



Analysis and Mining of Online Social Networks - Emerging Trends and Challenges

Sajid Yousuf Bhat and Muhammad Abulaish[#]

Department of Computer Science, Jamia Millia Islamia (A Central University)

Jamia Nagar, New Delhi - 110025, India

E-mail: s.yousuf.jmi@gmail.com, mAbulaish@jmi.ac.in

Abstract

Social network analysis (SNA) is a multi-disciplinary field dedicated to the analysis and modeling of relations and diffusion processes between various objects in nature and society, and other information/knowledge processing entities with an aim of understanding how the behavior of individuals and their interactions translate into large-scale social phenomenon. Due to exploding popularity of online social networks and availability of huge amount of user-generated content there is a great opportunity to analyze social networks and their dynamics at resolutions and levels not seen before. This has resulted in a significant increase in research literature at the intersection of the computing and social sciences leading to several techniques for social network modeling and analysis in the area of machine learning and data mining. Some of the current challenges in the analysis of large-scale social network data include social network modeling and representation, link mining, sentiment analysis, semantic SNA, information diffusion, viral marketing and influential node mining.

Introduction

The social world is a network of interactions and relationships that facilitates the flow and exchange of information and resources like norms, values, and ideas among individuals [White et al., 1976]. Such a view of the social world can be treated as a *social network*, and it can be defined as a social structure represented by a set of nodes and their inter-relationships, generally called *ties*. A node in a social network is usually called a *social actor*, and may represent a person, group, document, organization, or nation. A relation between a pair of nodes represents their ties reflecting friendship, kinship, dislike, common interest, acquaintance, financial exchange, physical connection, hyperlink, or co-location. Social network analysis (SNA) is one of the important techniques used in the field of sociology and also finds significant application in anthropology, biology, communication, economics, geography, and social computing [Scott 2000; Borgatti et al., 2009]. The increasing popularity of social networks is largely due to their relevance to various processes taking place in society, such as spread of cultural fads or diseases, formation of groups and communities, and recommendations. The process of social network analysis and modeling facilitates to understand how the behavior of individuals and their interactions translates into large-scale social systems. The application of SNA

[#] Corresponding author. E-mail: mAbulaish@jmi.ac.in

and social network concepts to a wide domain of research interests has gained huge popularity recently. For example, an application of SNA for transport planning is demonstrated in [Kowald et al., 2010]. Hulst (2008) highlighted the importance and applications of SNA for dealing with organized crime in adversary networks. With the variety of data such networks provide, the SNA tasks applicable (but are not limited to) include community (gang) identification, sentiment analysis and opinion mining, node influence analysis, and link prediction. The work of [Lewis et al., 2010] signified that the clusters or community analysis in protein interaction networks using SNA techniques can highlight functionally coherent groups of proteins and predict the level of function homogeneity within these protein groups or communities. Similarly, Swan et al. (1999) throw some light on the importance of community-wise management of knowledge and how managed intra-community and inter-community (across structural holes) interactions lead to significant and novel innovations. Using SNA for perceiving, controlling and distributing strategic and domain knowledge in organizations [Patak et al., 2007] and collaborative distance-learning [Reffay and Chainer, 2003] is also promising, considering the improvement of throughput for such systems.

Bonchi et al. (2011) presented a state-of-the-art survey on the business application of SNA in an online social network environment. They highlighted the exploitation of various SNA concepts like social contagion, influence, communities, and ranking for facilitating many business challenges like marketing, customer service, and managing resources (financial, human, knowledge). On the other hand, this paper aims to present some important challenges of social network analysis in a generic domain, and it highlights how online social networks facilitate the analysis and understanding of such challenges by providing a spectrum of opportunities and data that are highly related to the real-world. A huge amount of literature along the direction of social network analysis exists. Most of them concentrate on some specific aspect of the social networks (like community detection, information diffusion, etc.), in isolation or are mostly oriented towards sociological aspects relieving computer science. However, the present multidimensional online social networks provide a means of studying various data mining tasks related to social network analysis in a unified framework where each of them can benefit from the others. In this regard, we present a survey on the current state-of-the-art and challenges related to social network analysis in the light of online social networks and also present the design of a possible unified framework for some major tasks related to social network analysis.

This paper introduces some of the recent data mining tasks targeting social networks, taking into consideration data about the structure of social networks. Different data mining tasks and techniques require that the social networks be represented and modeled according to the needs of the analysis being performed in the task. As a result, various alternatives for social network modeling exist and can be categorized into different groups depending upon the techniques used and the level of analysis supported. Some of the popular social-network modeling techniques have been reviewed in section 2. For a detailed description of the graph-theoretic properties and social network metrics that form a basis for almost all kind of analyses of social networks readers should refer to [Mislove, 2009; Rupnik, 2006; Scott, 2000; Wasserman and Faust, 1994].

1 Social Network Properties and Data

Network systems have been traditionally considered to be random structures and despite links being considered to occur at random between nodes, most nodes were expected to have almost the same degree. However, significant contributions made by researchers like [Barabási and Albert, 1999; Albert et al., 1999] revealed that the vertex connectivity of large scale real-world networks actually follows a scale-free power-law distribution. That is, for large values of k , the fraction $P(k)$ of nodes having k connections to other nodes in the network follows the relation as shown in equation 1, where c is a normalization constant and γ is a parameter usually ranging between $2 < \gamma < 3$.

$$P(k) \sim ck^{-\gamma} \quad (1)$$

The growth of such networks involves a rich-get-richer scheme (preferential attachment) wherein new nodes have a higher probability to link to nodes of higher degree, or it can be stated that the

likelihood of a node acquiring a new link is in proportion to the node's degree [Barabási and Bonabeau, 2003].

Some special properties of social networks that differentiate social networks from other networks as highlighted by Newman and Park (2003) are: Firstly, social networks show positive correlations between the degrees of adjacent vertices (assortativity). More specifically, vertices of similar degree tend to be connected more with each other than with others. Secondly, social networks have non-trivial clustering or network transitivity. This property makes the networks to exhibit community structures, i.e., clusters of vertices/nodes that are more similar or connected within the group than to the rest of the network. Communities in social networks often map to important functional or interest groups of the underlying nodes and designing methods and techniques to identify them is a challenging task.

Milgram (1967) showed that in a well-defined population the average path length between two individuals so that they can meet each other was six hops, demonstrating that social networks can be classified as *small-world* which led to the famous phrase "*six degrees of separation*". Another famous concept of sociology is the weak link hypothesis by Granovetter (1973), according to which the degree of overlap between the friend neighborhoods of two individuals is observed to increase as a function of the strength of the tie connecting these two individuals. This specifies that strong ties are tightly clustered, whereas the weak ties represent longer-distance relationships, thus playing an important role for the flow of information and innovation. This phenomenon is known as *the strength of weak ties*. Based on this concept and the existence of groups in social networks, Burt (1992) defined *structural holes* as the topological scarcity or the weakness of links between the groups in a social network. In terms of the productivity perspective of organization control, structural holes seem to provide an opportunity of the brokerage of information flow between different working groups. This in turn provides an opportunity to control the projects on which various groups across a structural hole work. Within groups, opinion and behavior tend to be more homogeneous than otherwise. Thus, individuals who connect groups across structural holes often tend to have alternative ways of thinking and approaching a problem as they are possibly exposed to multiple activities. Brokerage between groups across the structural holes highlights alternative options that otherwise remain unexplored and based on this property, brokerage across a structural hole becomes social capital [Burt, 2004]. Burt (1997) signifies that an increase in the number of individuals performing the same task decreases the value of social capital and thus peers tend to lose the value of social capital to individuals (often managers) who have a very less number of peers.

The theory proposed by Granovetter (1973) related to the strength of weak ties highlights an important property of *contagion*, i.e., diffusion of diseases and innovations (belief, ideology, norm, technology, organizational form, fad, or fashion) in social and information networks through physical and/or virtual contacts. That is, the reach of a diffusion process (social distance covered) is significantly higher if it is passed through weak ties rather than strong. Moreover, *small-world* networks composed of a few long inter-community ties between tightly clustered communities facilitating rapid diffusion of information and disease.

1.1 Social Network Data Sources

For analyzing social networks, the primary requirement is the access to *social network data* which can be viewed as a social relational system characterized by a set of actors and their social ties [Wasserman and Faust, 1994]. Additional information in the form of actor attributes or multiple relations can be a part of the social relational system. Traditional sources of social network data included questionnaires, interviews, and observations. However, acquiring data through these sources is laborious and costly, and restricts the analysis to a smaller number of individuals, resulting in possibly significant individual biases. With the availability of large electronic datasets like the e-mail networks and telephone call graphs, and efficient computational resources, in the late 1990s, physicists entered the field of social networks (and complex networks in general) with a concern of analyzing topological properties of networks, developing new concepts, algorithms, and models. The advantage of such electronic datasets is that they are large, relatively easy to process, and accurate in the sense that subjective biases are absent [Kumpula, 2008].

Wu et al. (2008) argued that besides proxy datasets like e-mail networks, face-to-face (F2F) interactions also remain a powerful conduit for information exchange, especially for complex or tacit information. They incorporated wearable sociometric badges that can collect and analyze behavioral data from individuals over time by detecting people in close proximity, capturing face-to-face interaction time, and recording tonal variation and prosody using a microphone. Alternatively, Eagle et al. (2009) significantly showed that data gathered from the usage of mobile phones can be used to get a significant insight into the relational dynamics of user behavior. Besides communication information presented by communication networks like e-mail networks and telephone call graphs, the information provided by mobile phone data also spans to the behavior, location and proximity of mobile phone users using GPS, Bluetooth, cell tower IDs, application usage. With an aim of comparing the user behavior represented by the data collected from mobile phones with user-reported data collected from direct user survey, the analysis of Eagle et al. (2009) highlighted that the former, as a complement to the later, provides significant insights into purely cognitive constructs, such as friendship and individual satisfaction besides observable behavior. Such data can easily be mapped to the actual F2F interactions and relations between individuals.

1.2 The Online Social Network Boom

The growth of the World Wide Web has led to the evolution of different types of information sharing systems, which include online social networks (OSNs) like Facebook, MySpace, Flickr, LastFM, Digg, Bebo, Orkut, hi5, LinkedIn, LiveJournal, Twitter, etc. In the recent years, OSNs like Facebook have achieved significant popularity and now represent the most popular web sites. OSNs provide individuals a means of joining a network (become users), provide information which could define them or their preferences (profile), and enable them to publish any content which they like to share with other users. One of the important and outstanding features provided by these OSNs is to enable users to create links to other users with whom they associate. These user-centric features of OSNs enables a user to define and maintain social relations, find and link to other users with similar interests and preferences, and share, find and endorse content and knowledge contributed by a user itself or by other users [Mislove et al., 2007]. Considering the extreme popularity, huge membership and the enormous amounts of social network data generated by these online social networks, there exists a unique opportunity to study, understand, and leverage their properties. An in-depth analysis of online social network structure and growth can not only aid in designing and evaluating current systems, but it can also lead to better design of future online social network based systems and to a deeper understanding of the impact of online social networks on society.

The size of social networks is growing every day and the huge amount of data being produced by them is obviously leading to information explosion in the area of analyzing social networks. This necessitates the application of computational techniques to analyze the structure and nature of such complex networks more efficiently and accurately. Along with the sociologists developing social network analysis (SNA) methods to discover the properties of social networks, computer scientists are developing and applying data mining techniques to discover hidden patterns from social network data.

2 Social Network Representations

This section presents some of the basic techniques used for representing social networks. The amount of knowledge gained from a social network and the level of analysis that can be efficiently performed on it often depend upon the scheme used for representation. Representation issues are some of the preliminary issues faced. Certain representations naturally allow mathematical analysis while some stress only on theoretical exploration and reasoning. The usage of a particular representation scheme depends upon the problem at hand and the nature of tasks and operations that need to be performed on the network, and the level of characteristic details that need to be visualized about the network. Some of the common representations are presented in the following sub-sections.

2.1 Node-Link Representation

Graph-based models have been extensively used to analyze social networks by considering different types of graph such as undirected [Lahiri and Berger-Wolf, 2008; Bhattacharyya et al., 2009; Falkowski et al., 2008], directed/weighted [Chun et al., 2008], and bipartite [Tantipathananandh and Berger-Wolf, 2009]. Graph-based social network models consist of nodes to represent actors, and links to represent ties or relations. Sociologists refer to such graph representations of social networks as *sociograms*. Rendering a sociogram along with a summary of graph theoretical concepts for visualization provides a basic description of social network data. However, this might suffice for small graphs, but usually the data and/or research questions are too complex for this relatively simple approach [Jamali and Abolhassani 2006]. Alternatively, many node-link variants in 2D and 3D have been experimented in the information visualization (InfoVis) community¹, however for large graphs or graphs with high link density all these visualization techniques show highly overlapping edges resulting in occlusion. This makes it very difficult for any users to gain a visual picture of a graph, or to select or find a particular node or edge in a graph [Ghoniem et al., 2005].

2.2 Matrix Representation

The matrix representation as used by many researchers [Mislove, 2009; Tang and Lui 2010; Chen et al., 2009] to analyze social networks involves representing numerous important social network actor-based activities and concepts like interactions, friendship, citations, community subscriptions, information access, and interests in a matrix form. The most basic form of the matrix representation is binary and is called the *adjacency matrix* that uses 1 for an existing conceptual link between two objects and a 0 for a non-existing link. Other matrix representations may involve weights or intensities to represent the importance or priority of the links or their corresponding relationships. The *Laplacian matrix* $L = (l_{ij})_{n \times n}$ sometimes called *admittance matrix* or *Kirchhoff matrix* is a positive semi-definite matrix representation of a graph as shown in equation 2, where $\deg(v_i)$ is the degree of the node v_i .

$$l_{i,j} = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise .} \end{cases} \quad (2)$$

It is clear from equation 2 that the Laplacian matrix of a graph is simply taken as its degree matrix minus its adjacency matrix. The Laplacian matrix has at least one zero eigenvalue, and the number of such eigenvalues is equal to the number of disjoint parts in the graph. A normalized Laplacian matrix, as shown in equation 3, has been also considered by the researchers.

$$l_{i,j} = \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise .} \end{cases} \quad (3)$$

The normalization factor means that the largest eigenvalue is less than or equal to 2, with equality only when the graph is bipartite. Eigenvalues of the graph are called graph spectra and they give information about some basic topological properties of the underlying graph. Laplacian spectra of networks have been investigated enormously to understand the synchronization of coupled dynamics on networks [Jalan and Bandyopadhyay, 2008].

Graph representations using matrices provide an efficient alternative to the traditional node-link diagrams [Ghoniem et al., 2005] and the effectiveness of various matrix representations is significantly highlighted in Ghoniem et al. (2004). Ghoniem et al. (2005) have argued that even though matrices form a quick representation and provide higher resolution readability for various network analysis tasks involved, the matrix-based representation seems under-exploited. An increased familiarity of matrix representation, due to its wider use, has resulted in an improvement

¹ <http://www.infovis-wiki.net>

in its readability. On the other hand, path-related tasks are still challenging on matrix representations. To address matrix-based graph representations' weaknesses on the path-related tasks that are important for social network analysis, Henry and Fekete (2007) have proposed enhanced matrix visualization for graphs, MatLink, wherein node-links are overlaid on a matrix representation at its edges. It supports a fast and effective visualization of path-relationship between nodes using this representation by simply highlighting nodes using a mouse pointer. Node-link and matrix social network representations are the most popular and basic ones being used.

2.3 Semantic Representation of Social Networks

Tim Berners-Lee (the inventor of the WWW) envisioned the *Semantic Web*, wherein the basic concept is to enable sharing of data across wider contextual communities by enriching the Web with machine-lucid information for both its automatic and manual processing [Horrocks, 2008; Stumme et al., 2006]. The core concept of Semantic representation of data is *Ontology*, which is defined as a *formal, explicit specification of a shared conceptualization in the form of an explicitly specified vocabulary which describes the various aspects of the domain being modeled* [Gruber, 1993; Horrocks, 2008].

Gruber (2008) advocates for the possibilities and significance of combining semantic web framework with social media domain, with an aim of developing collective intelligence and knowledge systems. Similarly, Downes (2005) stresses on the need of semantic social networks for effective information retrieval, like personalized and efficient content search. According to Breslin and Decker (2007), the Semantic Web provides the representation and navigation mechanisms by linking people and objects to record and represent the heterogeneous ties that bind them to each other. This paper considers the fact that online social network data can be seen as a twofold structure: data describing the social network structure, and data describing the content produced by network members. Some of the ontologies that exist for representing this view of online social networks include:

- **Friend-of-a-Friend (FOAF)²**: This ontology is used to describe people, their relationships and their activity. It includes a large set of properties to describe a user profile (*family name, nick, interest* and so on), web usages (*online accounts, weblogs, memberships* and so on), relationships to connect people (*knows*), and so on [Erétéo et al., 2011].
- **RELATIONSHIP³**: This ontology specializes the *knows* property of the FOAF ontology by including properties to more precisely define the type of relationship between people [Erétéo et al., 2011].
- **Semantically-Interlinked Online Communities (SIOC)⁴**: This ontology is aimed at linking discussion posts to other related discussions, people, and topics emerging on platforms such as blogs, message boards, and mailing lists [Breslin and Decker, 2007] by including properties that specialize *OnlineAccount* and *HasOnlineAccount* from FOAF [Breslin et al., 2005]. Besides conventional discussion platforms, SIOC is evolving to describe new Web-based communication and content-sharing mechanisms [Breslin and Decker, 2007].
- **Simple Knowledge Organization Systems (SKOS)⁵**: This ontology involves organizing knowledge in a hierarchical manner (e.g., narrower, broader, related) and to link it to SIOC descriptions using *isSubjectOf* property [Erétéo et al., 2011].

The most popularly used ontology to represent the network of people in case of online social networks is the FOAF. Initially the FOAF data was produced by hand wherein the interested users would create a file that contained personal data, including email addresses, location, interests, a list of friends, etc., and make it available as a web-accessible resource. These early FOAF implementations suffered from inconsistent tag usage and occasional lack of clarity in the

² <http://www.foaf-project.org/>

³ <http://vocab.org/relationship/>

⁴ <http://sIOC-project.org/>

⁵ <http://www.w3.org/2004/02/skos/>

specifications and often lead to extending the working schema of FOAF whenever new and efficient conceptualizations were encountered.

