Review

ENCYCLOPEDIA WITH SEMANTIC COMPUTING AND ROBOTIC INTELLIGENCE
Vol. 1, No. 1 (2017) 1630001 (8 pages)
© World Scientific Publishing Company
DOI: 10.1142/S2425038416300019

World Scientific
www.worldscientific.com

Community detection in social networks

Hossein Fani*,†,‡ and Ebrahim Bagheri*
*Laboratory for Systems, Software, and Semantics (LS3)
Ryerson University, Toronto, ON, Canada

†University of New Brunswick, Fredericton, NB, Canada
‡hosseinfani@gmail.com

Received 29 June 2016; Accepted 20 July 2016; Published 17 March 2017

Online social networks have become a fundamental part of the global online experience. They facilitate different modes of communication and social interactions, enabling individuals to play social roles that they regularly undertake in real social settings. In spite of the heterogeneity of the users and interactions, these networks exhibit common properties. For instance, individuals tend to associate with others who share similar interests, a tendency often known as homophily, leading to the formation of *communities*. This entry aims to provide an overview of the definitions for an online community and review different community detection methods in social networks. Finding communities are beneficial since they provide summarization of network structure, highlighting the main properties of the network. Moreover, it has applications in sociology, biology, marketing and computer science which help scientists identify and extract actionable insight.

Keywords: Social network; community detection; topic modeling; link analysis.

1. Introduction

A social network is a net structure made up of social actors, mainly human individuals, and ties between them. Online social networks (OSN) are online platforms that provide social actors, i.e., users, in spatially disperse locations to build social relations. Online social networks facilitate different modes of communication and present diverse types of social interactions. They not only allow individual users to be connected and share content, but also provide the means for active engagement, which enables users to play social roles that they regularly undertake in real social settings. Such features have made OSNs a fundamental part of the global online experience, having pulled ahead of email.[1] Given individuals mimic their real world ties and acquaintances in their online social preferences,[2] the tremendous amount of information offered by OSNs can be mined through social network analysis (SNA) to help sociometrists, sociologists, and decision makers from many application areas with the identification of actionable insight.[3,4] For instance, despite the heterogeneity of user bases, and the variety of interactions, most of these networks exhibit common properties, including the small-world and scale-free properties.[5,6] In addition, some users in the networks are better connected to each other than to the rest. In other words, individuals tend to associate with others who share similar interests in order to communicate news, opinions or other information of interest, as opposed to establishing sporadic connections; a tendency termed homophily as a result of which *communities* emerge on social networks.[7]

Communities also occur in many other networked systems from biology to computer science to economics, and politics, among others. Communities identify proteins that have the same function within the cell in protein networks,[8] web pages about similar topics in the World Wide Web (WWW),[9] functional modules such as cycles and pathways in metabolic networks,[10] and compartments in food webs.[11]

The purpose of this entry is to provide an overview of the definition of community and review the different community detection methods in social networks. It is not an exhaustive survey of community detection algorithms. Rather, it aims at providing a systematic view of the fundamental principles.

2. Definition

The word community refers to a social context. People naturally tend to form groups, within their work environment, family, or friends. A *community* is a group of users who share similar interests, consume similar content or interact with each other more than other users in the network. Communities are either *explicit* or *latent*. Explicit communities are known in advance and users deliberately participate in managing explicit communities, i.e., users create, destroy, subscribe to, and unsubscribe from them. For instance, Google's social network platform, Google+[a], has *Circles* that allows users to put different people in specific groups. In contrast, in this entry, communities are meant to be latent. Members of latent communities do not tend to show explicit membership and their similarity of interest lies within their social interactions.

[a]plus.google.com

No universally accepted quantitative definition of the community has been formulated yet in the literature. The notion of *similarity* based on which users are grouped into communities has been addressed differently in social network analysis. In fact, similarity often depends on the specific system at hand or application one has in mind, no matter whether they are explicit connections. The similarity between pairs of users may be with respect to some reference property, based on part of the social network or the whole. Nonetheless, a required property of a community is *cohesiveness*. The more users gather into groups such that they are intra-group close (internal cohesion) and inter-group loose (external incoherence), the more the group would be considered as a community.

