overlapping communities in folksonomies or tripartite hypergraphs. In *Proceedings of the 20th international conference companion on World wide web*, pages 39–40. ACM, 2011.
35. M. Giatsoglou and A. Vakali. Capturing social data evolution using graph clustering. *IEEE Internet Computing*, 17(1):74–79, 2013.
36. M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
37. S. Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2010.

38. R. Guimerà, M. Sales-Pardo, and L. Amaral. Module identification in bipartite and directed networks. *Physical Review Letter*, 76(3):036102, 2007.
39. M. B. Hastings. Community detection as an inference problem. *Physical Review E*, 74:035102, Sep 2006.
40. W.C. Hill, L. Stead, M. Rosenstein, and G.W. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 194–201, Denver, Colorado, USA, 1995.
41. P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, June 1983.
42. P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.
43. P. Holme and J. Saramäki. *Temporal networks*. Springer, 2013.
44. H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. In *WWW*, pages 661–669. ACM, 2005.
45. A. Jacquin, L. Misakova, and M. Gay. A hybrid object-based classification approach for mapping urban sprawl in periurban environment. *Landscape and Urban Planning*, 84(2):152–165, 2008.
46. Simon Jones and Eamonn O'Neill. Feasibility of structural network clustering for group-based privacy control in social networks. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 9:1–9:13, New York, NY, USA, 2010. ACM.
47. R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, May 2004.
48. M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, and M.A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.
49. J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.
50. J.A. Konstan and J. Riedl. Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22((this issue)), 2012.
51. I. A. Kovács, R. Palotai, M. S. Szalay, and P. Csermely. Community landscapes: An integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PLoS ONE*, 5(9):e12528, 09 2010.
52. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493, 1999.
53. A. Lancichinetti, F. Radicchi, J.J. Ramasco, and S. Fortunato. Finding Statistically Significant Communities in Networks. *PLoS ONE*, 6(5), 2011.
54. I.X.Y. Leung, P. Hui, P. Liò, and J. Crowcroft. Towards real-time community detection in large networks. *Physical Review E*, 79:066107, Jun 2009.
55. X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 427–440, Berlin, Heidelberg, 2008. Springer-Verlag.
56. Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. Metafac: Community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 527–536. ACM, 2009.

57. Y.-R. Lin, H. Sundaram, and A. Kelliher. JAM: joint action matrix factorization for summarizing a temporal heterogeneous social network. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*, 2009.
58. F. Luo, J.Z. Wang, and E. Promislow. Exploring local community structures in large networks. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 233–239, Washington, DC, USA, 2006. IEEE Computer Society.
59. U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.
60. F.D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533:95–142, December 2013.
61. C.P. Massen and J.P.K. Doye. Identifying communities within energy landscapes. 71(046101), 2005.
62. Julian Mcauley and Jure Leskovec. Discovering social circles in ego networks. *ACM Trans. Knowl. Discov. Data*, 8(1):4:1–4:28, February 2014.
63. D. Miorandi and F. De Pellegrini. K-shell decomposition for dynamic complex networks. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 488–496. IEEE, 2010.
64. P.-A. Moëllic, J.-E. Haugeard, and G. Pitel. Image clustering based on a shared nearest neighbors approach for tagged collections. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, pages 269–278, New York, NY, USA, 2008. ACM.
65. T. Murata. Modularities for bipartite networks. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, pages 245–250. ACM, 2009.
66. T. Murata. Detecting communities from tripartite networks. In *Proceedings of the 19th international conference on World wide web*, pages 1159–1160. ACM, 2010.
67. T. Murata. Detecting communities from social tagging networks based on tripartite modularity. In *Proc. Workshop on Link Analysis in Heterogeneous Information Networks*, 2011.
68. N. Neubauer and K. Obermayer. Community detection in tagging-induced hypergraphs. In *Workshop on Information in Networks*, pages 24–25. New York University NY, USA, 2010.
69. M. Newman. *Networks: an introduction*. Oxford University Press, 2010.
70. M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, September 2003.
71. M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3), 2006.
72. M.E.J. Newman. Spectral methods for community detection and graph partitioning. *Physical Review E*, 88:042822, Oct 2013.
73. M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(026113), 2004.
74. J. Orwant. Heterogeneous learning in the doppelgänger user modeling system. *User Modeling and User-Adapted Interaction*, 4(2):107–130, 1995.
75. G. Paliouras. Discovery of web user communities and their role in personalization. *User Modeling and User-Adapted Interaction*, 22(1-2):151–175, 2012.
76. G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D. Spyropoulos. Clustering the users of large web sites into communities. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 719–726, Stanford, CA, USA, 2000.