One of the challenges faced by semantic web technology is *ontology learning or extraction* wherein the attempt is to automatically recreate a conceptual model/vocabulary from existing knowledge sources, in particular natural text. Mika (2005b) signifies that communities in social networks create their own ontologies that reflect their identities, languages, and collective intelligence through interactions on some particular interests. These ontologies can help to annotate the content created by the individuals in their respective communities. Mika (2005b) also visualizes a three layered architecture of the semantic web mapping to communities, ontologies, and content. This concept of ontology learning is basically related to the structure of online social networks which not only allow users to publish resources, but also allow them to annotate these resources by assigning short descriptive tags to them. Since a consensus of community users is defining the meaning for a resource (content), these social tags represent objects around which those users form more tightly connected social networks [Breslin and Decker, 2007]. This neologism for collaborative categorization of objects on online social networking sites, including web blogs and social bookmarking services like del.icio.us, using freely chosen keywords is termed as *folksonomy* [Mika, 2006]. A related problem to automatic ontology extraction for semantic web is automatic semantic annotation of web content. In this regard, Wu et al. (2006) show how folksonomy can be exploited to infer a global semantic model for semantically annotating web resources. Folksonomies can be improved by adding semantics that structure and link tags together. The proposed semantic model also aims in disambiguating tags and group synonymous tags together in concepts and can be used to efficiently search and discover semantically-related web resources. The main issue related to folksonomy, which the earlier mentioned works attempt to relate to is that tags lack an explicit structure and mostly do not relate to each other semantically. Folksonomies have generally been considered of consisting of two levels of descriptive tags one that tend to describe abstract concepts and can be understood by many and the other level represents tags that have a specific meaning within a particular context and community of people. This issue has been mostly faced with approaches to enrich folksonomies by bridging them with formal ontologies. The main property of folksonomies is that they consist a triadic structure wherein *people* associate *tags* to *resources*. Thus adding structure to the tags (e.g. link tags to domain ontology) implicitly adds structure to the set of users according to the structure of tags. For example, Mika (2006) applies social network analysis on different projections of the tripartite structure of folksonomies grouped similar communities of interest, i.e., groups of people sharing common tags. This in turn yields subsumption properties between the tags. The semantic knowledge of folksonomies can also guide in supporting folksonomy based social platforms [Erétéo et al., 2011].

Erétéo et al. (2011) have proposed a semantic web social network analysis framework based on extending the existing social data representation ontologies to *SemSNA* ontology. Their main aim is to enhance social network representations by combining the structure and content of networks by annotating the semantic social network with important social network indices like centrality, community membership and so on. Based on SPARQL formal definitions of SNA operators they compute semantically parameterized SNA features to annotate the graph nodes and record the results. Their analysis suggests the semantic representations of social networks have a significant effect on connecting and exchanging the social data and the knowledge embedded in the social network. The semantic representation provides a rich domain for representing the complete topological and meta-information describing real-world social networks like OSNs along multiple dimensions. Designing novel semantic framework(s) that allow the dynamic representation and analysis of online social networks seems to be a perfect future goal along this direction.

3 Social Network Models

Statistically social networks involve representing a set of n objects and their relationships using an $n \times n$ adjacency matrix X . Each entry x_{ij} in X represents a binary value (either 0 or 1) to record the presence or absence of relation between objects i and j , e.g., existence of a friend relation between

two individuals represented by i and j . In a general representation, x_{ij} can be a discrete value representing the intensity or strength of a relation between objects i and entity j on a particular scale. Moreover, an object O can have a set of attributes or properties $A=\{a_1, a_2, \dots, a_m\}$ that define the object, e.g., demographic properties of an individual. In this regard, a traditional goal is to learn a model which could explain the probability of the existence or absence of a relation between objects based on the set of their attributes O and the topological properties of individual objects (nodes) in the underlying network [Smyth 2003].

Considering the complex and irregular structure of social networks, naturally all models which bear resemblance to real social networks involve randomness in network construction rules [Kumpula, 2008]. The most common categorization of social network models is *topological models* and *distance-based models*. The difference being that the former utilizes only the network structure in its rules, whereas the latter assigns an intrinsic (random) coordinate for each node, and closer nodes are more likely to be linked to each other than distant nodes. In distance-based models the node coordinates can be interpreted as real geographic coordinates or alternatively as coordinates in an abstract *social space*, which may represent hobbies, opinions, occupation, etc. [Wong et al., 2006]. Topological models, on the other hand, try to model real networks by basing the network construction rules solely on the network topology and can be further classified based on whether the representation model incorporates dynamic behavior (evolving nature) of the social networks or not, i.e., whether the model is static or dynamic. Early researches in social network analysis have primarily focused on the static properties of these networks, neglecting the fact that most real-world interaction networks are dynamic in nature, but recent works have considered the dynamic nature of social networks to find evolving trends and patterns in them [Snijders et al., 2010]. Spiliopoulou (2011) presents a comprehensive survey on the dynamic community based models for social network evolution. It is desirable to identifying the sections of a network that tend to mutate, characterizing the type of mutation, predicting future mutations or events (e.g., link prediction, predicting split or merger of a network or its parts). In other terms, developing generic models for evolving networks is a challenge, which needs to be addressed. For instance, the rapid growth of online communities has dictated the need for analyzing large amounts of temporal data to reveal community structure, dynamics and evolution [Asur et al., 2009]. Jamali and Abolhassani (2006) have classified the social network models by considering formal methods [Hanneman and Riddle, 2005] for representing social networks. They classify social network models based on (i) descriptive methods and graphical representations, (ii) analysis procedures, often based on the decomposition of adjacency matrix, and (iii) statistical models based on probability distributions. Here, we have used a similar classification to make a distinction between various social network models. In addition we also present a review related to the models that have gained significant importance recently which include adaptive and game theoretic network models. Some of the popular statistical models for social networks are classified and explained below.

3.1 Exponential Random Graph Model or P^* Model

Exponential Random Graph Models (ERGMs) [Andersen et al. 1999; Wasserman and Pattison 1996] or the p^* models are built on the idea of $p1$ and $p2$ models and are considered as a promising way to model network structure via a series of substructures for cross-sectional data, and provide a general framework for descriptively modeling a static network. The $p1$ model [Holland and Leinhardt, 1981] defining the first probability distribution for binary dyadic data (i.e., data composed of two sets of objects A and B , in such a way that observations are basically observations of couples (a, b) , with $a \in A$ and $b \in B$) is a model for the four possible dyadic outcomes, one mutual, one null, and two asymmetric where the data (adjacency) matrix, x , is a realization of a random matrix X , in which each dyad, $D_{ij} = (X_{ij}, X_{ji})$ is an independent bivariate random variable with possible values as given in equation 4. The $p1$ model defines the nature of ties to be reciprocal and includes parameters which aim to define the tendency of an individual to extend and accept ties in the network. However $p1$ models are restrictive as they assume the dyads to be independent.

$$D_{ij} = \begin{cases} (1,1) & i \text{ links with } j \text{ and } j \text{ links with } i \\ (1,0) \text{ or } (0,1) & i \text{ links with } j \text{ or } j \text{ links with } i \\ (0,0) & \text{otherwise.} \end{cases} \quad (4)$$

The $p2$ model [Van-Duijn et al., 2004] in turn is a random effect variant of the $p1$ model in which the sender and receiver parameters are modeled as correlated random effects, a formulation that makes it possible to include actor and dyad-specific covariates as fixed sender, receiver, density, or reciprocity regression parameters. The $p1$ and $p2$ models for network structure focus on the dyads in the network. However, triangles (or triads) are important for several reasons to analyze social networks as given in [Handcock, 2002]. The idea of Markov Graphs [Frank and Strauss 1986], which is an extension to the $p1$ model, allow for triads in the network through the notion of conditional dependence.

Considering X as a random graph with N nodes, x_{ij} as a random variable representing the possibility of a tie between nodes i and j , and χ as the set of all such graphs, the exponential random graph models are defined based on the likelihood of the occurrence of such graphs as given by equation 5:

$$P_{\theta, \chi}(X = x) = \frac{\exp\{\theta' u(x)\}}{c(\theta, \chi)} \quad (5)$$

Here the parameters for the model are specified in the form of vector θ and the function $c(\theta, \chi)$ normalizes the value in the range of $[0, 1]$ such that the sum of all observed values yields one. The nature of the ERGM model is defined with the statistics specified by the vector $u(x)$ defining a particular realization x of a network [Toivonen et al., 2009]. A detailed description of these models can be found in [Wasserman and Robins, 2005].

Hanneke and Xing (2007) have extended EGRMs to *temporal* EGRMs for modeling networks evolving over discrete timestamps. However, one of the major problems associated with these models is that of inferential degeneracy (tendency to converge to either empty or complete graphs), as analyzed in [Handcock, 2003]. New specifications for the ERGM proposed in [Snijders et al., 2006] attempt to find a solution for the degeneracy via a different parameterization of the models. In [Robins et al., 2007], the authors have reviewed these new specifications and their experiments suggest that the resulting graphs tend to be more realistic and that the near degeneracy problem is avoided particularly in networks that show highly transitive relations.

3.2 Preferential Attachment Models

In 1999, Barabási and Albert noticed that real world social networks showed *power law* degree distributions, i.e., networks that grow from a small nucleus of nodes and follow a “rich-get-richer” scheme in terms of node degrees. They are also called scale-free random graphs. Barabási and Albert [Barabási and Albert, 1999] described a dynamic preferential attachment (PA) model specifically designed to generate scale-free networks and model the network growth as follows: From any current state of a network with n_0 nodes, each subsequent time step adds a new node with $m \leq n_0$ edges. The probability p_i that the new node will be connected to an existing node i depends on the degree $\text{deg}(i)$ of i according to the multinomial distribution of equation 6.

$$p_i = \frac{\text{deg}(i)}{\sum_j \text{deg}(j)}. \quad (6)$$

The preferential attachment model of Barabási and Albert results in a network with a power law degree distribution whose exponent is empirically determined to be $\gamma_{BA} = 2.9 \pm 0.1$ [Goldenberg et al., 2010].

3.3 Small World Models

The Watts-Strogatz (WS) small world model (Watts and Strogatz, 1998), for large networks incorporates some of the important characteristics of real-world networks including transitivity, limited degrees, and limited path lengths (geodesics). By taking a one-dimensional lattice of L vertices with each vertex being connected with its $2k$ nearest neighbors and periodic boundary conditions (the lattice is a ring), the model proceeds with “rewiring” each bond independently with some probability ϕ . The probability ϕ is varied such that the transition between order ($\phi = 0$) and randomness ($\phi = 1$) can be monitored closely. Rewiring here relates to relocating one end of a bond to a different vertex selected at random from the whole lattice, with conditions that there cannot be more than one bonds between two vertices and no loops can occur. This process generates $\phi Lk/2$ long-range edges which connect nodes that would in otherwise belong to discrete neighborhoods. The behavior of the network thus depends on three independent parameters: L , k and ϕ . In this model the average coordination number z remains constant ($z = 2$) during the rewiring process, but the coordination number of any particular vertex may change. The total number of rewired bonds, which are referred to as “shortcuts”, is ϕL on average. Newman and Watts [Newman and Watts, 1999], pointed out that the distribution of shortcuts in the WS small world model is not completely uniform which makes an average over different realizations of the randomness hard to perform. Furthermore, the average distance between pairs of vertices on the graph is poorly defined as there is a finite probability of a portion of the lattice becoming detached from the rest in the model. Newman and Watts [Newman and Watts, 1999], propose a variant of the WS small world model which overcomes these limitations. In this model, a new edge is inserted for a pair of nodes with a probability p without breaking any bond between any two nearest neighbors and also does not form isolated clusters. However, the conditions to form a bond (edge) are same as for the previous model. When $p = 0$, the Newman and Watts (NW) model forms a nearest-neighbor coupled network, and for $p = 1$, it forms a globally coupled network. Another variant proposed by Kleinberg, (2001), starts with an underlying finite-dimensional grid and adds shortcut edges where the probability that two nodes are connected by a long edge depends on the distance between them in the grid. That is, the probability that two non-adjacent nodes x and y are connected is proportional to $d(x; y)^{-\alpha}$. With α set to the dimension of the lattice, the greedy routing algorithm can find paths from one node to another in a polylogarithmic number of expected steps.

3.6 Game Theoretic Models

In an economical perspective of social networks, social actors tend to connect to each other with an aim of gaining incentives, payoffs or utilities from the connections they establish [Hellmann et al., 2012]. Moreover, forming a link may incur a cost and thus social actors can be seen to form links strategically in a social network [Narayanam and Narahari, 2011]. In case a social actor gains incentives which result from the actions of all the social actors in a social network, the network formation and growth can be appropriately modeled using the concept of game theory [Hellmann et al. 2012]. Game theory deals with modeling situations wherein two or more decision makers (players) seek to make mutually influencing decisions. For an introduction to game theory, the reader is referred to [Gintis, 2000]. In a basic sense, a game consists of the *players* deciding upon the next step(s), *Rules* or *Strategies* according to which the steps/decisions are taken, *Outcomes* resulting from each possible decision taken by a player and *Preferences* that a player expects or prefers out of the possible outcomes.

A network formation game is analogously defined as a 3-tuple $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ where N is the set of social actors (players), S_i is the set of strategies for each player i wherein a strategy $s_i \in S_i$ of a player i is the set of other players with which i wants to form a link, u_i is the incentive/utility of player i which depends on its neighborhood and the structure of the network [Yadati and Narayanam, 2011]. Initial game theoretic models for network formation include that of Aumann and Myerson (1988) where node pairs are given ordered opportunities to establish permanent links with mutual consent. The procedure is repeated until the remaining pairs reject these opportunities [Hellmann et al. 2012]. A better approach was later proposed by Myerson (1991) which involves that all the nodes (players) announce their strategies simultaneously such that a link I, j is established only if $I \in s_j$ and $j \in s_i$ [Jackson, 2005]. Some other models of network formation strive to maintain the

stable state of a network directly wherein stability is defined in terms of strategic equilibrium. For example, *Nash Equilibrium* defines a network to be in stable state if no node forms or deletes a link to another node unilaterally. Similarly, *Pairwise Stability*, defines that no node gains an incentive on deleting a link and that no pair of nodes exit in a network that want to form a mutual link resulting in a mutual incentive [Hellmann et al. 2012; Jackson, 2005]. Other alternatives include the *Strong Stability*, wherein the changes induced by a set S of nodes without the consent of the nodes $\notin S$ are such that any new links are only added within the set S and that at least one node of any deleted link is in S [Jackson, 2005]. Similarly, certain economic game models of network formation aim to maximize the efficiency of the network i.e. the function of the utilities of the nodes. For example, *Pareto Efficiency*, requires that for a network state g there should exist no network state g' where the utility of a node is greater than it has in g and the utilities of the remaining nodes is greater than or equal to that they have in g [Yadati and Narayanam, 2011].

Recently many studies related to game theoretic economic models were made and new related models were proposed [Jackson, 2010; Goyal, 2012]. Vallam et al. (2011) propose a game-theoretic model for network formation which involves localized payoffs and results in efficient pairwise stable networks. Recently evolutionary game theory has also found its applications in the dynamic analysis and formation of social networks. Evolutionary game theory on the dynamics of populations mainly involves the concept of a) Evolutionary stable strategy (ESS) and b) Analysis of the frequency of different strategies [Alexander, 2009]. A population is in an evolutionary stable state when all members adopt an ESS such that small occurrences of mutant strategies are not long lasting and vanish soon. Some studies along this direction include [Skyrms and Pemantle, 2000] which model the network dynamics in terms of the strength of link changes resulting from the repeated games played by the social actors. They also highlight that the nodes in dynamic networks tend to cluster and show cooperative strategies within them. Bala, and Goyal (2000) highlight that the mutual or one-sided incentives for a node to make and maintain links cause the network to quickly reach equilibrium. Although there exist in literature some attempts to model network formation using game theory, we argue that they are more centered on certain economic objectives. In terms of social network context, this line of work need more attention in the near future as the game theoretic formulation of a social network seems highly promising, naturally.

3.7 Adaptive Network Models

The network models discussed so far, including the preferential attachment models and the small-world models, are concerned with the *dynamics of networks* which deals with the topological dynamics of a network, i.e., the growth or change in the topology of a network over time according to simple local evolution rules which mimic the natural process of network formation. In addition to this line of work, the issue of *dynamics on networks* deals with the analysis of the state transition of nodes in a network (e.g., the transition of a person from infected to susceptible and vice versa in a contagion network, or increase in the volume of interaction between the connected individuals in a communication network). The basic assumption made here is that the network topology is static. The key ideas studies along this direction include the formation of opinions and diffusion of information which is discussed later in this paper (section 6).

It is, however, a notable fact that in real-world networks including social networks, the network states and topology co-evolve. We can say that the evolution of the network topology is linked to the network state and vice versa wherein there exists a feedback loop such that the characteristics of one network dimension defines the changes in the other. For example, the topology of a road network defines the flow of traffic while as the rate and nature of the flow of traffic defines the topological changes that need to be made in the road network (building new roads connecting key points and blocking some existing roads). Networks which exhibit such feedback loops in their nature are called *co-evolutionary* or *adaptive* networks [Gross and Blasius, 2008]. The study of adaptive networks is a young field and has only recently gained attention. Here we present a brief summary of some of the studies which aim to model the formation and growth of adaptive networks in nature and society.

One of the preliminary adaptive network models was proposed by Christiansen et al. (1998) which describes the evolution of an ecological network where nodes (populations) are linked through ecological interaction and the state of each node is defined in terms of its evolutionary fitness. The evolution of the network is modelled as the replacement of a node with a population of a new species having random fitness which in turn also affects the fitness of the neighboring nodes and the local topology. Similarly, Gross et al. (2006) model an adaptive network using a disease epidemic (Susceptible-Infected-Susceptible (SIS)) model wherein a susceptible node can isolate itself from an infected node by rewiring its links to other susceptible nodes which is shown to have a strong effect on the dynamics of the disease which in turn defines the rewiring process itself. Recently, Tero et al. (2010) proposed an adaptive network model inspired from the behavior of a single-celled amoeboid organism that forages for patchily distributed food sources and constructs a tubular network linking the discovered food sources. Their model emulates this network construction based on the feedback loops between the thickness of each tube and internal cytoplasmic flow. For more insight, the reader is referred to the original paper.

Adaptive network models seem to be the future hot topic for studying and analyzing the formation and growth of social networks especially the online social networks which present a unique opportunity to analyze both the dynamics *of* and *on* the networks. The adaptive online social network formation and growth models can exploit the feedback loops that exist between the states which include discussion topics, user opinions, application/game usage, group memberships and the network topology defined in terms of friendship links, interaction links or both. Moreover, despite the many advances in network modeling over the last decade, many unresolved issues still remain with big breakthroughs to be made in the areas of inference and dynamic modeling. Dynamic models are needed to test hypotheses related to network dynamics and model the significance of various constraints and metrics, which drive the dynamics, by defining and estimating parameters. For providing useful statistical inferences and insights, a network model needs to efficiently represent the dependencies between the network ties, and also between the behaviors of the actors. With the availability of the OSN datasets it is possible to create dynamic models that combine evolving topological and topic structures for a better understanding and modeling of the network evolution. Such hybrid models can also improve the predictability of links in both the static and dynamic settings of social networks. Evaluating and comparing the predictive ability of various models is also a consideration for future work along this direction. Moreover, most of the current network models tend to generate only undirected networks while as modeling directed network has received less attention. For more issues and challenges related to network modeling see [Snijders, 2011; Goldenberg et al., 2010].