Moreover, in *partitioned* communities, each user is a member of one and only one community. However, in real networks users may belong to more than one community. In this case, one speaks of *overlapping* communities where each user, being associated with a *mixture*, contributes partially to several or all communities in the network.

3. Application

Communities provide summarization of network structure, highlighting the main properties of the network at a macro level; hence, they give insights into the dynamics and the overall status of the network. Community detection finds application in areas as diverse as sociology, biology, marketing and computer science. In sociology, it helps with understanding the formation of action groups in the real world such as clubs and committees.[12] Computer scientists study how information is disseminated in the network through communities. For instance, community drives to connect like-minded people and encourages them to share more content. Further, grouping like-minded users who are also spatially near to each other may improve the performance of internet service providers in that each community of users could be served by a dedicated mirror server.[13] In marketing, companies can use communities to design targeted marketing as the 2010 Edelman Trust Barometer Report[b] found, 44% of users react to online advertisements if other users in their peer group have already done so. Also, communities are employed to discover previously unknown interests of users, alias implicit interest detection, which can potentially useful in recommender systems to set up efficient recommendations.[12]

In a very recent concrete application, Customer Relationship Management (CRM) systems are empowered to tap into the power of social intelligence by looking at the collective behavior of users within communities in order to enhance client satisfaction and experience. As an example, customers often post their opinions, suggestions, criticisms or support requests through online social networks such as Twitter[c] or Facebook.[d] Customer service representatives would quickly identify the mindset of the customer that has called into the call center by a series of short questions. For such cases, appropriate techniques are required that would look at publicly available social and local customer data to understand their background so as to efficiently address their needs and work towards their satisfaction. Important data such as the list of influential users within the community, the position of a given user in relation to influential users, the impact of users' opinions on the community, customer's social behavioral patterns, and emergence of social movement patterns are of interest in order to customize the customer care experience for individual customers.[14]

4. Detection

Given a social network, at least two different questions may be raised about communities: (i) how to identify all communities, and (ii) given a user in the social network, what is the best community for the given user if such a community exists. This entry addresses proposed approaches solving the former problem, known as community detection; also called community discovery or mining. The latter problem, known as community identification, is relevant but not aimed here.

The problem of community detection is not well-defined since its main element of the problem, the concept of community, is not meticulously formulated. Some ambiguities are hidden and there are often many true answers to them. Therefore, there are plenty of methods in the literature and researchers do not try to ground the problem on a shared definition.

4.1. *History*

Probably the earliest account of research on community detection dates back to 1927. At the time, Stuart Rice studied the voting themes of people in small legislative bodies (less than 30 individuals). He looked for *blocs* based on the degree of agreement in casting votes within members of a group, called Index of Cohesion, and between any two distinct groups, named Index of Likeness.[15] Later, in 1941, Davis *et al.*[16] did a social anthropological study on the social activities of a small city and surrounding county of Mississippi over 18 months. They introduced the concept of *caste* to the earlier studies of community stratification by social class. They showed that there is a system of colored caste which parsed a community through rigid social ranks. The general approach was to partition the nodes of a network into discrete subgroup positions (communities) according to some *equivalence* definition. Meantime, George Homans showed that

[b]www.edelman.co.uk/trustbarometer/files/edelmantrust-barometer-2010.pdf

[c]twitter.com

[d]www.facebook.com

social groups could be detected by reordering the rows and the columns of the matrix describing social ties until they form a block-diagonal shape.[17] This procedure is now standard and mainly addressed as *blockmodel* analysis in social network analysis. Next analysis of community structure was carried out by Weiss and Jacobson in 1955,[18] who searched for work groups within bureaucratic organizations based on attitude and patterns of interactions. The authors collected the matrix of working relationships between members of an agency by means of private interviews. Each worker had been asked to list her workers along with frequency, reason, subject, and the importance of her contacts with them. In addition to matrix's rows and columns reordering, work groups were separated by removing the persons working with people of different groups, i.e. *liaison* person. This concept of liaison has been received the name *betweenness* and is at the root of several modern algorithms of community detection.