77. G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007.
78. G. Palla, I. Dernyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.
79. S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554, 2012.
80. S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-based landmark and event detection for tagged photo collections. *IEEE MultiMedia*, 18(1):52–63, 2011.
81. P. Pons and M. Latapy. Computing communities in large networks using random walks (long version). *Computer and Information Sciences-ISCIS 2005*, pages 284–293, 2005. arXiv:arXiv:physics/0512106v1.
82. M. A. Porter, J.-P. Onnela, and P. J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 2009.
83. G.-J. Qi, C.C. Aggarwal, and T.S. Huang. Community detection with edge content in social media networks. In *IEEE 28th International Conference on Data Engineering (ICDE 2012)*, pages 534–545, 2012.
84. M. R. Quillian. Semantic memory. In M. MINSKY, editor, *Semantic Information Processing*, pages 27–70. MIT Press, Cambridge, Ma, 1968.
85. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658, 2004.
86. U.N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76:036106, Sep 2007.
87. E. Ramadan, A. Tarafdar, and A. Pothen. A hypergraph model for the yeast protein complex network. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, page 189. IEEE, 2004.
88. L. Rêgo Drumond, E. Diaz-Aviles, L. Schmidt-Thieme, and W. Nejdl. Optimizing multi-relational factorization models for multiple target relations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 191–200, 2014.
89. J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74:016110, 2006.
90. H. Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier*. New York: Addison-Wesley, 1993.
91. R. H. Richens. General program for mechanical translation between any two languages via an algebraic interlingua. *Mechanical Translation*, 3(2):37, 1956.
92. M.P. Rombach, M.A. Porter, J.H. Fowler, and P.J. Mucha. Core-periphery structure in networks. *SIAM Journal on Applied mathematics*, 74(1):167–190, 2014.
93. M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
94. S.E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.
95. D. Schuler. Community networks: Building a new participatory medium. *Communications of the ACM*, 37(1):38–51, 1994.
96. J. Scripps, P.-N. Tan, and A.-H. Esfahanian. Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007*

Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD '07, pages 26–35, New York, NY, USA, 2007. ACM.
97. S.B. Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.
98. U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 210–217, Denver, Colorado, USA, 1995.
99. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
100. S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.-P. Schnurr, R. Studer, and Y. Sure. Semantic community web portals. *Computer Networks*, 33(1-6):473–491, 2000.
101. M. Takaffoli, F. Sangi, J. Fagnan, and O. Zäıane. Community evolution mining in dynamic social networks. *Procedia-Social and Behavioral Sciences*, 22:49–58, 2011.
102. J. Tang, X. Hu, and H. Liu. Social recommendation: a review. *Social Network Analysis and Mining*, 3(4):1113–1133, 2013.
103. Lei Tang and Huan Liu. Community detection and mining in social media (synthesis lectures on data mining and knowledge discovery). *Morgan & Claypool Publishers*, 2010.
104. S. van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.
105. I. Vragović and E. Louis. Network community structure and loop coefficient method. *Physical Review E*, 74:016105, Jul 2006.
106. J. Wang and T.T. Lee. Paths and cycles of hypergraphs. *Science in China*, 42(1), 1998.
107. S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.
108. J. Xie, S. Kelley, and B.K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys*, 45(4):43:1–43:35, August 2013.
109. X. Xu, N. Yuruk, Z. Feng, and T.A.J. Schweiger. Scan: A structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 824–833, New York, NY, USA, 2007. ACM.
110. Z. Xu, V. Tresp, A. Rettinger, and K. Kersting. Social network mining with nonparametric relational models. In *Revised Selected Papers of the Second International Workshop on Advances in Social Network Mining and Analysis (SNAKDD)*, volume 5498 of *Lecture Notes in Computer Science*, pages 77–96. Springer, 2010.
111. J. Yang, J. J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *International Conference on Data Mining*, 2013.
112. H. Zhang, M.-Y. Kan, Y. Liu, and S. Ma. Online social network profile linkage. In *Information Retrieval Technology*, pages 197–208. Springer, 2014.
113. E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 531–540, New York, NY, USA, 2009. ACM.
114. D. Zhou, J. Huang, and B. Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems*, pages 1601–1608, 2006.