4 Social Link Prediction

Social networks are used to represent the various social systems in nature and society consisting of individuals, objects and the relations, links that bind them together. An important property of most of the social systems and their resulting social networks is that they are dynamic. New individuals and objects are added to a system and some existing ones removed or lost with time. Similarly, new relations and links between individuals and objects are created and some existing ones broken with time. Social networks need to reflect these changes that occur in the underlying social system and in doing so they are often considered highly dynamic structures. Studying and understanding the mechanisms which result in a particular change in a social network is often appealing as it can throw some light on predicting the behavior of social network with time.

A related challenge motivated from these considerations is the *link prediction problem*. The aim of link prediction is to predict the edges (relations, links) that can be induced in a social network S at a future time t' by a particular known state of S at a given time t . The applications of link prediction range from online recommendations to the detection of links between objects in a criminal case, from predicting possible interacting proteins or proteins with similar functionality to predicting possible future collaborations between researchers, from predicting friendship relations to predicting hyperlinks between web pages and so on. A natural approach for predicting links would

be to use *node-wise similarity*, i.e., to determine distance or similarity between two objects using the available node features. For example, social links between two individuals in a social network can be reliably predicted simply based on whether they are alike, i.e., based on the concept of *homophily*. Supervised learning techniques like decision-tree classifiers, SVM and so on can be trained on one subset of the edges and then used to predict links in its complement. Such a classifier can make good predictions for a set of unseen users or for future connections between known users [Hasan et al., 2006; Wang et al., 2011]. However, such an approach completely ignores the rich source of information that is the graph structure of the link graph. In contrast, according to Liben-Nowell and Kleinberg (2007) the link prediction problem asks *to what extent can the evolution of a social network be modeled using features intrinsic to the network itself*. As a result, most of the link prediction approaches are based on learning *topological patterns* from node-centric (local) properties or the overall network properties and the decision on the existence of a link is made based on the visible patterns in the given state of the network. Liben-Nowell and Kleinberg (2007) conducted a comprehensive study on some related link prediction methods which use the network topology to estimate a similarity-score between two nodes in a network and a list of resulting node pairs with decreasing similarity-scores. This list is then used to make a decision on the existence of a link. They have classified the link prediction methods into those based on *node neighborhoods*, *ensemble of all paths* and other high-level approaches. Based on their comparative analysis, they determine that no method is better than any other, instead all the methods perform well but many of methods outperform the random predictor, signifying that network topology itself contains important indications regarding the prediction of missing or future links. However, an important limitation in these techniques is that they are mainly based on the topological features of networks.

Intuitively, it seems that better performance can be achieved by using additional sources of information besides the topological information, such as the content or semantic attributes of the nodes. Along this direction Hasan et al. (2006) and O'Madadhani et al.(2005) incorporate features from multiple sources and train a classifier to decide if a link establishes or not. Hasan et al. (2006) use heterogeneous features including shortest distance, common neighbors and number of matching keywords of nodes to analyze a co-author dataset. On one of the datasets their analysis reveals that semantic features, which include keyword-match count, of the publications result in more gains in estimating the similarity. Similar results have been reported by O'Madadhani et al. (2005) on the use of content-based attributes, geographic proximity between authors, and similarity of journal publication patterns. Other approaches for link prediction include methods based on *probabilistic models* which try to incorporate multiple data elements from the network to learn a compact model whose output probability is then used for prediction. For example, Taskar et al. (2003) defined a joint probabilistic model over the entire graph which also includes the content attributes of the nodes besides overall link structure using discriminatively trained relational Markov networks and use the trained model to collectively classify the test data. Based on the assumption that the network structure is in a stationary state, Kashima and Abe (2006) proposed a parameterized probabilistic model of network evolution whose parameters are estimated using an EM algorithm and then used for link prediction. Although such model based approaches are powerful but usually computationally expensive which requires appropriate approximations to guarantee efficiency. Recently a comprehensive survey on such related methods was presented by Hasan and Zaki (2011) wherein they categorize link prediction methods into probabilistic relational models, linear algebraic models using rank-reduced similarity matrices, and Bayesian probabilistic models.

Murata and Moriyasu (2008) use weighted proximity measures of social networks for the task of link prediction based on the assumption that node proximity can be determined effectively by using both graph proximity measures and the link weights. Their approach was the first to take link weights in to consideration for the task of link prediction. They proposed three weighted similarity indices, as variants of the *common neighbors*, *adamic-adar* and *preferential attachment* indices, and their experimental results show that considering these weighted indices increases the performance of link prediction, especially in highly dense networks. However, when Lü and Zhou (2010) applied the weighted indices to the co-authorship network and to the US air transportation network, they found that the weighted indices performed even worse than the un-weighted ones, reminding of the *weak-*

ties theory [Granovetter, 1973], which claims that the links with small weights yet play a more important role in social networks. They suggest from their experiments that the weak ties play a significant role in the link prediction and the contributions of weak ties can remarkably enhance the prediction accuracy for some networks. Similarly, in [Liu and Lü, 2010], considering node similarity for link prediction, the authors have proposed a similarity index based on local random walk, which has lower computational complexity compared with other random-walk-based similarity indices with good or even better prediction. Zheleva et al. (2010) argue that for networks in which group structures exist and are known, the link prediction task can yield efficient results. They show how predictive models based on descriptive, structural and community features perform surprisingly well on challenging link-prediction tasks by overlaying friendship and family networks and using the features of the overlaid networks to accurately predict friendship relationships. Schifanella et al. (2010) show that the tagging activity of users reflects their group-participation and degree-centrality in the social network. Moreover, the activity-rates of users reveal a strong assortative mixing in the social network. Using the tagging activity of users from Flickr and Last.fm, they show that close neighbors tend to show more lexically-similar tagging activity around common topics and this feature can be exploited to predict links. They test this hypothesis on a Last.fm data set, and show that the links predicted between users based on high semantic similarity closely map to the underlying friendship links of users more accurately than Last.fm's suggestions based on listening patterns. Leroy et al. (2010) define *cold start link prediction* as a link prediction challenge wherein the primary source of links, i.e., the network structure is not known and only a secondary source of node related information is known. Unlike traditional link prediction methods which require a known state of a network, their method does not require an initial state of the network to predict the possible missing or future. More specifically, they assume that either the social network explicitly exists, but is kept secret by its owner, or it does not exist at all. They propose a two-phase method based on a *bootstrap probabilistic graph* for cold start link prediction. In the first phase, based on some limited information (potentially, weakly correlated with the link structure of the network), the method predicts the existence of links resulting in a probabilistic graph, i.e., a graph where each edge is labeled with a probability representing the confidence of the prediction, or in other terms, the uncertainty of the existence of a link. The second phase takes as input the probabilistic graph and refines it by adopting graph-theoretic measures given in [Liben-Nowell and Kleinberg, 2007] as used in the classical link prediction settings.

Although, link prediction is not a new problem in information science, traditional methods have not caught up with the new development of network science especially the new perspectives and tools resulted from the studies of complex networks. For example, the hierarchical and modular structure of social networks (section 5), could efficiently guide the link prediction task and in turn throw light on the community wise evolution of social networks. A major challenge related to link prediction is the heterogeneity found in most of the real world social networks. For example, online social networks usually involve different node types like, users, images, URLs, tags and so on. Similarly, links may also involve attributes representing polarity (positive or negative), friend and foe relations and so on. Furthermore, link prediction in weighted and directed networks along with predicting link weights also requires significant attention. The cold start link prediction problem which deals with predicting links that could be induced into social network when a new node is added is another big challenge that has received less attention till now.

5 Community Analysis

The description of the structure of complex networks is often studied at different levels ranging from the microscopic characteristics of individual nodes (degree, centrality and so on) to the macroscopic description in terms of statistical properties of the whole network (degree distribution, total clustering coefficient and so on). In between these two levels a “mesoscopic” description tries to explain the community structure in complex networks. Communities are considered to be the sets of nodes in a network that have denser connectivity to each other than to the rest of the network (see Fig. 1) and are important because they can often be closely related to functional units of a system, e.g., groups of individuals interacting with each other in a society [Girvan and Newman 2002; Arenas

et al., 2004], World Wide Web (WWW) pages related to similar topics [Flake et al., 2002], and compartments in food webs [Krause et al., 2003]. The basic task in the analysis of communities is *community detection*, which has received a lot of attention in the recent past, and the field is still rapidly evolving [Fortunato and Castellano, 2007].

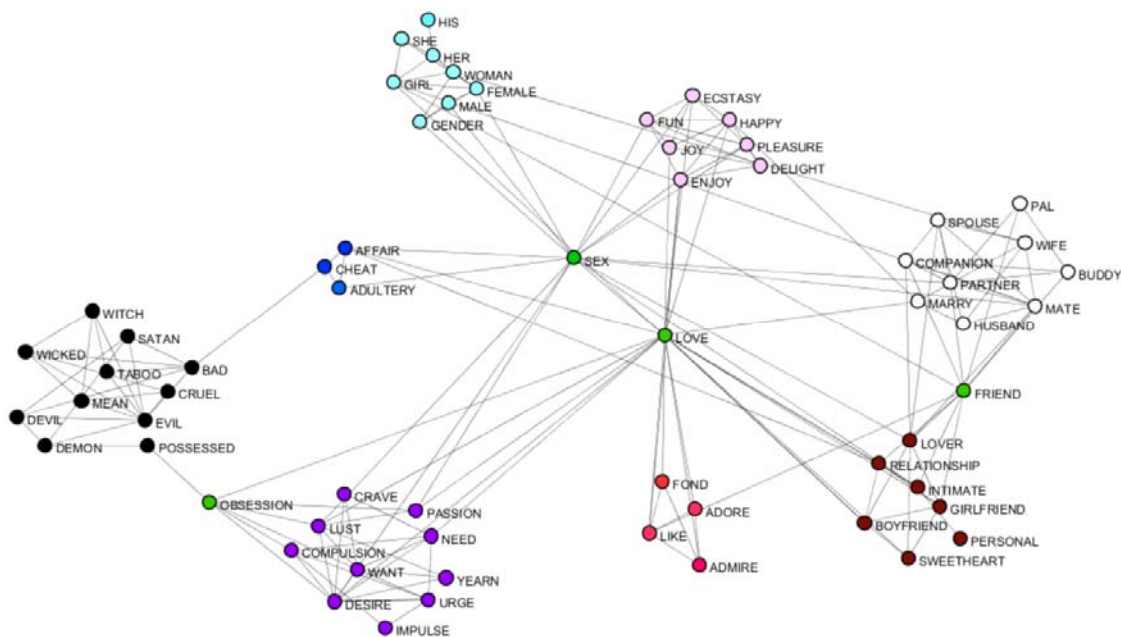


Fig. 1. Community Structure identified in a word association network by the method proposed in [Bhat and Abulaish, 2012]

Detecting communities in a network depends on various factors like, whether the definition of community relies on global or local network properties, whether nodes can simultaneously belong to several communities, whether the link weights are utilized, and whether the definition allows for hierarchical community structure. An open challenge related to community detection is to cope up with overlapping communities that occur when a particular node in a network belongs simultaneously to several communities. Some of the approaches to deal with overlapping communities have been proposed in [Lancichinetti et al., 2009; Palla et al., 2005]. Another challenge related to community detection is due to presence of networks containing hierarchical structures. In such networks, a community may be part of even larger communities. Newman and Girvan (2004) have worked in this direction and introduced the concept of modularity as a measure for the goodness of a partitioning as sometimes it is better to study community structure considering nested hierarchy rather than choosing a single community partitioning [Clauset et al., 2008]. Furthermore, real world social networks tend to change dynamically, for example, in online social networks, each day new users join the network and new connections occur between existing members, while some existing ones leave or become dormant. For analyzing such communities it is desirable to understand the evolution and dynamics of the community structure. We provide a more detailed description on different aspects of community analysis in the following sub-sections.

5.1 Methods and Approaches for Community Detection

At a higher level of abstraction, the problem of community detection can be divided into *local community mining* and *global community mining*. In local community mining the problem is to determine a local community C from a network for which the complete connectivity information of the nodes is unknown and we are given only a partial set of nodes forming C and their connectivity information. The community structure of C can be revealed by recursively crawling the neighboring nodes of the given partial set of nodes and then based on the connectivity of the traversed nodes, determine the *core* and the *boundary* of the community C . In global community mining, which is based on the assumption that whole information regarding the node connectivity of a network is known, the problem is to identify all the communities represented by their corresponding nodes

that are present in the network [Chen, 2010]. In the field of community mining most of the works are oriented towards identification of global communities.

$$x_{i,j} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2} \quad (7)$$

The principal and the most popular technique used by sociologists in their study of social networks for finding communities is *hierarchical clustering* [Hastie et al., 2001; Scott, 2000]. Hierarchical clustering methods involve discovering natural partitions of social networks, using various metrics of similarity, e.g., *Euclidean distance* [Wasserman and Faust, 1994] defined in equation 7 compares the neighbors that two vertices share. In this equation x is the similarity measure and A is the adjacency matrix. Hierarchical clustering techniques do not provide a way to decide upon a proper cut in the generated dendrogram, i.e., to choose among the created partitions a more appropriate community structure. It should also be noted that the accuracy of results of a community detection method also depend upon the particular similarity measure used. Based on the approach of generating clusters, hierarchical clustering methods can be classified into two categories – *agglomerative* and *divisive* [Hastie et al., 2001]. Agglomerative methods are based on a bottom up approach that involves iteratively merging clusters if their similarity is sufficiently high. After defining a measure for vertex similarity, each node is assigned to its own community forming n communities for n nodes. Now, an edge is added one at a time for the pair of nodes which shows the highest similarity. This approach often finds only the strongly connected cores of communities and fails to identify the less densely connected boundary nodes for the communities. An example of agglomerative method is [Clauset et al., 2004], which is based on *modularity optimization* [Newman and Girvan, 2004] and starts with a state in which each vertex is the sole member of one of n communities. Communities are repeatedly joined together in pairs, choosing at each step the join that results in the greatest increase (or smallest decrease) in modularity. This method can be applied to very large networks. In contrast to agglomerative methods, *divisive methods* follow top-down approach wherein all the nodes are initially assigned to a single large community which is iteratively divided into smaller communities by removing edges with low similarity values and thus causing the split in a large community. Divisive methods usually classify all of the vertices (including peripheral nodes, outliers) into communities, which may lower the accuracy of community detection. Based on the sociological notion of *betweenness centrality* [Wasserman and Faust, 1994], Girvan and Newman (2002) proposed a community finding algorithm based on divisive clustering approach, which progressively removes edges from the network. The algorithm calculates the betweenness of all edges in the network and removes the one with the highest betweenness. This process is repeated until no edge remains or a stopping criterion is met.

Besides the above deterministic methods for community detection, other methods for the detection of groups from networks have also been based on the concept of *stochastic block-modeling* [Nowicki and Snijders, 2001]. In SNA, Block-modeling is commonly used to divide the vertices in a network into categories or classes wherein nodes assigned to the same category or class share some common properties or show some degree of equivalence. The equivalence defined within a class is mainly based on the topological properties of nodes such as *structural equivalence* [Lorrain and White, 1971] and *regular equivalence* [Everett and Borgatti, 1994]. Structural equivalence assigns two vertices to the same category or class if they have all the neighbors common or at least show a higher degree of overlap between the set of their respective neighbors. On the other hand regular equivalence defines two vertices to be similar if they show similar connections properties with vertices of some other categories. In stochastic block-modeling, objects are assigned positions defined in terms of IID (independent and identically distributed) random variables, and a particular type of link between two objects is in turn defined as another random variable which depends only on the positions of the object pair it links. Extending the general stochastic block-modeling approach of Nowicki and Snijders (2001) that uses Gibbs sampling to infer the object positions, Wolfe and Jensen (2004) allow an object to attain multiple position categories so as to model the multiple roles which an object may possess in different contexts. Wang et al. (2005) and Wang et al. (2006) propose a generalized stochastic block-modeling approach which allows detecting groups of individuals whose activity is centered around certain topics based on the relations between

individuals and their respective demographic properties. There also exist numerous other methods and techniques for community detection which include methods based on maximum likelihood [Clauset et al., 2008], mathematical programming [Agarwal and Kempe, 2008], inference [Hastings, 2006], and latent space clustering [Handcock et al., 2007].

Extending the DBSCAN algorithm [Ester et al., 1996] to undirected and unweighted graph structures, Xu et al. (2007) proposed SCAN (Structural Clustering Algorithm for Networks) to find clusters, hubs, and outliers in large networks based on structural similarity, which uses the neighborhood of vertices as clustering criteria. Similarly, considering only the weighted interaction graph of the online social networks, Falkowski et al. (2007) extended the DBSCAN algorithm to weighted interaction graph structures of online social networks. The basic idea of density-based clustering methods is that if the neighborhood of a given radius ε of a point p contains more than μ objects, then a new cluster with p as a core object is created. The process then iterates to find *density-reachable* objects from these core objects and defines a *density-connected* cluster using *density-connectivity* relations between nodes [Ester et al., 1996]. Some important features of density-based community detection methods include less computation, detection of outliers and natural scalability to large networks. However, the main drawback of traditional density-based community detection methods is that they require the global neighborhood threshold ε and the minimum cluster size μ to be specified by the users. The methods are particularly sensitive to the parameter ε , which is difficult to determine prior. Actually, how to determine the optimal value for parameter ε automatically for the density-based clustering methods (e.g., DBSCAN and SCAN) is a longstanding and challenging task [Sun et al., 2010]. The method proposed by Sun et al. (2010) reduces the number of possible values to consider for ε significantly by considering only the edge weights of a Core- Connected Maximal Spanning Tree (CCMST) of the underlying network. In order to find an optimal value for ε from the remaining domain, they use modularity as a quality function to automatically select the value for ε which yields the community structure resulting best modularity. Similarly, Huang et al. (2010) proposed a two stage parameter free extension of density-based clustering by first finding smaller communities using the highest local structural similarity value of ε for a pair of nodes and a constant value for μ , and then iteratively optimizing the modularity measure [Newman and Girvan, 2004] upon joining these smaller communities. In the first stage it uses a density-based approach to detect micro-communities by considering dense pairs (i.e., pairs of nodes whose structural similarity is largest among their adjacent neighbor nodes). In the second stage, it iteratively joins the micro-communities by considering the gain in modularity.