4.2. *Contemporaries*

Existing community detection approaches can be broadly classified into two categories: *link-based* and *content-based* approaches. Link-based approaches, also known as topology-based, see a social network as a graph, whose nodes are users and edges indicate explicit user relationships. On the other hand, content-based approaches mainly focus on the information content of the users in the social network to detect communities. Also called topic-based, the goal of these approaches is to detect communities formed toward the topics extracted from users' information contents. Hybrid approaches incorporate both topological and topical information to find more meaningful communities with higher quality. Recently, researchers have performed a longitudinal study on the community detection task in which the social network is monitored at regular time intervals over a period of time.[19,20] Time dimension opens up a new *temporal* version of community detections. The following sections include the details of some of the seminal works in each category.

4.2.1. *Link analysis*

Birds of a feather, flock together. People tend to bond with similar others. The structures of ties in a network of any type, from friendship to work to information exchange, and other types of relationship are grounds on this tendency. Therefore, links between users can be considered as important clues for inferring their interest similarity and subsequently finding communities. This observation which became the earliest reference guideline at the basis of most community definitions was studied thoroughly long after its usage by McPherson *et al.*[7] as the homophily principle: '*Similarity breeds connection*'.

In link-based community detection methods, the social network is modeled by a graph with nodes representing social actors and edges representing relationships or interactions. Required cohesiveness property of communities, here, is reduced to *connectedness* which means that connections within each community are dense and connections among different communities are relatively sparse. Respectively, primitive graph structures such as components and cliques are considered as promising communities.[21] However, more meaningful communities can be detected based on graph partitioning (clustering) approaches, which try to minimize the number of edges between communities so that the nodes inside one community have more intra-connections than inter-connections with other communities. Most approaches are based on iterative bisection: continuously dividing one group into two groups, while the number of communities which should be in a network is unknown. With this respect, Girvan–Newman approach has been used the most in link-based community detection.[22] It partitions the graph by removing edges with high betweenness. The edge betweenness is the number of the shortest paths that include an edge in a graph. In the proposed approach, the connectedness of the communities to be extracted is measured using modularity (Sec. 5). Other graph partitioning approaches include max-flow min-cut theory,[23] the spectral bisection method,[24] Kernighan–Lin partition,[25] and minimizing conductance cut.[26]

Link-based community detection can be viewed as a data mining/machine learning clustering, an unsupervised classification of users in a social network in which the proximity of data points is based on the topology of links. Then, unsupervised learning which encompasses many other techniques such as k-means, mixture models, and hierarchical clustering can be applied to detect communities.

4.2.2. *Content analysis*

On the one hand, in spite of the fact that link-based techniques are intuitive and grounded on sociological homophily principle, they fall short in identifying communities of users that share similar conceptual interests due to two reasons, among others. Firstly, many of the social connections are not based on users' interest similarity but other factors such as friendship and kinship that do not necessarily reflect inter-user interest similarity. Secondly, many users who have similar interests do not share connections with each other.[27] On the other hand, with the ever growing of online social networks, a lot of user-generated content, known as social content, is available on the networks, besides the links among users. Users maintain profile pages, write comments, share articles, tag photos and videos, and post their status updates. Therefore, researchers have explored the possibility of utilizing the topical similarity of social content to detect communities. They have proposed content- or topic-based community detection methods, irrespective of the social network structure, to detect like-minded communities of users.[28]

Most of the works in content-based community detection have focused on probabilistic models of textual content for detecting communities. For example, Abdelbary *et al.*[29] have identified users' topics of interest and extracted topical communities using Gaussian Restricted Boltzmann Machines. Yin *et al.*[30] have integrated community discovery with topic modeling in a unified generative model to detect communities of users who are coherent in both structural relationships and latent topics. In their framework, a community can be formed around multiple topics and a topic can be shared among multiple communities. Sachan *et al.*[12] have proposed probabilistic schemes that incorporate users' posts, social connections, and interaction types to discover latent user communities in Twitter. In their work, they have considered three types of interactions: a conventional tweet, a reply tweet, and a retweet. Other authors have also proposed variations of Latent Dirichlet Allocation (LDA), for example, Author-Topic model[31] and Community-User-Topic model,[32] to identify communities.