The area of community detection has grown enormously in the recent years in terms of the number of community detection algorithms proposed. However, it is difficult to find any consensus in the definition of a community used or the application areas proposed, among most of them. A main issue that still needs to be explored is that of the validation of the community structure found by these algorithms. This calls for defining benchmark graphs with known community structure for testing the algorithms. Although various methods for creating synthetic graphs with known community structures exist, this process does not guarantee the performance on real networks.

5.2 Community Evaluation and Modularity Optimization

We have seen many approaches for community detection but one of the main challenges that most of them have to face is that in real world they need to identify community structures from networks for which the actual underlying community structure is hidden and there is no ground truth for the mining problem with which the quality of the detected community structure could be compared. Even though a community mining algorithm demarcates communities in a network, we need to answer a challenging question as to how do we determine if the identified community structure significantly maps to the actual hidden community structure of the underlying network?. Community mining algorithms are often seen to identify a community structure even in random networks which are expected to have no significant community structure at all, so how can we measure the structure that is found for these *structureless* networks?. How can we compare between different community results to find the *best* ones for a given network [Chen, 2010]?

As a basic solution to the above mentioned problems, various objective functions have been proposed with an aim of finding optimal solutions based on different criteria. Some of the objective functions that have been used so far include *ratio cut minimization* and *normalized cut minimization*. Ratio cut minimization [Wei and Cheng, 1989] involves minimizing the fraction of all possible edges leaving a cluster, whereas the normalized cut minimization [Shi and Malik, 2000] seeks to minimize the cut relative to the number of edges in a cluster instead of its size. Given an undirected graph $G=(V, E)$, let S be a subgraph of G representing a cluster with the number of nodes $n_s = |S|$ and the number of edges $m_s = |\{(u, v): u \in S \wedge v \in S\}|$. Let $d(S)$ and $d(G/S)$ be the sum of the degrees of all nodes of S and of the rest of the graph G/S respectively, $c_s = |\{(u, v): u \in S, v \notin S\}|$ is the number of edges on the boundary of S (cut size of S). The *ratio cut* and *normalized cut* objective function can be expressed by equations 8 and 9, respectively. Both the ratio cut and the normalized cut minimization find clusters of almost similar size, i.e., the number of nodes and/or edges in the resulting communities is almost the same.

$$R(S) = \frac{c_s}{n_s(n - n_s)} \quad (8)$$

$$N(S) = \frac{c_s}{d(S)} \quad (9)$$

Finding the minimum ratio cut of a graph by checking every possible collection of clusters is computationally prohibitive. However, many methods have been proposed to find an approximation to the minimum ratio cut over the whole graph. Another solution for evaluating community structure is defined as a measure called *Modularity* represented as Q introduced by Girvan and Newman (2004) originally to define a stopping criterion for one of their algorithms in order to choose the best community structure from a hierarchy of communities. Based on the intrinsic link structure of a network, modularity is a measure of goodness of a given partition of the network into communities [Newman, 2004; Newman, 2006a; Newman, 2006b; Newman and Girvan, 2003; Newman and Girvan 2004]. The idea of modularity Q is to compare the number of links inside communities to the expected number of links in a random reference network which contains no community structure. More precisely, the modularity Q uses equation 10 in which l_{mm} is the number of links inside community m , L is the number of links in the entire network, K_m is the sum of degrees of nodes comprising community m , and the sum is over all communities. The term $K_m^2/2L$ corresponds to the expected number of links inside the community m for a randomized graph of the same size and same degree sequence as the original network. The total number of communities is given by represented by q .

$$Q = \sum_{m=1}^q (l_{mm} - \frac{K_m^2}{2L}) \quad (10)$$

Modularity exploits the differences in node degrees as it aims to find the difference between the fraction of edges which exist between the nodes within a community, and the fraction of edges which is expected to exist between these nodes if the edges are assigned at random based on the node degrees in the underlying network. In this way, higher values of modularity for a partition scheme represent communities for which more edges of the nodes in the community occur within the community than expected by a random assignment of edges between the nodes. The maximum value of Q is 1 and it can also obtain negative values, which corresponds to assigning nodes into communities such that communities are sparser than the random reference. It can be assumed that high values of modularity indicate good partitions. In particular, Girvan and Newman (2004) suggested that the partition of a network which maximizes modularity Q is the best representation of the community structure of the network. Since Q is independent of the method of obtaining the communities, for many later community detection methods it became the objective function to be maximized leading to modularity optimization based methods for community detection. It should be noted that an exhaustive optimization of Q is impossible as there exist a large number of ways in which a graph can be portioned (even if the graph is small). Furthermore, modularity optimization is

an NP-complete problem [Brandes et al., 2006], which indicates that a polynomial time solution with respect to the size of the graph is not feasible. In this regard several algorithms have been proposed which aim to reach approximately close values of maximum modularity in the least time possible. One of the first algorithms for modularity maximization is the greedy method proposed by Newman (2004). Newman’s method follows an agglomerative approach wherein at each level smaller groups are merged to form larger groups only when there is a gain in modularity. A similar approach is proposed by Blondel et al. (2008) which also takes the link weights into consideration and involves computing the gain in weighted modularity on merging. Other approaches for modularity optimization include methods based on simulated annealing [Kirkpatrick et al., 1983] like [Guimerà and Amaral, 2005], extremal optimization [Boettcher and Percus, 2001] like [Duch and Arenas, 2005], and spectral methods [Newman, 2006a; Newman, 2006b] which optimize modularity by considering two partitions at a time using spectral bisection. Instead of using the Laplacian matrix, they use a *modularity matrix* whose elements are determined by equation 11 in which k_i and k_j are degrees of the nodes i and j respectively and m is the total number of nodes in the graph and A is the adjacency matrix.

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \quad (11)$$

Although modularity optimization methods have proved highly effective in practice for community evaluation [Danon et al., 2005], there are three major problems for the Q measure. Firstly, modularity requires information about the entire structure of the graph, which is unrealistic in case of large networks like the WWW. As a solution to this problem, Clauset (2005) have proposed a measure of local community structure, called *local modularity*, for graphs which lack global knowledge. Similarly, Radicchi et al. (2004) proposed a divisive hierarchical method, where links are iteratively removed based on the value of their edge clustering coefficient, which is a local measure. This approach involves less computation than that of edge betweenness used in [Girvan and Newman, 2002] and thus yields a significant improvement in the complexity of the algorithm. Moreover, the stopping criterion of the procedure depends on the properties of the communities themselves and not on the values of a quality function like modularity. Secondly, modularity-based methods have a resolution limit and may fail to identify smaller (possibly important) communities [Fortunato and Barthélemy, 2007]. Possible solutions include recursive algorithms based on modularity optimization [Ruan and Zhang, 2008]. Finally, as pointed out by Scripps et al. (2007), modularity only measures existing links on the network, but does not explicitly consider the absent links between two nodes in the same community. In this regard, it seems better to optimize a quality function which takes into consideration both the topological structure of the networks and some secondary information related to node properties or their temporal [Shalizi et al., 2007]. How and what secondary information related to nodes, edges or the network as a whole can be used along with the network topology for the community detection task is among the challenging issues related to community analysis in social networks [Newman, 2008; Traud et al., 2008].

A important issue related to community analysis is the interpretation of the detected community structure in networks, i.e., what does a detected community structure in a particular network represent or what do we do with the detected communities. While most of the algorithms are based on the structural information of the social networks, it is still difficult to say that structural communities closely map to the underlying functional communities in most of the real world social networks. Alternatively, it seems promising to explore new techniques and methods which tend to incorporate any available secondary information related to the nodes, edges or the network itself, e.g., demographic information of nodes, edge weights or textual content interactions and so on along with the network’s primary topological characteristics. In case of online social networks it is possible to use the content information available in the form of comments, messages, tags and so on to define topic hierarchies and the orientation of group opinions about them to guide the clustering process.

5.3 Overlapping and Hierarchical Community Detection

One of the challenges in community detection as observed by Zhang et al. (2007) is that the communities may show overlapping behavior, i.e., some nodes may show affinity towards multiple communities. The membership of an entity in many groups is very common in real world networks. For example, in a social network, a person may participate in many interest groups, see Fig. 2. Identifying such overlapping communities from social networks has gained significant attention recently. In this section, we introduce some of the existing techniques to detect overlapping communities. As mentioned by Chen (2010), a natural way to discover overlapping communities is to first globally partition the network and then locally expand the discovered communities to locate overlapping components. For example, Wei et al. (2008) first partition the network into seed groups of overlapping community structure using existing spectral clustering methods. A locally optimal expansion process is then applied to greedily optimize Newman's Modularity Q measure. Similarly, Baumes et al. (2005) initialize community cores using Link Aggregate (LA) algorithm and then refine the peripheries by an Iterative Scan (IS) procedure.

The most popular method for identifying overlapping communities is the *Clique Percolation Method* (CPM) proposed by Palla et al. (2005), which is based on the concept of a k -clique. A k -clique is a subgraph of k nodes. The method relies on the observation that communities seem to consist of several small cliques that share many nodes with other cliques in the same community. A k -clique community is defined as the largest connected subgraph obtained by taking the union of a k -clique with all k -cliques which are adjacent (two k -cliques sharing $k - 1$ nodes where k is a given parameter representing the clique size) to it. Rewiring one end of some links in a k -clique to some other node of an adjacent k -clique (called *rolling*) can also be used to identify a k -clique community as shown in [Derényi et al., 2005]. The choice of k has a significant effect on the found community structure. Typically used values of k are between 3 and 6 and the values of 2 and 1 map to bond and node percolation respectively. High values of k yield tight, internally cohesive communities, whereas small values of k yield sparse and larger communities. The intermediate k -cliques are allowed to share nodes between them and thus the resulting communities can show overlaps at common nodes. Kumpula et al. (2008) have proposed an enhanced variation of the CPM called the Sequential Clique Percolation (SCP) algorithm. SCP involves adding edges to an empty network in a sequential manner. On the addition of each new edge, the formation of any new k -cliques is checked by searching for $(k - 2)$ -cliques in local neighborhoods of both the nodes at the ends of the newly added edge. The core of the SCP algorithm consists of constructing k -clique communities continuously when new k -cliques appear. For each new k -clique there are two possible cases – the k -clique can either form its own community or it can overlap with one or more existing communities. In the latter case, all overlapping communities merge to form a single community.

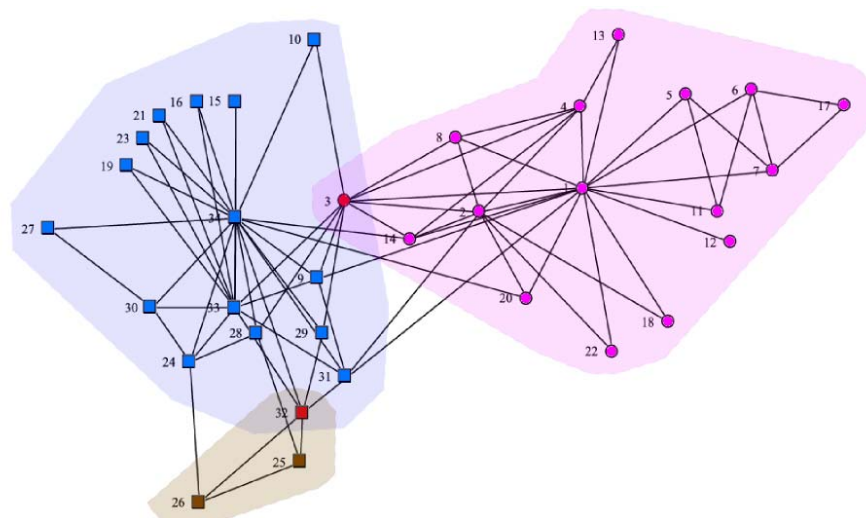


Fig. 2. Overlapping community structure found in the Zachary's network [Zachary, 1977] by the method proposed in [Bhat and Abulaish, 2012]

Zhang et al., (2007) propose a hybrid method by embedding the vertices of an arbitrary graph into a d -dimensional space using spectral mapping in order to utilize the fuzzy c-means algorithm on graphs through optimizing a quality function. However, the eigenvector calculations involved in the algorithm render it computationally expensive to use on large networks. Wei et al. (2008) first partition the network into seed groups of overlapping community structures using existing spectral clustering method. A locally optimal expansion process is then applied to greedily optimize Newman's modularity measure. McDaid and Hurley (2010) presented an overlapping community detection method MOSES by combining local optimization with Overlapping Stochastic Block Modeling [Latouche et al., 2009] using a greedy maximization strategy. Here communities are created and deleted, and nodes are added or removed from communities, in a manner that maximizes a likelihood objective function. For an in-depth introduction to other overlapped community detection methods readers are advised to see [Fortunato, 2010; Chen, 2010].

Besides overlapping community structures, networks often contain significantly different community structures at different levels of granularity wherein smaller communities are embedded within some larger communities, see Fig. 3, i.e., there exists a community hierarchy within the network [Simon, 1962]. In order to provide appropriate information about the modular structure of a network, it is desirable to detect overlapping communities along with their hierarchical organization. In [Lancichinetti et al., 2009], the authors have proposed a method for simultaneously uncovering both the hierarchical and the overlapping community structure from networks. The method is based on locally optimizing a fitness function for each node to decide upon the assignment of a node to a particular community. A node is allowed to show affinity to multiple communities and thus can be assigned to multiple communities resulting in a possible overlapping community structure for the underlying network. Moreover, the method uses a resolution parameter whose value defines the average size of the communities to be detected and thus varying this parameter between optimal higher and lower allows exploring the hierarchical levels of the community structure for the underlying network.

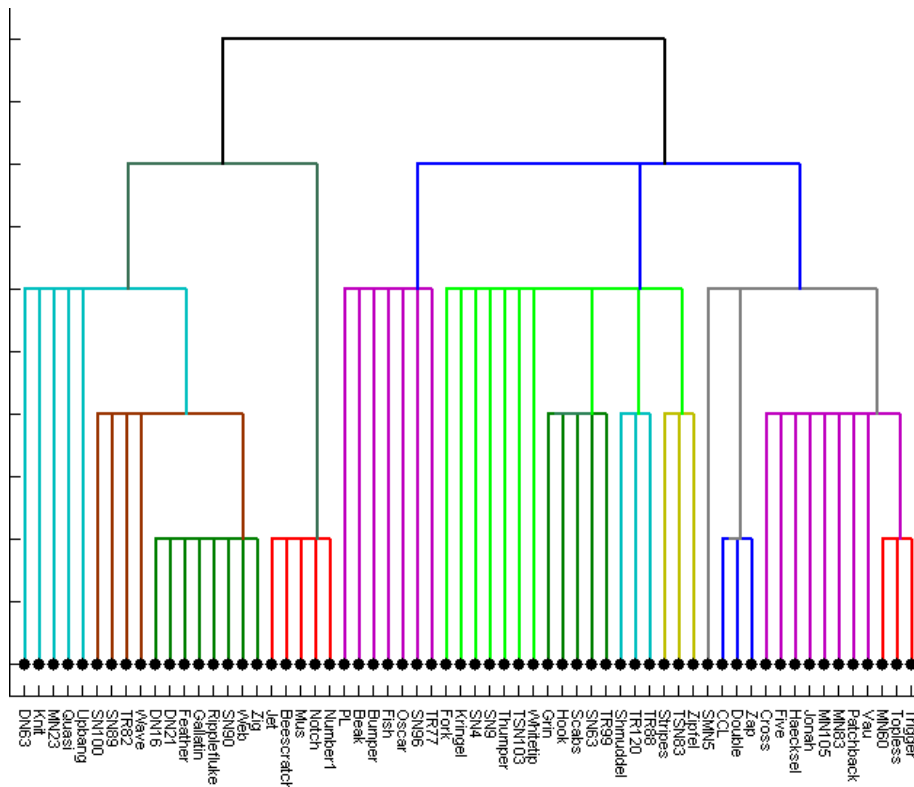


Fig. 3. Overlapping and hierarchical community structure identified in a Dolphin network [Lusseau et al., 2003] by the method proposed in [Bhat and Abulaish, 2012]

Reichardt and Bornholdt (2006) show that there exists an analogy between detecting communities in networks and finding the ground state of a spin glass model in the sense that the energy of the spin system is equivalent to the quality function of the clustering with the spin states being the group indices. In this regard edges should exist between nodes which have the same spin state, and cease to exist between the ones with different spin states. A single parameter relates the weight given to missing and existing links in the quality function and allows for an assessment of overlapping and hierarchical community structures. In line with CPM, Kumar et al. (2009) propose a method (HOC) to identify hierarchical and overlapping communities by finding maximal cliques from the underlying network. However, unlike CPM, HOC uses the overlapping neighborhood criteria to define the similarity between two arbitrary nodes in a network. For HOC, if two nodes have the overlapping neighborhood ratio greater than a threshold, they belong to the same community. The community detection framework, *Infomap*, presented by Rosvall and Bergstrom (2008), reformulates community detection as minimizing the description length of a random walk across the network. The total description length consists of the length for encoding community transitions and the length for encoding movements within communities. Infomap considers smaller description for the trajectory of random walk to be more reasonable for defining a community partition. Rosvall and Bergstrom (2011) extend Infomap to find hierarchical community structures from networks. Recently, Lancichinetti et al. (2011) presented OSLOM which locally optimizes the statistical significance of clusters defined with respect to a random graph generated by the configuration model during community expansion. OSLOM is able to detect a hierarchical community structure by reapplying the algorithm on intermediate supernetworks of detected communities. The methods proposed by Lancichinetti et al. (2009), Kumar et al. (2009) and Reichardt and Bornholdt (2006) provide a tunable parameter (resolution parameter) whose value determines the size of the detected communities. This allows visualizing the community structure at different resolutions and thus forms a community hierarchy. These methods are called multi-resolution methods and some other multiresolution methods include [Pons, 2006; Arenas et al., 2008]. See [Fortunato, 2010] for a description of these methods.

A two stage algorithm proposed by [Shen et al., 2009] for detecting overlapping and hierarchical community structures in a network involves identifying all maximal cliques in the network, which along with each subordinate vertex (single vertices that do not belong to any clique) are taken as an initial set of communities. A dendrogram is then created in an iterative way using an agglomerative approach. In second phase, a proper cut point for the dendrogram is determined by finding the maximal value of an extended modularity measure which also considers the number of communities to which a node belongs to. Instead of assigning nodes to communities, Ahn et al. (2010), assume that *links*, rather than nodes, are characterized by a single attribute (such as the community assignment). They study hierarchical organization of overlapping communities following a link clustering approach.

5.4 Dynamic Community Detection and Community Evolution

As mentioned earlier, one of the important properties of the real world social networks is that they tend to change dynamically and this property has until recently been largely ignored in terms of community detection. Recently, several datasets and methods of recording the network dynamics have become available, enabling to monitor and analyze the evolution of real-world networks [Kumar et al. 2006; Leskovec et al., 2008] which also makes it possible to analyze the evolutionary changes related to communities. The main evolutionary characteristic events related to the lifetime of a community include *birth* (a new community emerges), *growth* (an existing community gains more nodes), *shrinkage* (an existing community loses some member nodes), *merging* (more than one existing nodes join to form a new community), *split* (an existing community breaks into more than one community), and *death* (an existing community ceases to exist by losing all member nodes or at least all of its core nodes).