Another stream of work models the content-based community detection problem into a graph clustering problem. These works are based on a similarity metric which is able to compute the similarity of users based on their common topics of interest and a clustering algorithm to extract groups of users (latent communities) who have similar interests. For example, Liu *et al.*[33] have proposed a clustering algorithm based on topic-distance between users to detect content-based communities in a social tagging network. In this work, LDA is used to extract hidden topics in tags. Peng *et al.*[34] have proposed a hierarchical clustering algorithm to detect latent communities from tweets. They have used predefined categories in SINA Weibo and have calculated the pairwise similarity of users based on their degree of interest in each category.

Like link-based methods, content-based community detection methods can be turned into data clustering in which communities are sets of points. The points, representing users, are close to each other inside versus outside the community with respect to a measure of distance or similarity defined for each pair of users. In this sense, *closeness* is the required cohesiveness property of the communities.

4.2.3. *Link jointly with content*

Content-based methods are designed for regular documents and might suffer from short, noisy, and informal social contents of some social networks such as Twitter or the like microblogging services. In such cases, the social content alone is not the reliable information to extract true communities.[35] Presumably, enriching social contents with social structure, i.e. links, does help with finding more meaningful communities. Several approaches have been proposed to combine link and content information for community detection. They have achieved better performance, as revealed in studies such as Ref. 36 and 37. Most of these approaches devise an integrated generative model for both link and content through shared latent variables for community memberships.

Erosheva *et al.*[38] introduce Link-LDA, an overlapping community detection to group scientific articles based on their abstract (content) and reference (link) parts. In their generative model, an article is assumed to be a couple model for the abstract and the reference parts each of which is characterized by LDA. They adopt the same bag-of-words assumption used in abstract part for the reference part as well, named bag-of-references. Thus, articles that are similar in the abstract and the references tend to share the same topics. As opposed to Link-LDA in which the citation links are treated words, Nallapti *et al.*[39] suggest to explicitly model the topical relationship between the text of the citing and cited document. They propose Pairwise-Link-LDA to model the link existence between pairs of documents and have obtained better quality of topics by employing this additional information. Other approaches that utilize LDA to join link and content are Refs. 40 and 41. In addition to probabilistic generative models, there are other approaches such as matrix factorization and kernel fusion for spectral clustering that combine link and content information for community detection.[42,43]

4.2.4. *Overlapping communities*

The common approach to the problem of community detection is to partition the network into disjoint communities of members. Such approaches ignore the possibility that an individual may belong to two or more communities. However, many real social networks have communities with overlaps.[44] For example, a person can belong to more than one social group such as family groups and friend groups. Increasingly, researchers have begun to explore new methods which allow communities to overlap, namely *overlapping* communities. Overlapping communities introduces a further variable, the membership of users in different communities, called *covers*. Since there is an enormous number of possible covers in overlapping communities comparing to standard partitions, detecting such communities is expensive.

Some overlapping community detection algorithms utilize the structural information of users in the network to divide users of the network into different communities. The dominant algorithm in this trend is based on clique percolation theory.[45] However, LFM and OCG are based on local optimization of a fitness function over user's out/in links.[46,47] Furthermore, some fuzzy community detection algorithms calculate the possibility of each node belonging to each community, such as SSDE and IBFO.[48,49] Almost all algorithms need prior information to detect overlapping communities. For example, LFM needs a parameter to control the size of communities. There are, also, some probabilistic

approaches in which communities are latent variables with distributions on the entire user space such as Ref. 50.

Recent studies, however, have focused on links. Initially suggested by Ahn *et al.*,[51] link clustering finds communities of links rather than communities of users. The underlying assumption is that while users can have many different relationships, the relationships within groups are structurally similar. By partitioning the links into non-overlapping groups, each user can participate in multiple communities by inheriting the community assignment of its links. Link clustering approach significantly speeds up the discovering of overlapping communities.

4.2.5. *Temporal analysis*

The above methods do not incorporate temporal aspects of users' interests and undermine the fact that users of communities would ideally show similar contribution or interest patterns for similar topics throughout the time. The work by Hu *et al.*[19] is one of the few that considers the notion of temporality. The authors propose a unified probabilistic generative model, namely GrosToT, to extract temporal topics and analyze topics' temporal dynamics in different communities. Fani *et al.*[20] follow the same underlying hypothesis related to topics and temporality to find time-sensitive communities. They use time series analysis to model user's temporal dynamics. While GrosToT is primarily dependent on a variant of LDA for topic detection, the unique way of user representation in Ref. 20 provides the flexibility of being agnostic to any underlying topic detection method.

5. Quality Measure

The standard procedure for evaluating results of a community detection algorithm is assessing the similarity between the results and the ground truth that is known for benchmark datasets. These benchmarks are typically small real-world social networks or synthetic ones. Similarity measures can be divided into two categories: measures based on pair counting and measures based on information theory. A thorough introduction of similarity measures for communities has been given in Ref. 52. The first type of measures based on pair counting depends on the number of pairs of vertices which are classified in the same (different) communities in the ground truth and the result produced by the community detection method. The *Rand index* is the ratio of the number of user pairs correctly classified in both the ground truth and the result, either in the same or in different communities, over the total number of pairs.[53] The *Jaccard index* is the ratio of the number of user pairs classified in the same community in both ground truth and the result, over the number of user pairs which are classified in the same community of result *or* ground truth. Both the Rand and the Jaccard index are adjusted for random grouping, in that a null model is introduced. The normal value of the index is subtracted from the expectation value of the index in the null model, and the result is normalized to [0, 1], yielding 0 for independent partitions and 1 for identical partitions. The second type of similarity measures models the problem of comparing communities as a problem of message decoding in information theory. The idea is that, if the output communities are similar to the ground truth, one needs very little information to infer the result given the ground truth. The extra (less) information can be used as a measure of (dis)similarity. The *normalized mutual information* is currently very often used in this type of evaluation.[54] The normalized mutual information reaches 1 if the result and the ground truth are identical, whereas it has an expected value of 0 if they are independent. These measures have been recently extended to the case of overlapping communities such as the work by Lancichinetti *et al.*[55]

Ground truth is not available in most cases of the real-world applications and there is no well-defined criterion for evaluating the resulting communities. In such cases, quality functions are defined as a quantitative measure to assess the communities. The most popular quality function is the *modularity* introduced by Newman and Girvan.[22] It is based on the idea that a random network is not expected to have a modular structure. The communities are going to emerge as the network deviate from a random network. Therefore, the more density of links exists in the actual community with compare to the expected density when the users were connected randomly, the more modular a community is. Simply, modularity of a community is the number of links within communities minus expected number of such links. Evidently, the expected link density depends on the chosen null model. One simple null model would be a network with the same number of links as the actual network and links are placed between any pair of users with the uniform probability. However, this null model yields a Poissonian degree distribution which is not a true descriptor of real networks. With respect to the modularity, high values imply *good* partitions. So, a community with maximum modularity in a network should be the near best one. This ignites a class of community detection which is based on modularity maximization. While the application of modularity has been questioned,[4] it continues to be the most popular and widely accepted measure of the fitness of communities.

As another quality function, the conductance of the community was chosen by Leskovec *et al.*[26] The conductance of a community is the ratio between the cut size of the community and the minimum between the total degree of the community and that of the rest of the network. So, if the community is much smaller than the whole network, the conductance equals the ratio between the cut size and the total degree of the community. A *good* community is characterized by a low cut size and a large internal density of links which result in low values of the conductance. For each real network, Leskovec *et al.* have carried out a systematic analysis on the quality of communities that have various sizes. They

derived the network community profile plot (NCPP), showing the minimum conductance score among subgraphs of a given size as a function of the size. They found that communities are well defined only when they are fairly small in size. Such small communities are weakly connected to the rest of the network, often by a single edge (in this case, they are called whiskers), and form the periphery of the network. The fact that the best communities appear to have a characteristic size of about 100 users is consistent with Dunbar conjecture that 150 is the upper size limit for a working human community.[56]