The analysis of the evolution of communities in large social networks, such as membership, growth and disbandment, is becoming increasingly prominent especially for social networking sites such as MySpace and Facebook. By monitoring friendship links and community membership on a social network site, and co-authorship and conference publication in a bibliography database, Backstrom

et al. (2006) study the relation between the evolution of the community structure and the topological structure of the underlying network. Kairam et al. (2012) investigate about the life and death of online groups based on various group based network features. They find that communities possessing densely connected hierarchical sub-groups and a set of loosely connected nodes tend show long life and higher rate of membership gain. Although, groups increase the diffusion rate, groups formed through the diffusion process attain smaller sizes with the passage of time. Moreover, their analysis suggests that historical growth features can help in closely estimating the growth in near future but the structural features of the network tend to perform better in estimating the growth to a distant future (at least for small groups).

In order to identify community structure from networks, most of the approaches either consider a single snapshot at any time or all the nodes and edges present within a particular time window. But in case of dynamic networks such approaches tend to miss the important behavior of communities, i.e., their evolution with time, which represents one of the most important properties of networks and communities to be observed [Tantipathananandh et al., 2007]. The typical dynamic community detection problem as formulated by [Backstrom et al., 2006, Tantipathananandh et al., 2007] is to observe the social interactions of a subset of individuals of a network at each time step along a discrete time scale. Consequently, several subgraphs are formed which reveal the underlying community structure and the changes that occur to it over time. However, the bipartite mapping of communities for two subgraphs in these methods assumed a zero to one or one to one mapping between the communities and hence do not identify merge or split events. Alternatively, Greene et al., (2010) propose a heuristic threshold-based method allowing many-to-many mapping between communities across different time-steps, thus also enabling the detection of merge and split events. This approach is independent of the choice of the underlying static community finding algorithm applied to the individual step graphs. To perform the mapping between the communities at two time-steps, they define the similarity between two communities based on the jaccard coefficient given by equation 12, wherein C_i and C_j are two communities at two consecutive time steps respectively.

$$sim(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \quad (12)$$

If the similarity (or overlap) for two communities C_i and C_j is found greater than a certain threshold (between 0 and 1), an evolutionary relation is established between them. Similar approach is followed in CHRONICLE [Kim and Han, 2009] is a two stage extension of SCAN [Xu et al., 2007] for identifying community evolution in dynamic networks but it does not consider the edge weights (often a important supplement for community detection) when available.

Wang et al. (2008b) have presented a core-based algorithm for tracking community evolution which depends on core nodes to establish the evolving relationships among communities at different snapshots. Instead of overlapping level of nodes or edges between two communities, their algorithm heavily relies on core nodes that are distinguished from ordinary nodes on the basis of both the community topology and the node weight as follows. Other methods introduced for the analysis of communities and their temporal evolution in *dynamic networks* include [Asur et al., 2009; Kumar et al., 2006; Lin et al., 2007; Palla et al., 2007; Spiliopoulou et al., 2006]. However, as pointed out by [Lin et al., 2009], a common weakness in these studies is that first the communities are identified separately and then a mapping across communities to identify their evolutionary characteristics is performed. An alternative approach would require that the identified community structure itself provides information regarding its evolutionary events and a historical event log helps in deciding upon an appropriate community structure. Lin et al. (2009) analyze the evolution of communities by first identifying an initial community structure using a stochastic block model and then adapting this initial community structure by following a probabilistic model for capturing the community evolution. For each time-step graph of a dynamic network, the community structure is determined by considering both by the current network state and the historic community evolution patterns. However, the main limitations of their method include the need for specifying number of communities a priori and low scalability due to large number of matrix computations required.

Falkowski et al. (2008) propose a framework for studying community dynamics where a preliminary community structure adapts to dynamic changes in a social network. A similar approach is proposed by Bhat and Abulaish (2012), but unlike Falkowski et al. (2008), their concern is on tracking the evolution of overlapping communities and does not need an ageing function to remove old interactions from the network. Their distance function is based on average reciprocated interactions in node-neighborhoods. Moreover, the proposed framework is applicable to directed/undirected and weighted/un-weighted networks, whereas the method of Falkowski et al. (2008) applies only to un-directed and weighted networks. For un-weighted networks, the proposed framework by Bhat and Abulaish (2012) assigns a unit weight to each edge in the network without altering the meaning or representation of the network. Other related dynamic community models and detection methods for dynamic networks are studied in [Spiliopoulou, 2011]. Studying the evolution of discussion topics can help in identifying the evolving community structures in dynamic social networks which is another challenge that need more work to be done. Studying overlapping and hierarchical community structures in a dynamic perspective in evolving social networks may also throw some light on how communities merge or split, and the role of nodes with multiple memberships (overlapping nodes) in this context. Moreover, analyzing communities in weighted and directed networks also needs some more attention.

6 Diffusion and Influence in Social Networks

With the increasing popularity of online social networks that form an efficient means of analyzing information diffusion there has been a good amount of research regarding the analysis and modeling of information flow in these networks. Diffusion in networks has been studied across a wide range of real-world scenarios with an aim to find trends, patterns and models that represent the spread, propagation or diffusion of entities like fads, diseases, computer viruses, knowledge, products, etc. Early work in the area of diffusion in networks comes from the field of epidemiology [Bailey, 1975], where it deals with the spread of disease among individuals connected as a network; and from computer science [Newman et al., 2002], involving the study of the spread of computer viruses over the individuals connected via e-mail networks. This early work is primarily based on the SIR (Susceptible, Infected/Infectious and Recovered/Removed) and the SIRS (Susceptible, Infected/Infectious, Recovered/Removed and Susceptible) models.

In the field of sociology, the role of the *word-of-mouth* in the propagation of innovation in social networks has been extensively studied. Similarly, in marketing science literature many models for studying the diffusion of new products have been proposed [Mahajan et al., 1990]. For example, the primitive Bass model [Bass, 1969] predicts adoption based on relative populations of *innovators* that are not influenced by the decisions of others and *imitators* whose adoption depends on the total number of adoptions in the system. In the field of social network analysis, the flow of information in a social network is modeled as the propagation of *innovation* in the social network and was introduced by Ryan and Gross (1943) and later consolidated by Rogers (1995). Unlike the disease spread models, information diffusion in social networks is based on *complex social contagion* [Morris, 2000; Centola and Macy, 2007] that has properties like thresholds to infection, i.e., individuals wait for several of their friends to adopt the behavior before involving themselves.

With an increased popularity of the new enhanced forms of communication (like mobile phones, e-mail, online social networks, etc.) in recent years, researchers have been able to measure social contagion efficiently and exploit it to propose effective models for information diffusion in social networks. Katona et al. (2011) analyze an OSN dataset to identify individual word-of-mouth effects with an aim to discover how the local communication network structure affects the diffusion process. They show that dense groups/communities facilitate higher rate of word-of-mouth influence and that influencers who occupy structural holes in the network have, on average, higher influential power. Moreover, people with many friends have a lower average influence than those with fewer friends. They also suggest that demographic data is useful to identify strong influencers and together with global network variables are also useful to identify adopters as well. Along a similar direction, Bhat and Abulaish (2013) analyze the influential significance of overlapping nodes,

i.e., nodes which belong to multiple communities in a social network. Their analysis highlights that highly overlapping nodes in a social network represent the best influential nodes in the network in terms of their betweenness centrality and that outliers and single membership nodes can be easily discarded as least influential. Considering the case of Facebook, Bakshy et al. (2012) empirically re-highlight the strength of weak ties and signify that weak ties, defined directly in terms of interaction propensities, diffuse novel information that would not have otherwise spread and thus play an important role in facilitating information diffusion. Their analysis also indicates that most information diffusion on Facebook is surprisingly driven by simple contagion and not by complex contagion. Lerman et al. (2012) conducted empirical analysis of user activity on the online social networking sites Digg and Twitter, in the light of information diffusion cascades in these networks. Their analysis reveals that as Digg networks are dense and contain community structures, many of the cascades appear to spread through an interconnected community. On the other hand as Twitter does not contain significant community structures, its cascades are more tree-like. Based on a viral email experiment involving 31,183 individuals, Iribarren and Moro (2009) relate the unexpected slow pace of information diffusion in an email network to the heterogeneity found in the response time of human activity patterns. Similar results related to the effects of heterogeneity in both the number of recommendations made by individuals and of the time they take to transmit the information are reported in [Iribarren and Moro, 2008]. Aral et al. (2007) analyze a social network of an organization represented by ten months of email traffic observed over two five month periods. Besides, they also use some information related to accounting data on project co-work relationships between the workers in the organization. Their analysis reveals some interesting observations which include: In general it is not only the social structure but also the information content that determine the movement of information; the productivity of an information worker is directly proportional to her access of more novel and timely information. The analysis of Sadilek et al. (2012) and Corley et al. (2009) throws some light on how the content and demographic information available from online social media can be helpful in detecting and tracking disease spread in the underlying population. Some basic mathematical models that have been proposed in literature for analyzing information diffusion in social networks are explained in the following sub-sections.

6.1 Threshold Models

In threshold models of information diffusion, a node v in a network may adopt an action (become active) only if a certain number of nodes in v 's neighborhood are active, i.e., when the number of active nodes in the neighborhood of a node v exceeds a certain *threshold* the node v becomes active. The most simple example is the *linear threshold model* [Granovetter, 1987] in which a node v belonging to the network has a nonnegative weight $w_{v,u}$ for every node u in v 's neighborhood $N(v)$ with a property that $\sum_{u \in N(v)} w_{v,u} \leq 1$. Given a threshold value θ_v (value can be a fixed value or chosen at the start of the process) and an initial set A_1 of active nodes, it follows a sequence of steps in such a way that at any time t , every node that was active at time $t-1$ remains active and each node v that was inactive at time $t-1$ becomes active at time t if and only if $\sum_{u \in N(v)} w_{v,u} X_{u,t-1} \geq \theta_v$ where $X_{u,t-1}$ is 1 if u was active at time $t-1$ and 0, otherwise. Thus, the weight $w_{v,u}$ represents the extent to which node v is influenced by node u , and the threshold θ_v represents the personal tendency of v to adopt a new action of its neighbors.

6.2 Cascade Models

In cascade models of information diffusion [Goldenberg et al., 2001a; Goldenberg et al., 2001b], each individual in a social network who adopts an action (becomes active) has a single probabilistic chance to activate each inactive node in its neighborhood. For example, in the *independent cascade model* [Goldenberg et al., 2001b], given an initial set of active nodes in the network the process proceeds in a series of time steps where at each time step a node u that has just become active may attempt to activate each inactive node v in its neighborhood irrespective of the set of neighbors of u that have attempted to activate v in the past. Whether or not u becomes active, v and u have no further contact throughout the remainder of the process. The process terminates when no new activations can be made.

Kempe et al. (2003) give a generalized model for the independent cascade and the linear threshold models where the probability with which an active node u attempts to activate an inactive node v depends on the neighbors of u who have already attempted activate u . However, the probability that an individual v is active after the activation process does not depend on any sequence of the activation attempts of the neighbors and removes any ambiguity related to any simultaneous activation attempts made by the active neighbors.

6.3 Recent Works on Information Diffusion

Jackson and Yarvi (2005) have analyzed that for the spread of a behavior in a social network there exists a threshold for the number of initial active nodes (initial adopters) where ‘tipping’ occurs. It means that a large number of initial active node results in an increased adoption rate which reaches to a large sub-population. On the other hand a smaller number of seed nodes lead to the collapse of adoption behavior leading to very few or no new active nodes. Furthermore, once the tipping point is surpassed, they observe that the initial adoption rate is high, followed by a slower adoption rate towards the end. Moreover, the network structure such as the degree distributions of the nodes affects the tipping point and the adoption rate. A similar observation has been made by Iribarren and Moro (2008), who argue that most of the initial diffusion process takes place due to super-spreading events and since there exists heterogeneity in the scheduling of information transmission by individuals [Barabási, 2005; Aiello et al., 2000; Gruhl et al., 2004], the diffusion process slows down towards the end in logarithmic time. Garg et al. (2009) have considered the role of peers (that represent non-explicit relationships of users in a network) on the diffusion of niche information in a social network using a dataset from Last.fm. The peers of a user they considered had very small life span of connection as it was dependent on the online user’s evolving taste in music. Their analysis shows that there is a positive influence of online peers (non-explicit relationships) on the diffusion of niche information at a more granular level and found that users are 6 times more likely to discover a new track as a result of peer influence and discover 2.7 niche tracks as a result of that influence. Using blog contents posted by online social network users, Gruhl et al. (2004) have studied information diffusion in terms of topics propagation from one blog to another. They have characterized information diffusion along two dimensions – *topics* and *individuals*. In topic-based characterization process they found that the blog world contains numerous topics identifiable by their respective set of user postings, which are mostly composed of a union of active discussions (called *chatter*) mediated by the authors of the topic and short-term, high-intensity discussions (called *spikes*). In individual-based characterization, they characterize users on the basis of their respective posts related to the birth, activeness and death of a particular topic. They propose an SIRS based model to analyze the process of information diffusion so as to identify individuals who play an important role in the formation and spread of *infectious* topics.

Based on a *blog world* representation of online social networks, where a function called *blogroll* enables bloggers to specify *explicit relationships* with other bloggers, and *trackback* and *scrap* functions allow bloggers to make their posts linked to other bloggers' posts and to copy other bloggers' posts to their blogs respectively, Kwon et al. (2009a) analyzed the diffusion of information in these networks. In contrast to the assumption based on social network theory that information diffusion in social networks occurs through the established relations between members [Brown and Reinegen, 1987], the analyses of Kwon et al. (2009a) show that a majority of information diffuses between blogs that have no explicit relationships. Furthermore, some posts show an explosive increase in the number of blogs that trackback or scrap these posts. The reasons for such explosive information diffusion are identified as i) listing blogs on the main page of a blog world service portal, and ii) diffusion through search engines. Based on these observations, Kwon et al. (2009b) proposed an information diffusion model using the existing *independent cascade model* [Goldenberg et al., 2001b] as a basis and added to it the *virtual space* (i.e., the main page of the blog service provider) that exposes posts to a lot of bloggers with no prior relationship. Zhao et al. (2010) seek to determine the role of *weak-ties* in the process of spreading information within the blog world. They show that although many important features of the network structure are defined by the weak-ties, exclusively republishing blogs through weak-ties (scrap function of the blog world) cannot facilitate diffusion of information through the network. On the other hand republishing blogs selecting at

random cannot facilitate the diffusion process. Moreover, when the blogs are selected at random for republishing, removing weak ties leads to a sharp decrease in the rate and range of the diffusion. This concludes that for blog-networks, it is difficult to analyze the delicate role played by weak-ties in the information diffusion process. Studying the communication patterns of e-mail networks over time, Kossinets et al. (2008) formulated the notion of ‘distance’ between two nodes in a social network based on the minimum time it takes for information to diffuse from one node to the other. They identify the backbone of the network (a subgraph of a network on which information has the potential to flow the quickest) as a sparse graph with a concentration of both highly embedded edges and long-range bridges – reflecting the relationship between tie strength and connectivity in social networks. By analyzing the content of the blog posts, Stewart et al. (2007) modeled the problem of discovering information diffusion paths from the blogosphere as a problem of frequent pattern mining by representing a blog community collected in a certain time period as a blog sequence database. They define Information Diffusion Paths (IDP) as sequences of blogs that frequently discuss similar topics sequentially, and around similar time points.

Based on *Continuous-Time Markov Chains* [Norris, 1997], the information diffusion model proposed by Song et al. (2007) deals with predicting the flow of information where they measure the likelihood of the propagation of information from a specific sender to a specific receiver during a certain time period and propose a recommendation algorithm that predicts the most likely node to receive the information during a limited time period. Using diffusion-rate of information in a social network by estimating the expected time for information to diffuse to a specific user, the model also ranks nodes/ users based on how quickly information flows to them. By considering the picture popularity distributed over the Flickr social network, Cha et al. (2009) have shown that social links are the dominant method of information propagation and that information spreading is limited to individuals who are within close proximity of the picture uploaders. Furthermore, information takes a long time to reach from one node to other and hence the popularity of Flickr pictures steadily increases over many years. The role of *social influence* in the diffusion of information in social networks is studied by Oh et al. (2008) using a dataset from YouTube. Their analysis reveals interesting relationships between social influence resulting from a user’s network-position and the initial and latter stages of information diffusion in a social network. Habiba et al. (2010) define *spread* of a diffusion process in terms of the number of nodes expected to be affected by a stochastic diffusion process over time. They consider a node to qualify as a *best spread blocker* if its removal leads to a decrease in the spread of a diffusion process. Their results show that on both static and dynamic networks, local measures like the node degree, etc., perform at par with other measures in identifying the blockers. Instead of utilizing the topological information of a network and with an aim of predicting the magnitude and the rate of information diffusion, Yang and Leskovec (2010) propose a *Linear Influence Model* wherein (considering the network as implicit) the number of newly infected/influenced nodes by a node p at any given time t is a function of the number of nodes infected/influenced by p before time t . They model the influence functions as a regression task in a non-parametric way and show that they can be estimated using a simple least squares procedure.

6.4 Influential Node Mining and Viral Marketing

The social networks of customers have long been expected to have a potential for being exploited to increase brand/product awareness as word-of-mouth (WOM) flows through these networks facilitating social contagion. Moreover, the adoptions and opinions of certain customers (opinion leaders and influential customers) often influence the adoption behavior of other customers making their identification desirable for product manufacturers and marketing agencies [Wuyts et al., 2010]. In this direction, Trusov et al. (2009) quantify the effect of WOM referrals based on an OSN dataset which contains the records of new members which join the site. Their analysis reveals that the addition of new members to OSNs is strongly facilitated by WOM referrals whose impact is found to be almost 20 times higher than marketing events and 30 times higher than media appearances.