6. Future Direction

No doubt community detection has matured and social network analysts have achieved an in-depth knowledge of the communities, their emergence and evolution, in real social networks, but interesting challenges still yet to be addressed. As hinted in the introduction (Sec. 1), the quest for a single *correct* definition of network communities and a single *accurate* community detection method seem to be futile which is not necessarily a problem. Years of endeavors have resulted in many methods for finding communities based on a variety of principles. The picture that is emerging is that the choice of community detection algorithm depends on the properties of the network under study. In many ways, the problem of community detection has a parallel in the more mature topic of clustering in computer science, where a variety of methods exist, each one with standard applications and known issues. As a consequence, one challenge for the task of community detection is about distinguishing between existing ones. One way is to compare algorithms on real networks where network's metadata is available. In the case of social networks, for example, we can use demographic and geographic information of users. One example of such a metadata-based evaluation has been done in Ref. 57, but a standard framework for evaluation has yet to be emerged.

Moreover, networks are dynamic with a time stamp associated with links, users, and the social contents. As seen, almost all measures and methods ignore temporal information. It is also inefficient to apply such community detection algorithms and measures to static snapshots of the social network in each time interval. In order to truly capture the properties of a social network, the methods have to analyze data in their full complexity fueled with time dimension. This trend has slowly started.[19,20] What makes the problem more complex is the fact that online social network data arrive in a streaming fashion, esp. the social contents. The states of the network need to be updated in an efficient way *on the fly*, in order to avoid a bottleneck in the processing pipeline. To date, only a small number of work has approached this problem directly.[58]

The last area is the computational complexity in which the current methods will need dramatic enhancement, particularly with the ever-increasing size of current online social networks. As an example, Facebook has close to one billion active users who collectively spend twenty thousand years online in one day sharing information. Meanwhile, there are also 340 million tweets sent out by Twitter users. A community detection algorithm needs to be efficient and scalable, taking practical amount of time to finish when applied on such large-scale networks. Many existing methods are only applicable to small networks. Providing fast and scalable versions of community detection methods is one proposed direction worthy of significant future efforts. One solution to deal with large-scale networks is the sampling. The goal is to reduce the number of users and/or links while keeping the underlying network structure. Network sampling is done in the preprocessing step and is independent of the subsequent steps in community detection algorithms. Hence, it provides performance improvement to all community detection algorithms. Although sampling seems to be straightforward and easy to implement, it has a direct impact on the results of community detection, in terms of both accuracy and efficiency. It has been shown that naively sampled users or links by uniform distribution will bring bias into the output sampled network, which will affect the results negatively.[59] One of the challenges going forward in social network analysis will be to provide sampling methods and, particularly, to take sampling into account in the overall performance analysis of community detection methods.

References

[1]F. Benevenuto *et al.*, Characterizing user behavior in online social networks, *Proc. 9th ACM SIGCOMM Conf. Internet Measurement Conf.* (ACM, 2009).
[2]P. Mateos, Demographic, ethnic, and socioeconomic community structure in social networks, *Encyclopedia of Social Network Analysis and Mining* (Springer New York, 2014), pp. 342–346.
[3]R. Claxton, J. Reades and B. Anderson, On the value of digital traces for commercial strategy and public policy: Telecommunications data as a case study, World Economic Forum Global Information Technology Report, 2012-Living in a Hyperconnected World, World Economic Forum, (pp. 105–112)
[4]A. Stevenson and J. Hamill, Social media monitoring: A practical case example of city destinations, *Social Media in Travel, Tourism and Hospitality* (Ashgate, Farnham, 2012), pp. 293–312.
[5]D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393**, 440 (1998).
[6]A. L. Barabasi and R. Albert, Emergence of scaling in random networks, *Science* **286**, 509 (1999).
[7]M. McPherson, L. Smith-Lovin and J. M. Cook, Birds of a feather: Homophily in social networks, *Ann. Rev. Sociol.* **415** (2001).
[8]J. Chen and B. Yuan, Detecting functional modules in the yeast protein–protein interaction network, *Bioinformatics* **22**, 2283 (2006).
[9]Y. Dourisboure, F. Geraci and M. Pellegrini, Extraction and classification of dense communities in the web, *Proc. 16th Int. Conf. World Wide Web* (ACM, 2007).