Viral marketing refer to the marketing techniques that are based on utilizing existing social networks in order to increase brand awareness or achieve other marketing objectives like increasing product sales by incorporating a self-replicating *viral* process that is analogous to the spread of pathological

or computer viruses. Viral marketing programs involve identifying individuals with high *social networking potential* (size of an individual's social network and their ability to influence that network) or *network value*, and creating viral messages that motivate this set of population to form a countable customer base for the brand. Network value of a customer is defined as the expected increase in sales to others that results from marketing to that customer [Domingos, 2005]. Some of the influence factors identified by Domingos (2005) that collectively influence the network value of a customer in a positive manner are:

- high connectivity in the network
- interest in the product
- leadership or asymmetric influence over the network
- higher level of *cascading-influence*

The problem of viral marketing was formalized by Kempe et al. (2003) as selecting the optimal individuals to be seeded with a product in an arbitrary network given a fixed marketing budget. This strategy involves encouraging the word-of-mouth by distributing discounted or free products to targeted consumers assuming that they will then discuss the product with their friends and encourage them to buy the product. However, what customers to seed with these initial products in order to maximize the amount and rate of product adoption are not obvious. Sun and Tang (2011) present a survey on the some existing models for social influence analysis. They present the notion of influence in terms of some important social network primitives including node degree, edge betweenness, structural holes, homophily and so on. They also review the notion of influence in terms of the actions and interactions of individuals in a social network and how community structures help in explaining influence in a social network. Moreover, they present some existing models for maximizing the influence spread in social networks which besides the Threshold (section 6.1) and the Cascade (section 6.2) models include *High-degree heuristic*, *Low-distance heuristic* and *Degree discount heuristic*. Here the high-degree and Low-distance heuristics consider nodes with higher degree (thus possibly can influence more nodes) and nodes with the shortest paths to other nodes as seed nodes respectively. The Degree discount heuristic is similar to High-degree heuristic but involves discounting the degree for a new potential seed node p by the number of already selected seed nodes that participate in defining p 's degree.

Online social networks like Facebook, Twitter, Orkut, etc., provide efficient platforms to advertise and market products to consumers as these platforms allow their users to create virtual networks which provide a formalization of social interactions of individuals. However, despite a huge marketing scope in these online social networks, it has been difficult to use this platform successfully for marketing [Baruh, 2009] often due to privacy considerations as the full network described by these online platforms is not known. As a solution to this problem, Stonedahl et al. (2010) have introduced the LVMP (*local viral marketing problem*) similar to the global viral marketing problem proposed by Kempe et al. (2003). The only difference is that in LVMP the structural knowledge of the global network is not available, rather only the characteristics of each vertex that provide summary statistics about the vertex and its role in the network are used. Instead of selecting individuals in order to maximize the influence, Hartline et al. (2008) and Arthur et al. (2009) have studied the problem of *revenue maximization*, i.e., the process of selecting an individual for offering an optimal price such that maximum revenue can be extracted from each next buyer. Their methods are based on the marketing strategies called *influence-and-exploit strategies* that initially *influence* a population by giving free or discounted products to set of influential buyers and then extract revenue from the remaining buyers using a 'greedy' pricing strategy. However, in contrast to Hartline et al. (2008) approach in which an influenced node is allowed to market product to any arbitrary target nodes and also determine the price of the item without considering the network structure, Arthur et al. (2009) incorporate structural information about the network and the timing of an offer is determined through cascading the recommendations. As a dual to the influence maximization problem, Kimura et al. (2008 & 2009) have addressed a *contamination minimization problem*, which involves minimizing the diffusion of unwanted information (under independent cascade model) by preventing the flow of information through a small subset of edges a network. In

particular, they have shown that unlike the case of removing nodes, blocking the edges which connect nodes with higher number of outgoing edges does not always work effectively for minimizing the reach of a spreading process. They also propose an effective approach to find an approximate solution by following a greedy strategy involving *bond percolation*. Percolation theory proposed in [Broadbent and Hammersley, 1957] examines how connectivity is disrupted within spatially structured systems [Stauffer and Aharony, 1991] and refers to a class of models that describe the properties of a system given the networking among its constituents

Besides the above research works, a number of researchers have specifically concentrated on identifying leaders and influential bloggers from online social networks to assist viral marketing of products and other allied tasks over these platforms. In support of these issues, a survey of over 200 million bloggers by McCann (2009) reveals that 31.7% blog about opinions related to products and brands which are viewed by 71% of regular internet users. Similarly, a survey performed by Nelsenwire (2009) on a relatively smaller population of internet users in some countries finds that 70% of people give significance to the online opinions of other people on the products they intend to purchase. For the managerial task of market analysis, blogs represent vital sources of information in the form of directly accessible insights on products (own and/or rival) and related customer feedback [Richardson and Domingos, 2002; Lawrence et al., 2010]. Agarwal et al. (2008) have proposed a model to identify influential bloggers by analyzing the influence of their blog posts based on four blog properties: *recognition*, *activity generation*, *novelty*, and *eloquence*, that form the parameters for the model and can be tuned to obtain different breeds of influential bloggers. Furthermore, their experimental results also confirm that the influential bloggers in a blogging community are not necessarily active bloggers. Goyal et al. (2008) have proposed a frequent pattern mining approach to discover leaders in social networks based on analyzing various user actions in the social network. Their method requires preliminary knowledge about the underlying social graph and an action log containing all user actions and their time stamps. By computing an *influence graph* and *influence cube* from available action log data and introducing various thresholds, they present frequent pattern mining based algorithms to identify *leaders*, *tribe leaders* and *confidence leaders*. Similarly, considering the log-in behavior of users of an OSN, Trusov et al. (2010) use a Bayesian approach to determine the influential nodes from an ego-centered influence network estimated from the log-in sequences followed by friend relations.

6.5 Conflicting Issues in Viral Marketing

An important challenge in the direction of viral marketing is due to the contradicting claims by Watts and Dodds (2007) that the *influentials* hypothesis is incorrect and by Goldenberg et al. (2009) that the *influentials* hypothesis is correct. More specifically, a computer simulation by Watts and Dodds (2007) suggested that seeding well-connected people to maximize the spread of information works only under certain conditions and should be less preferred as seeds or early adopters for large referral cascades. On the other hand, the study of Goldenberg et al. (2009) on the impact of social position on information-probability indicates that opinion leaders (hubs) in the network may adopt early not because they are innovative but because they are better informed than others by early exposure to innovations through their multiple social links. The analysis of Hinz et al (2011) and Iyengar et al. (2011) along this issue highlights that seeding hubs (high-degree seeding) for viral marketing results in higher number of referrals because hubs are more actively involved in the diffusion process due to the higher number of links. Moreover, in contrast to Goldenberg et al. (2009) the analysis of Iyengar et al. (2011) significantly highlights that opinion leaders associate to early adoption even after controlling for contagion, and are equally sensitive to contagion as non-leaders. In this section we present some more important and relevant work related to influential node mining and viral marketing and how online social networks present these opportunities. In addition to this, the analysis of Leskovec et al. (2007) highlights some interesting and challenging issues related to the traditional diffusion models in the context of recommendation which tend to change the perception of information diffusion and hence the approaches towards viral marketing. Unlike the traditional models of information diffusion, they observe that a repeated exposure of an individual to some product or resource decreases the chance of its adoption by the individual. It means that providing excessive incentives to selected seed customers could have negative effects.

As discussed earlier, traditional diffusion models consider that each individual either has the same chance of infecting each of its neighbors (cascade models), or that an individual does not get infected unless the number of its neighbors who are infected is greater than a threshold (threshold models). In either case, the chance of a node getting infected is directly proportional to number of its infected neighbors. However, Leskovec et al. (2007) argue that although the chance of a product p , to be purchased by an individual i , increases with the number of i related suggestions directed at p , the chance of adoption quickly decreases to a relatively low level. This indicates that even though a product may be suggested by many friends to an individual i , it is possible that the individual i does not buy the suggested product as it may seem to be of no significant use. Earlier, we also mentioned that often highly connected nodes are considered good candidates for selecting influential nodes. However, Leskovec et al. (2007) argue that nodes with high degrees can give productive recommendations only up to a certain level since the success per recommendation declines as a high-degree node sends out a significantly large number of recommendations for a certain product. In this regard they present a stochastic model which supports the diffusion of influence/recommendation through long paths, but also considers the possibility that the recommendation paths can get blocked at short lengths as discussed earlier.

Marketing is one of the major application areas for the analysis of the diffusion of information and influence in social networks. However, besides marketing, areas including behavioral science and the study of the spread of computer viruses and spam, etc., also find themselves striving for finding information diffusion pathways and influential nodes from online social networks in order to have a better understanding for controlling the possible epidemics of their outburst. Although, many issues relating to the social process of diffusion have been addressed, some young directions still need more attention. For example, studying diffusion trends within and across the communities in a social network, and analyzing the community structure of influential nodes for information diffusion seem promising. Similarly, addressing link prediction issues in the light of influence and information diffusion tendency of the nodes in a social network might also highlight new factors that can possibly determine the direction and weights of new links that can possibly appear in the social network in near future. The field also calls for identifying new practical application areas, and demonstration of the working of various diffusion models and node influence measuring/prediction methods.

7 Future Directions and Conclusion

Most of the existing current literature on social network analysis tasks is oriented towards network modeling and community detection, followed by link prediction, information diffusion and influential node mining. However, currently all of these tasks face some common challenges and issues. One of the most common issues is the dynamic nature of the real-world social networks that tend to change with time. Incorporating this feature of social networks into statistical models can help in efficiently dealing with the link prediction problem. Besides the existence of various dynamic community detection methods, analyzing overlapping and hierarchical community structures in evolving networks still needs some more attention. Another issue related to most of the tasks addressed in this paper is that of the heterogeneity found in real-world social networks. Often social networks involve different types of nodes representing users, resources (URLs, videos, etc.), and so on. Similarly, social links can also be associated with different attributes like polarity, relationship types (friend, foe, family, etc.) and so on. Community detection algorithms need to deal with this heterogeneity by ensuring that the nodes and links of similar types are grouped within same communities besides considering the topological structure of the social network. It may also be required to analyze as how/why nodes and links change their discriminating attribute values with the evolution of social network. Link mining and information diffusion models may need to address such issues. Opinion and sentiment-aware community detection, link prediction and information diffusion also seem to be promising as very less work has been done along this direction. Another issue related to the tasks involved in social network analysis is that most of the works have been done considering only undirected and un-weighted nature of social networks. Incorporating weights and directions to links for social network mining tasks is still an open challenge. For example, predicting both direction and weight of missing or future links in a social network. Moreover, a huge

amount of literature along the direction of social network analysis exists, and most of them concentrate on some specific aspect of the social networks (like community detection, information diffusion, etc.), in isolation or are mostly oriented only towards sociological aspects relieving computer science. However, the present multi-dimensional online social networks provide a means of studying various data mining tasks related to social network analysis in a unified framework where each of them can benefit from the others. For example, modeling information diffusion process in conjunction with centrality analysis and community detection can help to achieve more cost-effective viral marketing. Establishing relationships between opinion orientation, community structures and diffusion pathways in a population can lead to significant advances in the analysis of large real-world social networks. Similarly, using both topological network structure and the non-structural information like user-generated content for SNA tasks seems more promising than using only the topological information. Online social networks are a rich source of both structural and non-structural data and it seems they form an efficient platform for formulating frameworks that include multiple social network activities and data mining tasks being performed in collaboration with each other to yield more meaningful and realistic results.

In this regard we propose a conceptual framework (see Fig. 4) which allows a unified analysis of online social network data in which various SNA tasks can be made to supplement each other for a better interpretation and analysis of the underlying processes of online social networks. Community detection module can help to understand which individuals tend to have higher intensities of relationships than the rest of the network. For interaction networks, identified communities can represent clusters of individuals who interact with each other more frequently than the rest of the network and for friendship networks communities can represent the clusters of friends who have higher affinity with each other than the rest of the network. The main advantage could be to understand the relationship (similarity and difference) between different communities of interaction and friendship networks. For example, we may tend to answer the question: *do the groups of close friends interact more frequently or the higher intensities of interaction among a group of individuals are independent of the level of friendship among the same individuals.*

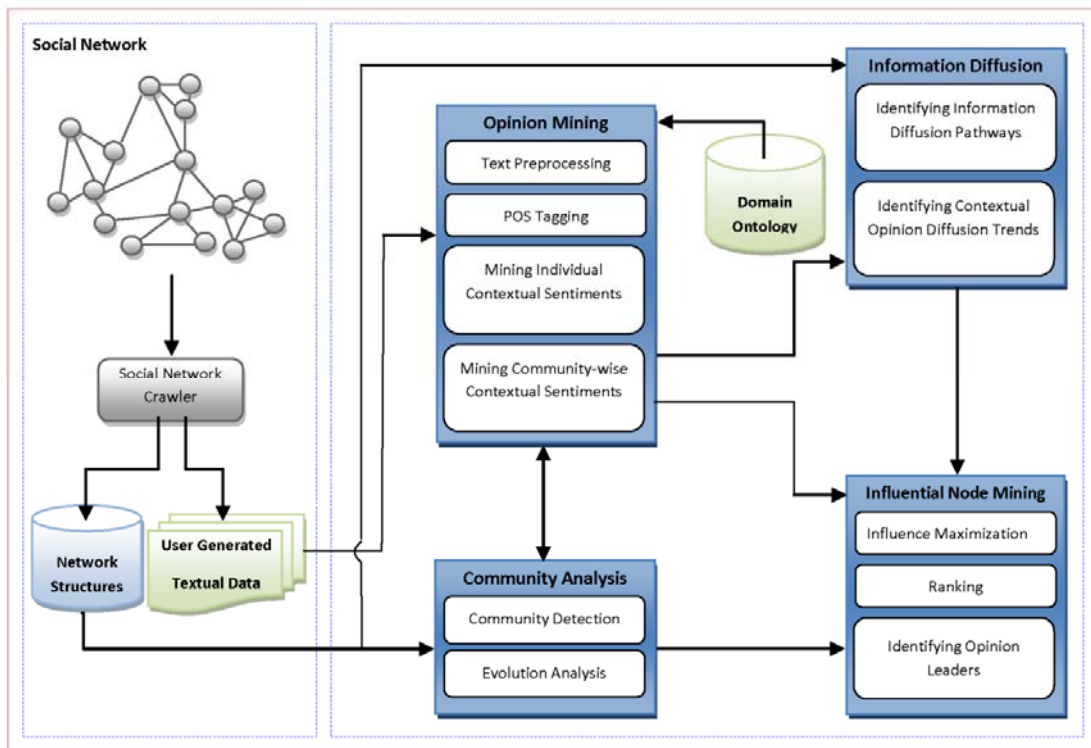


Fig. 4. A unified conceptual framework for online social network analysis

Another task related to the proposed framework is to analyze user-generated content generated by the individuals of a social network. This can help in analyzing sentiments of individuals which represent their opinions about various objects like products, events, government policies, etc. Text mining techniques along with information retrieval (IR) and natural language processing (NLP) can be used for this purpose. Opinion mining techniques are usually based on the consideration that a document, sentence, or feature/aspect is opinionated about one single issue or object and the problem is to classify the opinion as *positive*, *negative* or *neutral* or on a multi-point scale to determine the *strength* of the expressed opinion; for example, to determine whether a user specifies the picture quality of a camera phone as *neutral*, *low*, *medium*, or *high*, where neutral is treated as the lack of opinion. However, the user-generated content can be oriented towards numerous heterogeneous objects or contexts and it may be required to analyze the sentiments of individuals in different contexts depending upon the temporal requirements. In order to efficiently represent and interpret a particular context, the idea is to utilize domain knowledge. Starting with a seed ontology, text mining technique along with IR and NLP can be used to enrich it further through learning concepts and their relationships from textual data source in an automatic way. The identified communities from interaction and friendship networks can be exploited to analyze the sentiments and opinions learned from textual data efficiently at both individual and community levels. The opinion strength of a community can be computed as the average strength of the opinions expressed by the individuals in that community. Furthermore, comparing opinion strengths and sentiments of the identified communities may reveal interesting relationships between the levels of interaction, friendship, and strengths of the opinions and orientation of sentiments. At this point, it can also be possible to identify central individuals and communities that tend to have a higher impact on contextual sentiment orientations of the individuals and communities.

The framework also facilitates to analyze and model the information diffusion process in both interaction and friendship networks. This process involves studying how far and at what rate the diffusion/spread of information can take place in a network. Information diffusion modeling of the interaction and friendship networks can help to determine which among the two networks has a higher scope of spreading information among the individuals, or does there exist a significant difference in the information diffusion process of the two networks. Furthermore, considering a particular context, it is also possible to understand how particular contextual opinions diffuse in the interaction, friendship, and community networks (where a node represents an identified community and the weighted edges represent the degree of similarity between them). For such a modeling, the opinion similarity between any two nodes will contribute in determining the probability of the spread between them.

Another SNA task supported by the proposed framework is to identify influential nodes (individuals) in the interaction and friendship networks. This has importance in domains like viral marketing which involves identifying individuals with high social networking potential (size of an individual's social network and their ability to influence that network). In order to identify influential nodes in the interaction and friendship networks, the individual nodes can be ranked by estimating diffusion probabilities from observed information diffusion data using independent cascade model discussed earlier. Furthermore, contextual opinion leaders (individuals and communities) can be identified by analyzing the interaction, friendship and community networks in terms of the diffusion of contextual opinions in the networks. It can also help in determining the relationship between the influential nodes identified separately in interaction and friendship networks. An important issue not significantly highlighted here is that the amount of data made available by online social networks is huge and it is growing exponentially. It calls for dealing with the scalability issues for the existing methods to face various challenges discussed in this paper, and requires improving existing methods or designing new fast and efficient methods to tackle the huge amount of heterogeneous and dynamic data.

The field of social network analysis is not new, but still it is rapidly growing with new challenges being faced due to the growth of new multidimensional social networks. This paper attempts to present a review of some of the latest and important aspects of social network analysis which include methods for social network modeling, analysis and mining. Existing techniques for many

analysis tasks are discussed and presented in a summarized way, which could be a useful source for the researchers in social network analysis area to get insight about the related problems and existing state-of-the art techniques

References

- AGARWAL, G. AND KEMPE, D. 2008. Modularity maximizing network communities using mathematical programming. *The European Physical Journal B* 66, 409–418.
- AGARWAL, N., LIU, H., TANG, L. AND YU, P. S. 2008. Identifying the influential bloggers in a community. In *Proceedings of the international Conference on Web Search and Web Data Mining* (Palo Alto, California, USA, February 11 - 12, 2008). WSDM '08. ACM, New York, NY, 207-218.
- AGRAWAL, R., RANTZAU, R. AND TERZI, E. 2006. Context-sensitive ranking. In *Proceedings of the 2006 ACM SIGMOD international Conference on Management of Data* (Chicago, IL, USA, June 27 - 29, 2006). SIGMOD '06. ACM, New York, NY, 383-394.
- AHN, Y. Y., BAGROW, J. P. AND LEHMANN, S. 2010. Link communities reveal multi-scale complexity in networks. *Nature* 466, 761-764, (5 August 2010), doi:10.1038/nature09182
- AIELLO, W., CHUNG, F. AND LU, L. 2000. A random graph model for power law graphs. *Experimental Math* 10, 53-66.
- ALEXANDER, J. M. 2009. *Evolutionary Game Theory*, Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, CSLI, Stanford University.
- ALBERT, R., JEONG, H., AND BARABÁSI, A. L. 1999. The Diameter of the WWW. *Nature* 40, 6749 130–131. arXiv:cond-mat/9907038. doi:10.1038/43601.
- ANDERSON, C., WASSERMAN, S. AND CROUCH, B. 1999. A p* primer: logit models for social networks. *Social Networks* 21, 1, 37–66, 1999. ISSN 0378-8733, DOI: 10.1016/S0378-8733(98)00012-4.
- ARAL, S., BRYNJOLFSSON, E. AND ALSTYNE, M. W. V. 2007. Productivity Effects of Information Diffusion in Networks. In *Proceedings of the 28th Annual International Conference on Information Systems*, Montreal, CA.
- ARENAS, A., DANON, L., DÍAZ-GUILERA, A., GLEISER, P. M. AND GUIMERÁ, R. 2004. Community analysis in social networks. *European Physical Journal B* 38, 2, 373.
- ARENAS, A., FERNANDEZ, A. AND GOMEZ, S. 2008. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics* 10, 053039. doi: 10.1088/1367-2630/10/5/053039
- ARTHUR, D., MOTWANI, R., SHARMA, A. AND XU, Y. 2009. Pricing Strategies for Viral Marketing on Social Networks. In *Proceedings of the 5th international Workshop on internet and Network Economics* (Rome, Italy, December 14 - 18, 2009). S. Leonardi, Ed. Lecture Notes In Computer Science, vol. 5929. Springer-Verlag, Berlin, Heidelberg, 101-112.
- ASUR, S., PARTHASARATHY, S., AND UCAR, D. 2009. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data* 3, 4 (Nov. 2009), 1-36.
- AUMANN, R. AND MYERSON, R. 1988. Endogenous Formation of Links Between Players and Coalitions: An Application of the Shapley Value. In: Roth, A. (ed.) *The Shapley Value*, Cambridge University Press, 175-191.
- BACKSTROM, L., HUTTENLOCHER, D., KLEINBERG, J., AND LAN, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM, New York, NY, 44-54.
- BAILEY, N. T. J. 1975. *The Mathematical Theory of Infectious Diseases and its Applications* (2nd edition). Griffin, London.
- BAKSHY, E., ROSENN, I., MARLOW, C. AND ADAMIC, L. 2012. The role of social networks in information diffusion. In *WWW'12 Proceedings of the international conference on WWW*.
- BALA, V., AND GOYAL, S. 2000. A noncooperative model of network formation. *Econometrica* 68, 5, 1181-1229.

- BARABÁSI, A. L. AND ALBERT, R. 1999. Emergence of scaling in random networks, *Science* 286, 509-512.
- BARABÁSI, A. L. AND BONABEAU, E. 2003. Scale-free networks. *Scientific American* 288, 50-59.
- BARABÁSI, A. L. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435, 207.
- BARUH, L. 2009. Social Media Marketing: Web X.0 of Opportunities. In *Handbook of Research on Social Interaction Technologies and Collaboration Software: Concepts and Trends*, T. DUMOVA AND R. FIORDO, Eds. Idea Group, Inc., 2009, 33-45.
- BASS, F. M. 1969. A new product growth for model consumer durables. *Management Science* 15, 5, 215-227.
- BAUMES, J., GOLDBERG, M. K., KRISHNAMOORTHY, M. S., MAGDON-ISMAIL, M. AND PRESTON, N. 2005. Finding communities by clustering a graph into overlapping subgraphs. In *AC'05: Proceedings of the IADIS International Conference on Applied Computing*, N. GUIMARAES, AND P. T. ISAIAS, Eds. Algarve, Portugal, 97-104.
- BHAT, S. Y., AND ABULAISH, M. 2012. OCTracker: A Density-Based Framework for Tracking the Evolution of Overlapping Communities in OSNs. In *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'2012)*, 501-505. IEEE.
- BHAT, S. Y. AND ABULAISH, M. 2013. Overlapping Social Network Communities and Viral Marketing, In *International Symposium on Computational and Business Intelligence (To Appear)*, New Delhi, India.
- BHATTACHARYYA, P., GARG, A. AND WU, S. F. 2009. Social Network Model Based on Keyword Categorization. In *ASONAM '09: Proceedings of International Conference on Advances in Social Network Analysis and Mining*, 170-175, 20-22 July 2009. doi: 10.1109/ASONAM.2009.46.
- BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R. AND LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, P10008.
- BOETTCHER, S. AND PERCUS, A. 2001. Optimization with extremal dynamics. *Physical Review Letters* 86, 5211-5214.
- BONCHI, F. , CASTILLO, C., GIONIS, A. AND JAIMES, A. 2011. Social Network Analysis and Mining for Business Applications. *ACM Trans. Intell. Syst. Technol* 2, 3, Article 22 (May 2011), 37 pages.
- BORGATTI, S. P., MEHRA, A., BRASS, D. J. AND LABIANCA, G. 2009. Network Analysis in the Social Sciences. *Science* 13, 323 (5916), 892-895. DOI:10.1126/science.1165821
- BRANDES, U., DELLING, D., GAERTLER, M., GÖRKE, R., HOEFER, M., NIKOLOSKI, Z. AND WAGNER, D. 2006. On Modularity: NP-Completeness and Beyond. *Technical Report 2006-19*, ITI Wagner, Faculty of Informatics, Universität Karlsruhe, TH.
- BRESLIN, J. AND DECKER, S. 2007. The Future of Social Networks on the Internet: The Need for Semantics, *IEEE Internet Computing* 11, 6, 86-90, Nov.-Dec. 2007.
- BRESLIN, J., HARTH, A., BOJARS, U., AND DECKER, S. 2005. Towards semantically interlinked online communities. In *ESWC '05 Proceedings of the 2nd European Semantic Web Conference*, 71-83. Springer-Verlag.
- BROADBENT, S., HAMMERSLEY, J. 1957. Percolation processes: I. Crystals and mazes. In *Proceedings of the Cambridge Philosophical Society* 53, 629-641.
- BROWN, J. AND REINEGEN, P. 1987. Social Ties and Word-of-Mouth Referral Behavior. *Journal of Consumer Research* 1, 3, 350-362.
- BURT, R. S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard Univ Press.
- BURT, R. S. 1997. The contingent value of social capital. *Administrative Science Quarterly* 42, 339-365.
- BURT, R. S. 2004. Structural Holes and Good Ideas. *American Journal of Sociology* 110, 2, 349-399.
- CENTOLA, D., AND MACY, M. 2007. Complex contagions and the weakness of long ties1. *American Journal of Sociology* 113, 702-734.
- CHA, M., MISLOVE, A. AND GUMMADI, K. 2009. A measurement-driven analysis of information propagation in the Flickr social network. In *WWW'09: Proc. of the Eighteenth Int. Conf. on World Wide Web*, 721-730.

- CHEN, J., ZAÏANE, O. R., GOEBEL, R. 2009. Detecting Communities in Social Networks Using Max-Min Modularity. In *SDM 2009: Proceedings of the SIAM Data Mining Conference*, 978-989.
- CHEN, J. 2010. *Community Mining - Discovering Communities in Social Networks*, Ph.D. thesis, University of Alberta, Spring 2010, Edmonton, Alberta.
- CHICKERING, D. M. AND HECKERMAN, D. 2000. A decision-theoretic approach to targeted advertising. In *Proceedings of Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, Stanford University, Stanford, CA, 82-88.
- CHRISTENSEN, K., DONANGELO, R., KOILLER, B., AND SNEPPEN, K. 1998. Evolution of random networks. *Physical Review Letters* 81, 11, 2380-2383.
- CHUN, H., KWAK, H., EOM, Y., AHN, Y., MOON, S., AND JEONG, H. 2008. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM Conference on internet Measurement (Vouliagmeni, Greece, October 20 - 22, 2008)*. IMC '08. ACM, New York, NY, 57-70.
- CLAUSET, A., NEWMAN, M. E. J. AND MOORE, C. 2004. Finding community structure in very large networks, *Physical Review E* 70, 066111.
- CLAUSET, A., MOORE, C. AND NEWMAN, M. E. J. 2008. Hierarchical structure and the prediction of missing links in networks, *Nature* 453, 7191, 98-101.
- CLAUSET, A. 2005. Finding local community structure in networks. *Physical Review E* 72, 026132.
- CORLEY, C. D., MIKLER, A. R., SINGH, K. P. AND COOK, D. J. 2009. Monitoring influenza trends through mining social media. *Proceedings of the International Conference on Bioinformatics and Computational Biology (BIOCOMP09)*, 340-346. Las Vegas, NV, USA.
- DANON, L., DUCH, J. DÍAZ-GUILERA, A. AND ARENAS, A. 2005. Comparing community structure identification. *Journal of Statistical Mechanical*, P09008.
- DERÉNYI, I., PALLA, G. AND VICSEK, T. 2005. Clique percolation in random networks. *Physical Review Letters* 94, 160202.
- DOMINGOS, P. AND RICHARDSON, M. 2001. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Francisco, California, August 26 - 29, 2001)*. KDD '01. ACM, New York, NY, 57-66.
- DOMINGOS, P. 2005. Mining social networks for viral marketing. *IEEE Intelligent Systems* 20, 80-82.
- DOWNES, S. 2005. Semantic networks and social networks. *The Learning Organization* 12, 5, 11-417
- DUCH, J. AND ARENAS, A. 2005. Community detection in complex networks using extremal optimization. *Physical Review E* 72, 2, 027104.
- EAGLE, N., PENTLAND, A. AND LAZER, D. 2009. Inferring social network structure using mobile phone data, In *Proceedings of the National Academy of Science* 106, 15274-15278.
- ERÉTÉO, G., LIMPENS, F., GANDON, F., CORBY, O., BUFFA, M., LEITZELMAN, M. AND SANDER, P. 2011. Semantic Social Network Analysis, a concrete case. In *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena*. IGI Global: Hershey, PA.
- ESTER, M., KRIEGEL, H, JÖRG, S. AND XU, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the International Conference on Knowledge Discovery from Data*, 226-231.
- EVERETT, M. G. AND BORGATTI, S. P. 1994. Regular equivalence: General theory. *Journal of Mathematical Sociology* 19, 29-52.
- FALKOWSKI, T., BARTH, A. AND SPILIOPOULOU, M. 2007. DENGGRAPH: a density-based community detection algorithm. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, Washington, DC, USA. 112-115.
- FALKOWSKI, T., BARTH, A. AND SPILIOPOULOU, M. 2008. Studying community dynamics with an incremental graph mining algorithm. In *AMCIS'08: Proceedings of the 14th Americas Conference on Information Systems* 29, Toronto, Canada.
- FLAKE, G., LAWRENCE, S., GILES, C. AND COETZEE, F. 2002. Self-organization and identification of web communities. *IEEE Computer Society* 35, 3, 66.

- FORTUNATO, S. AND BARTHÉLEMY, M. 2007. Resolution limit in community detection. In *Proceedings of the National Academy of Science (USA)* 104, 1, 36-41.
- FORTUNATO, S. AND CASTELLANO, C. 2007. Community structure in graphs. *arXiv:0712.2716*, [physics.soc-ph].
- FORTUNATO, S. 2010. Community detection in graphs. *Physics Reports* 486, 75-174. DOI: 10.1016/j.physrep.2009.11.002.
- FRANK, O. AND STRAUSS, D. 1986. Markov graphs. *Journal of the American Statistical Association* 81, 832-842.
- GARG, R., TELANG, R. AND SMITH, M. 2009. Peer Influence and Information Diffusion in Online Networks: An Empirical Analysis, In *CIST'09: Conference on Information Systems and Technology* (October 10-11), San Diego, California, USA.
- GHONIEM, M., FEKETE, J. D. AND CASTAGLIOLA, P. 2005. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization* 4, 2, 114-135.
- GHONIEM, M., JUSSIEN, N. AND FEKETE, J. D. 2004. VISEXP: visualizing constraint solver dynamics using explanations. In *FLAIRS'04: Seventeenth international Florida Artificial Intelligence Research Society conference*, Miami, Florida, USA.
- GINTIS, H. 2000. *Game theory evolving: A problem-centered introduction to modeling strategic behavior*. Princeton University Press.
- GIRVAN, M. AND NEWMAN, M. E. J. 2002. Community structure in social and biological networks. In *Proceedings of the National Academy of Sciences (PNAS)* 99, 7821-7826.
- GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E., AND AIROLDI, E. M. 2010. A Survey of Statistical Network Models. *Foundation and Trends in Machine Learning* 2, 2 (Feb. 2010), 129-233.
- GOLDENBERG, J., HAN, S., LEHMANN, D. R. AND HONG, J. W. 2009. The role of hubs in the adoption process. *Journal of Marketing* 73, 2, 1-13.
- GOLDENBERG, J., LIBAI, B. AND MULLER, E. 2001a. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review* 1, 9.
- GOLDENBERG, J., LIBAI, B. AND MULLER, E. 2001b. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 3, 12, 211-223.
- GOYAL, A., BONCHI, F. AND LAKSHMANAN, L. V. 2008. Discovering leaders from community actions. In *Proceeding of the 17th ACM Conference on information and Knowledge Management* (Napa Valley, California, USA, October 26 - 30, 2008). CIKM '08. ACM, New York, NY, 499-508.
- GOYAL, S. 2012. *Connections: an introduction to the economics of networks*. Princeton University Press.
- GRANOVETTER, M. S. 1973. The strength of weak ties. *American Journal of Sociology* 78, 6, 1360-1380.
- GRANOVETTER, M. 1987. Threshold models of collective behavior. *American Journal of Sociology* 83, 6, 1420-1443.
- GREENE, D., DOYLE, D., AND CUNNINGHAM, P. 2010. Tracking the evolution of communities in dynamic social networks. In *Proceedings of ASONAM '10: the 2010 International Conference on Advances in Social Networks Analysis and Mining*. Washington, DC, USA: IEEE Computer Society, pp. 176-183.
- GROSS, T., AND BLASIUS, B. 2008. Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface* 5, 20, 259-271.
- GROSS, T., D'LIMA, C. J. D., AND BLASIUS, B. 2006. Epidemic dynamics on an adaptive network. *Physical review letters* 96, 20, 208701.
- GRUBER, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 199-220.
- GRUBER, T. R. 2008. Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semant.* 6, 1 (February 2008), 4-13.

- GRUHL, D., GUHA, R., LIBEN-NOWELL, D. AND TOMKINS, A. 2004. Information Diffusion Through Blogspace. In *Proceedings of the 13th international conference on World Wide Web*, ACM Press, New York, 491-501.
- GUIMERÀ, R. AND AMARAL, L. A. N. 2005. Functional cartography of complex metabolic networks. *Nature* 433, 7028, 895-900.
- HABIBA, YU, Y., BERGER-WOLF, T. Y. AND SAIA, J. 2010. Finding Spread Blockers in Dynamic Networks, In *Advances in Social Network Mining and Analysis*, L. GILES et al. Eds. Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 5498/2010, 55-76.
- HANDCOCK, M. S. 2002. Statistical Models for Social Networks: Inference and Degeneracy. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, R. BREIGER, K. CARLEY, AND P. E. PATTISON, Eds. National Research Council of the National Academies. Washington, DC: The National Academies Press, 229 – 240.
- HANDCOCK, M. 2003. Assessing degeneracy in statistical models of social networks. *Technical Report 39*, University of Washington.
- HANDCOCK, M. S., RAFTER, A. E. AND TANTRUM, J. M. 2007. Model-based clustering for social networks. *Journal of the Royal Statistical Society A* 170, 301–354.
- HANNEKE, S. AND XING, E. P. 2007. Discrete temporal models of social networks. In *Proceedings of the 2006 Conference on Statistical Network Analysis* (Pittsburgh, PA, USA). E. AIROLDI, D. M. BLEI, S. E. FIENBERG, A. GOLDENBERG, AND E. P. XING, Eds. Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg, 115-125.
- HANNEMAN, R. AND RIDDLE, C. 2005. *Introduction to Social Network Methods* (online textbook). University of California, Riverside, CA. Available at <http://faculty.ucr.edu/~hanneman/>
- HARTLINE, J., MIRROKNI, V., AND SUNDARARAJAN, M. 2008. Optimal marketing strategies over social networks. In *Proceeding of the 17th international Conference on World Wide Web* (Beijing, China, April 21 - 25, 2008). WWW '08. ACM, New York, NY, 189-198.
- HASAN, M. A., CHAOJI, V., SALEM, S. AND ZAKI, M. 2006. Link prediction using supervised learning. In *SDM'06: Proceedings of the Workshop on Link Analysis, Counter-terrorism and Security* (at SIAM Data Mining Conference), Bethesda, MD.
- HASAN, M. A. AND ZAKI, M. J. 2011. A survey of link prediction in social networks. In *Social Network Data Analytics*, Springer US, 243-275.
- HASTINGS, M. B. 2006. Community detection as an inference problem. *Physical Review E* 74, 035102.
- HASTIE, T., TIBSHIRANI, T. AND FRIEDMAN, J. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In *The Mathematical Intelligencer* 27, 2, 83-85, Springer, Berlin, DOI: 10.1007/BF02985802.
- HELLMANN, T., STAUDIGL, M. AND ROGERS, B. W. 2012. Evolution of Social networks. Working paper No. 470, Institute of Mathematical Economics, Bielefeld University, Germany.
- HENRY, N. AND FEKETE, J. D. 2007. MatLink: enhanced matrix visualization for analyzing social networks. In *Proceedings of the 11th IFIP TC 13 international Conference on Human-Computer interaction - Volume Part II* (Rio de Janeiro, Brazil, September 10 - 14, 2007). C. BARANAUSKAS, P. PALANQUE, J. ABASCAL, AND S. D. BARBOSA, Eds. Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg, 288-302.
- HINZ, O., SKIERA, B., BARROT, C. AND BECKER, J. U. 2011. Seeding Strategies for Viral Marketing: An Empirical Comparison. *Journal of Marketing* 75, 6, 55-71.
- HOLLAND, P. W., AND LEINHARDT, S. 1981. An exponential family of probability distributions for directed graphs. (With discussion.) *Journal of the American Statistical Association* 76, 33-65.
- HORROCKS, I. 2008. Ontologies and the semantic web. *Commun. ACM* 51, 12 (December 2008), 58-67.
- HUANG, J., SUN, H., HAN, J., DENG, H., SUN, Y. AND LIU, Y. 2010. Shrink: a structural clustering algorithm for detecting hierarchical communities in networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management, ser. CIKM '10*. New York, NY, USA: ACM, 219-228.