[10]R. Guimera and L. A. N. Amaral, Functional cartography of complex metabolic networks, *Nature* **433**, 895 (2005).
[11]A. Krause *et al.*, Compartments revealed in food-web structure, *Nature* **426**, 282 (2003).
[12]M. Sachan, D. Contractor, T. A. Faruquie and L. V. Subramaniam, Using content and interactions for discovering communities in social networks, *21st Int. Conf. World Wide Web (WWW'12)*, (2012), pp. 331–340.
[13]B. Krishnamurthy and J. Wang, On network-aware clustering of web clients, *ACM SIGCOMM Computer Commun. Rev.* **30**, 97 (2000).
[14]Y. Richter, E. Yom-Tov and N. Slonim, Predicting customer churn in mobile networks through analysis of social groups, SDM, *SIAM* (2010), pp. 732–741.
[15]S. A. Rice, The identification of blocs in small political bodies, *Am. Political Sci. Rev.* **21**, 619 (1927).
[16]A. Davis *et al.*, *Deep South: A Sociological Anthropological Study of Caste and Class* (University of Chicago Press, 1941).
[17]G. C. Homans, The Human Croup (1950).
[18]R. S. Weiss and E. Jacobson, A method for the analysis of the structure of complex organizations, *Am. Sociol. Rev.* **20**, 661 (1955).
[19]Z. Hu, Y. Junjie and B. Cui, User Group Oriented Temporal Dynamics Exploration, AAAI (2014).
[20]H. Fani *et al.*, Time-sensitive topic-based communities on twitter, *Canadian Conf. Artificial Intelligence*. Springer International Publishing (2016).
[21]S. Fortunato, Community detection in graphs, *Phys. Rep.* **486**, 75 (2010).
[22]M. Girvan and M. E. J. Newman, Community structure in social and biological networks, *Proc. Nat. Acad. Sci.* **99**, 7821 (2002).
[23]L. R. Ford and D. R. Fulkerson, Maximal flow through a network, *Can. J. Math.* **8**, 399 (1956).
[24]A. Pothen, H. D. Simon and K.-P. Liou, Partitioning sparse matrices with eigenvectors of graphs, *SIAM J. Matrix Anal. Appl.* **11**, 430 (1990).
[25]B. W. Kernighan and L. Shen, An efficient heuristic procedure for partitioning graphs, *Bell System Tech. J.* **49**, 291 (1970).
[26]J. Leskovec, J. Kleinberg and C. Faloutsos, Graphs over time: Densification laws, shrinking diameters and possible explanations, *Proc. Eleventh ACM SIGKDD Int. Conf. Knowledge Discovery in Data Mining* (ACM, 2005).
[27]Q. Deng, Z. Li, X. Zhang and J. Xia, Interaction-based social relationship type identification in microblog, *Int. Workshop on Behavior and Social Informatics and Computing* (2013), pp. 151–164.
[28]N. Natarajan, P. Sen and V. Chaoji, Community detection in content-sharing social networks, *IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining* (2013), pp. 82–89.
[29]H. A. Abdelbary, A. M. ElKorany and R. Bahgat, Utilizing deep learning for content-based community detection, *Science and Information Conf.* (2014), pp. 777–784.
[30]Z. Yin, L. Cao, Q. Gu and J. Han, Latent community topic analysis: Integration of community discovery with topic modeling, *J. ACM Trans. Intell. Syst. Technol. (TIST)* **3**(4) (2012).
[31]M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth, The author-topic model for authors and documents, *20th Conf. Uncertainty in Artificial Intelligence* (2004), pp. 487–494.
[32]D. Zhou, E. Manavoglu, J. Li, C. L. Giles and H. Zha, Probabilistic models for discovering e-communities, *15th Int. Conf. World Wide Web* (2006), pp. 173–182.
[33]H. Liu, H. Chen, M. Lin and Y. Wu, Community detection based on topic distance in social tagging networks, TELKOMNIKA *Indonesian J. Electr. Eng.* **12**(5), 4038.
[34]D. Peng, X. Lei and T. Huang, DICH: A framework for discovering implicit communities hidden in tweets, *J. World Wide Web* (2014).
[35]T. Yang *et al.*, Combining link and content for community detection: A discriminative approach, *Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (ACM, 2009).
[36]D. Cohn and T. Hofmann, The missing link — a probabilistic model of document content and hypertext connectivity, NIPS (2001).
[37]L. Getoor, N. Friedman, D. Koller and B. Taskar, Learning probabilistic models of link structure, *J. MLR* **3** (2002).
[38]E. Erosheva, S. Fienberg and J. Lafferty, Mixed membership models of scientific publications, *Proc. Natl. Acad. Sci.* 101 (2004).
[39]R. M. Nallapati, A. Ahmed, E. P. Xing and W. W. Cohen, Joint latent topic models for text and citations, KDD (2008).
[40]L. Dietz, S. Bickel and T. Scheffer, Unsupervised prediction of citation influences, In ICML (2007).
[41]A. Gruber, M. Rosen-Zvi and Y. Weiss, Latent topic models for hypertext, UAI (2008).
[42]S. Zhu, K. Yu, Y. Chi and Y. Gong, Combining content and link for classification using matrix factorization, SIGIR (2007).
[43]S. Yu, B. D. Moor and Y. Moreau, Clustering by heterogeneous data fusion: Framework and applications, NIPS workshop (2009).
[44]J. Xie, S. Kelley and B. K. Szymanski, Overlapping community detection in networks: The state of the art and comparative study, *ACM Comput. Surv.* **45** (2013), doi: 10.1145/2501654.2501657.
[45]G. Palla, I. Derényi, I. Farkas and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature* **435**, 814 (2005).
[46]A. Lancichinetti, S. Fortunato and J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, *New J. Phys.* **11**, 033015 (2009), doi: 10.1088/1367-2630/11/3/033015.
[47]E. Becker, B. Robisson, C. E. Chapple, A. Guénoche and C. Brun, Multifunctional proteins revealed by overlapping clustering in protein interaction network, *Bioinformatics* **28**, 84 (2012).
[48]M. Magdon-Ismail and J. Purnell, SSDE-Cluster: Fast overlapping clustering of networks using sampled spectral distance embedding and gmms, *Proc. 3rd Int. Conf. Social Computing (SocialCom/PASSAT)*, Boston, MA, USA. NJ, USA: IEEE Press (2011), pp. 756–759, 10.1109/PASSAT/SocialCom.2011.237.
[49]X. Lei, S. Wu, L. Ge and A. Zhang, Clustering and overlapping modules detection in PPI network based on IBFO, *Proteomics* **13**, 278 (2013).
[50]W. Ren *et al.*, Simple probabilistic algorithm for detecting community structure, *Phys. Rev. E* **79**, 036111 (2009).
[51]Y.-Y. Ahn, J. P. Bagrow and S. Lehmann, Link communities reveal multiscale complexity in networks, *Nature* **466**, 761 (2010).
[52]M. Meilă, Comparing clusterings — an information based distance, *J. Multivariate Anal.* **98**, 873 (2007).
[53]W. M. Rand, Objective criteria for the evaluation of clustering methods, *J. Am. Stat. Assoc.* **66**, 846 (1971).

[54]L. Danon *et al.*, Comparing community structure identification, *J. Stat. Mech., Theory Exp.* **09**, P09008 (2005).
[55]A. Lancichinetti, S. Fortunato and J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, *New J. Phys.* **11**, 033015 (2009).
[56]R. Dunbar, *Grooming, Gossip, and the Evolution of Language* (Harvard University Press, Cambridge, USA, 1998).
[57]Y. Y. Ahn, J. P. Bagrow and S. Lehmann, Link communities reveal multi-scale complexity in networks, *Nature* (2010).
[58]C. C. Aggarwal, Y. Zhao and S. Y. Philip, On clustering graph streams, SDM (SIAM, 2010), pp. 478–489.
[59]Y. Ruan *et al.*, Community discovery: Simple and scalable approaches, *User Community Discovery* (Springer International Publishing, 2015), pp. 23–54.