- HULST, R. C. V. D. 2008. Introduction to Social Network Analysis (SNA) as an investigative tool. *Trends in Organized Crime* 12, 2, 101-121.
- IRIBARREN, J. L. AND MORO, E. 2008. Information diffusion epidemics in social networks. Preprint: *arXiv:0706.0641v1*, [physics.soc-ph]
- IRIBARREN, J. L. AND MORO, E. 2009. Impact of Human Activity Patterns on the Dynamics of Information Diffusion, *Physical Review Letters* 103, 3, 038702.
- IYENGAR, R., DEN BULTE, C. V. AND VALENTE, T. W. 2011. Opinion Leadership and Social Contagion in New Product Diffusion. *Marketing Science* 30, 2, 195-212.
- JACKSON, M. O. AND YARIV, L. 2005. Diffusion on Social Networks, *Économie Publique* 16, 1, 3-16.
- JACKSON, M. O. 2005. A survey of network formation models: Stability and efficiency. *Group Formation in Economics: Networks, Clubs and Coalitions*, ed. G. Demange and M. Wooders, 11-57.
- JACKSON, M. O. 2010. *Social and economic networks*. Princeton University Press.
- JALAN, S., AND BANDYOPADHYAY, J. N. 2008. Random matrix analysis of network Laplacians, *Physica A: Statistical Mechanics and its Applications* 387, 2-3(15), 667-674, ISSN 0378-4371, DOI: 10.1016/j.physa.2007.09.026.
- JAMALI, M. AND ABOLHASSANI, H. 2006. Different Aspects of Social Network Analysis. In *Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence* (December 18 - 22, 2006). Web Intelligence. IEEE Computer Society, Washington, DC, 66-72.
- KAIRAM, S. R., WANG, D. J. AND LESKOVEC, J. 2012. The life and death of online groups: predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12)*. ACM, New York, NY, USA, 673-682.
- KASHIMA, H. AND ABE, N. 2006. A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction. In *Proceedings of the Sixth international Conference on Data Mining* (December 18 - 22, 2006). ICDM. IEEE Computer Society, Washington, DC, 340-349.
- KATONA, Z., ZUBCSEK, P. AND SARVARY, M. 2011. Network Effects and Personal Influences: The Diffusion of an Online Social Network, *Journal of Marketing Research XLVIII (June 2011)*, 425-443, American Marketing Association.
- KEMPE, D., KLEINBERG, J., AND TARDOS, É. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Washington, D.C., August 24 - 27, 2003). KDD '03. ACM, New York, NY, 137-146.
- KUMPULA, J. 2008. *Community Structures in Complex Networks: Detection and Modeling*, Ph.D. Thesis, Helsinki University of Technology, Finland.
- KUMPULA, J. M., M. KIVELÄ, K. KASKI, AND J. SARAMÄKI. 2008. Sequential algorithm for fast clique percolation. *Physical Review E* 78, 2, 026109.
- KIM M.-S. AND HAN, J. 2009. Chronicle: A two-stage density-based clustering algorithm for dynamic networks," in *Proceedings of the 12th International Conference on Discovery Science*, ser. DS '09. Berlin, Heidelberg: Springer-Verlag. 152-167.
- KIMURA, M., SAITO, K. AND MOTODA, H. 2008. Minimizing the spread of contamination by blocking links in a network. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 1175-1180.
- KIMURA, M., SAITO, K., AND MOTODA, H. 2009. Blocking links to minimize contamination spread in a social network. *ACM Transactions Knowledge Discovery from Data* 3, 2 (Apr. 2009), 1-23.
- KIRKPATRICK, S., GELATT, C. D. AND VECCHI, M. P. 1983. Optimizing by simulated annealing. *Science* 220, 671-680.
- KLEINBERG, J. M. 2001. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS) 14*. MIT Press, Cambridge, MA.
- KOSSINETS, G., KLEINBERG, J., AND WATTS, D. 2008. The Structure of Information Pathways in a Social Communication Network. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 435-443.

- KOWALD, M., FREI, A., HACKNEY, J. K., ILLENBERGER, J. AND AXHAUSEN, K. W. 2010. Collecting Data on Leisure Travel: The Link Between Leisure Contacts and Social Interactions, *Procedia - Social and Behavioral Sciences* 4, 38-48, ISSN 1877-0428.
- KRAUSE, A. E., FRANK, K. A., MASON, D. M., ULANOWICZ, R. E. AND TAYLOR, W. W. 2003. Compartments revealed in food-web structure. *Nature* 426, 6964, 282.
- KUMAR, P., WANG, L., CHAUHAN, J. AND ZHANG, K. 2009. Discovery and visualization of hierarchical overlapping communities from bibliography information. In *Proceedings of the 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, ser. DASC '09*. Washington, DC, USA: IEEE Computer Society, 664-669.
- KUMAR, R., NOVAK, J., AND TOMKINS, A. 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM, New York, NY, 611-617.
- KWON, Y., KIM, S., AND PARK, S. 2009a. An analysis of information diffusion in the blog world. In *Proceeding of the 1st ACM international Workshop on Complex Networks Meet information & Knowledge Management* (Hong Kong, China, November 06 - 06, 2009). CNIKM '09. ACM, New York, NY, 27-30.
- KWON, Y. S., KIM, S. W., PARK, S., LIM, S. H. AND LEE, J. B. 2009. The information diffusion model in the blog world. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (Paris, France, June 28 - 28, 2009). SNA-KDD '09. ACM, New York, NY, 1-9.
- LAHIRI, M. AND BERGER-WOLF, T. Y. 2008. Mining periodic behavior in dynamic social networks. In *ICDM '08: Proceedings of the International Conference on Data Mining*, Pisa, Italy, 373-382. doi: 10.1109/ICDM.2008.104.
- LANCICHINETTI, A., FORTUNATO, S. AND KERTÉSZ, J. 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* 11, 3, 033015.
- LANCICHINETTI, A., RADICCHI, F., RAMASCO, J. J. AND FORTUNATO, S. 2011. Finding statistically significant communities in networks. *PLoS ONE* 6, 5.
- LATOUCHE, P., BIRMELE, E. AND AMBROISE, C. 2009. Overlapping stochastic block models. *Tech. rep. arXiv:0910.2098*. URL <http://arxiv.org/abs/0910.2098>
- LAWRENCE, R., MELVILLE, P., PERLICH, C., SINDHWANI, V., MELIKSETIAN, S., HSUEH, P. Y. AND LIU, Y. 2010. Social Media Analytics- The Next Generation Of Analytics-Based Marketing Seeks Insights From Blogs, *OR/MS Today*, February, 2010, 26-30, Lionheart Publishing Inc., Marietta, USA.
- LERMAN, K., GHOSH, R. AND SURACHAWALA, T. 2012. Social contagion: An empirical study of information spread on Digg and Twitter follower graphs. [arXiv:1202.3162v1](https://arxiv.org/abs/1202.3162v1) [cs.SI]
- LEROY, V., CAMBAZOGLU, B. B. AND BONCHI, F. 2010. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA, July 25 - 28, 2010). KDD '10. ACM, New York, NY, 393-402.
- LESKOVEC, J., ADAMIC, L. A. AND HUBERMAN, B. A. 2007. The dynamics of viral marketing. *ACM Trans. Web* 1, 1, Article 5 (May 2007).
- LESKOVEC, J., BACKSTROM, L., KUMAR, R., AND TOMKINS, A. 2008. Microscopic evolution of social networks. In *Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA, August 24 - 27, 2008). KDD '08. ACM, New York, NY, 462-470.
- LEWIS, A. C. F., JONES, N. S., PORTER, M. A. AND DEANE, C. M. 2010. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology* 4, 100
- LIBEN-NOWELL, D. AND KLEINBERG, J. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58, 7 (May. 2007), 1019-1031.
- LIN, Y., CHI, Y., ZHU, S., SUNDARAM, H. AND TSENG, B. L. 2009. Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data* 3, 2 (Apr. 2009), 1-31.
- LIN, Y., SUNDARAM, H., CHI, Y., TATEMURA, J., AND TSENG, B. L. 2007. Blog Community Discovery and Evolution Based on Mutual Awareness Expansion. In *Proceedings of the IEEE/WIC/ACM*

- international Conference on Web intelligence* (November 02 - 05, 2007). Web Intelligence. IEEE Computer Society, Washington, DC, 48-56.
- LI, P., LI, Z., LIU, H., HE, J. AND DU, X. 2009. Using Link-Based Content Analysis to Measure Document Similarity Effectively. In *Advances in Data and Web Management*, Lecture Notes in Computer Science, 5446/2009, 455-467, DOI: 10.1007/978-3-642-00672-2_40.
- LIU, W. AND LÜ, L. 2010. Link prediction based on local random walk. *Europhysics Letters* 89, 58007.
- LORRAIN, F. AND WHITE, H. C. 1971. The structural equivalence of individuals in social networks. *Journal of Mathematical Sociology* 1, 49-80.
- LÜ, L. AND ZHOU, T. 2010. Link prediction in weighted networks: The role of weak ties, *Europhysics Letters* 89, 18001.
- LUSSEAU, D., SCHNEIDER, K., BOISSEAU, O. J., HAASE, P., SLOOTEN, E. AND DAWSON, S. M. 2003. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology* 54, 396-405.
- MAHAJAN, V., MULLER, E. AND BASS, F. M. 1990. New product diffusion models in marketing: A review and directions for research. *Journal of Marketing* 54, 1, 1-26.
- MCCANN, U. 2009. *Power to the people: Social media tracker wave 4*. Retrieved from <http://universalmccann.bitecp.com/wave4/Wave4.pdf>
- MCD AID, A. AND HURLEY, N. 2010. Detecting highly overlapping communities with model-based overlapping seed expansion. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '10. Washington, DC, USA: IEEE Computer Society, 2010, 112-119.
- MIKA, P. 2005b. Social networks and the semantic web: The next challenge. *IEEE Intelligent Systems* 20, 1, 80-93
- MIKA, P. 2006. Ontologies are us: A unified model of social networks and semantics, *Web Semantics: Science, Services and Agents on the World Wide Web* 5, 1, March 2007, Pages 5-15, ISSN 1570-8268.
- MISLOVE, A. E. 2009. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*, Ph.D. Thesis, Rice University, Houston, Texas.
- MISLOVE, A. E., MARCON, M., GUMMADI, K. P., DRUSCHEL, P., AND BHATTACHARJEE, B. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on internet Measurement* (San Diego, California, USA, October 24 - 26, 2007). IMC '07. ACM, New York, NY, 29-42.
- MILGRAM, S. 1967. The small world problem. *Psychology Today* 2, 60.
- MORRIS, S. 2000. Contagion. *Review of Economic Studies* 67, 57-78.
- MURATA, T. AND MORIYASU, S. 2008. Link prediction based on structural properties of online social networks. *New Generation Computing* 26, 245-257.
- MYERSON, R. 1991. *Game Theory: Analysis of Conflict*, Harvard University Press: Cambridge, MA.
- NARAYANAM, R., AND NARAHARI, Y. 2011. Topologies of strategically formed social networks based on a generic value function—Allocation rule model. *Social Networks* 33, 1, 56-69.
- NELSENWIRE. 2009. Global Advertising: Consumers Trust Real Friends and Virtual Strangers the Most. *Nielsen Global Online Consumer Survey*, <http://blog.nielsen.com/nielsenwire/consumer/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most/>
- NEWMAN, M. E. J. AND GIRVAN, M. 2003. Mixing patterns and community structure in networks. In *Statistical Mechanics of Complex Networks*, R. PASTOR-SATORRAS, J. RUBI, AND A. DIAZ-GUILERA, Eds. Springer-Verlag, Berlin, Lecture Notes in Physics, 625/2003, 66-87, DOI: 10.1007/978-3-540-44943-0_5.
- NEWMAN, M. E. J. AND GIRVAN, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69, 026113.
- NEWMAN, M. E. J. AND PARK, J. 2003. Why social networks are different from other types of networks. *Physical Review E* 68, 036122.

- NEWMAN, M. E. J., FORREST, S. AND BALTHROP, J. 2002. Email Networks and the Spread of Computer Viruses. *Physical Review E* 66, 3, 035101-035104.
- NEWMAN, M. E. J. AND WATTS, D. J. 1999. Renormalization group analysis of the small-world network model, *Physics Letters A* 263, 341-346.
- NEWMAN, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Physical Review E* 69, 066133.
- NEWMAN, M. E. J. 2006a. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74, 3, 036104.
- NEWMAN, M. E. J. 2006b. Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences* 103, 23, 8577-8582.
- NEWMAN, M. E. J. 2008. The physics of networks, *Physics Today* 61, 11, 33-38.
- NIE, Z., ZHANG, Y., WEN, J. AND MA, W. 2005. Object-level ranking: bringing order to Web objects. In *Proceedings of the 14th international Conference on World Wide Web* (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM, New York, NY, 567-574.
- NORRIS, J. R. 1997. *Markov Chains*, Cambridge University Press, Cambridge.
- NOWICKI, K. AND SNIJDERS, T. A. B. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* 96, 455, 1077-1087.
- OH, J., SUSARLA, A. AND TAN, Y. 2008. Examining the diffusion of user-generated content in online social networks. *Social Science Research Network eLibrary*.
- O'MADADHAIN, J., HUTCHINS, J. AND SMYTH, P. 2005. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations Newsletter* 7, 2, 23-30.
- PALLA, G., BARABÁSI, A. L. AND VICSEK, T. 2007. Quantifying social group evolution. *Nature* 446, 664-667. doi:10.1038/nature05670.
- PALLA, G., DERÉNYI, I., FARKAS, I. AND VICSEK, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 7043, 814-818.
- PATAK, N., MANE, S., SRIVASTAVA, J. AND CONTRACTOR, N. 2007. Knowledge Perception Analysis in a Social Network. In *Proc. 5th Workshop on Link Analysis, Counterterrorism and Security, SIAM International Data Mining Conference*.
- PONS, P. 2006. Post-processing hierarchical community structures: quality improvements and multi-scale view. eprint, *arXiv:cs/0608050v1*.
- RADICCHI, F., CASTELLANO, C., CECCONI, F., LORETO, V. AND PARISI, D. 2004. Defining and identifying communities in networks. In *Proceedings of the National Academy of Sciences* 101, 9, 2658.
- REFFAY, C., AND CHANIER, T. 2003. How social network analysis can help to measure cohesion in collaborative distance-learning. Computer Supported Collaborative Learning. *Bergen : Kluwer Academic Publishers*. Retrieved Feb. 23, 2010, from <http://archive-edutice.ccsd.cnrs.fr/edutice-00000422>
- REICHARDT, J. AND BORNHOLDT, S. 2006. Statistical mechanics of community detection. *Phys. Rev. E* 74, 016110.
- RICHARDSON, M. AND DOMINGOS, P. 2002. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM, New York, NY, 61-70.
- ROBINS, G., SNIJDERS, T., WANG, P., HANDCOCK, M. AND PATTISON, P. 2007. Recent Developments in Exponential Random Graph (p^*) Models for Social Networks. *Social Networks* 29, 2, 192-215.
- ROGERS, E. M. 1995. *Diffusion of Innovations* (fourth edition). Simon & Schuster, New York.
- ROSVALL, M. AND BERGSTROM, C. T. 2008. Maps of random walks on complex networks reveal community structure. In *Proceedings of the National Academy of Sciences of the United States of America* 105, 4, 1118-1123.
- ROSVALL, M. AND BERGSTROM, C. T. 2011. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLOS ONE* 6, e18209.

- RUAN, J. AND ZHANG, W. 2008. Identifying network communities with a high resolution. *Physical Review E* 77, 016104.
- RUPNIK, J. 2006. Finding community structure in social network analysis – overview. In *IS 2006: Proceedings of the International multiconference on Information society*, Ljubljana, Slovenija.
- RYAN, B. AND GROSS, N. C. 1943. The diusion of hybrid seed corn in two Iowa communities, *Rural Sociology* 8, 15-24.
- SADILEK, A., KAUTZ, H., AND SILENZIO, V. 2012. Modeling spread of disease from social interactions. In *Proceedings of Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*.
- SCHIFANELLA, R., BARRAT, A., CATTUTO, C., MARKINES, B. AND MENCZER, F. 2010. Folks in Folksonomies: Social Link Prediction from Shared Metadata. In *Proceedings of the 3rd ACM International Conference on Web search and data mining*, NY, USA, 271–280.
- SCOTT, J. 2000. *Social network analysis: A handbook* (2nd edition), Sage, London.
- SCRIPPS, J., TAN, P., AND ESFAHANIAN, A. 2007. Exploration of Link Structure and Community-Based Node Roles in Network Analysis. In *Proceedings of the 2007 Seventh IEEE international Conference on Data Mining* (October 28 - 31, 2007). ICDM. IEEE Computer Society, Washington, DC, 649-654.
- SHALIZI, C. R., CAMPERI, M. F., AND KLINKNER, K. L. 2007. Discovering functional communities in dynamical networks. In *Proceedings of the 2006 Conference on Statistical Network Analysis* (Pittsburgh, PA, USA). E. AIROLDI, D. M. BLEI, S. E. FIENBERG, A. GOLDENBERG, AND E. P. XING, Eds. Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg, 140-157.
- SHEN, H., CHENG, X., CAI, K. AND HU, M. B. 2009. Detect overlapping and hierarchical community structure in networks. *Physica A* 388, 8, 1706-1712.
- SHI, J. AND MALIK, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8, 888-905.
- SIMON, H. A. 1962. The architecture of complexity. In *Proceedings of the American Philosophical Society* 106, 6, 467.
- SKYRMS, B., AND PEMANTLE, R. 2000. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences* 97,16, 9340-9346.
- SMYTH, P. 2003. Statistical modeling of graph and network data. In *Proceedings of IJCAI Workshop on Learning Statistical Models from Relational Data*, August 2003, Acapulco, Mexico.
- SNIJDERS, T.A.B. 2001. The statistical evaluation of social network dynamics. In *Sociological Methodology* 31, 1, M. SOBEL, AND M. BECKER, Eds. Blackwell Publishers Inc, 361–395.
- SNIJDERS, T. A. B. 2011. Statistical Models for Social Networks, *Annual Review of Sociology* 37, 129–51.
- SNIJDERS, T., PATTISON, P., ROBINS, G. AND HANDCOCK, M. 2006. New specifications for exponential random graph models. *Sociological Methodology* 36, 1, 99–153.
- SNIJDERS, T. A. B., VAN DE BUNT, G.G., STEGLICH, C. E.G. 2010. Introduction to stochastic actor-based models for network dynamics. *Social Networks: Dynamics of Social Networks* 32, 1 (January 2010), 44-60, ISSN 0378-8733, DOI: 10.1016/j.socnet.2009.02.004.
- SONG, X. D., CHI, Y., HINO, K. AND TSENG, B. L. 2007. Information Flow Modeling based on Diffusion Rate for Prediction and Ranking. In *Proceedings of the 16th International Conference on World Wide Web*, ACM, New York, 191-200.
- SPILOPOULOU, M., NTOUTSI, I., THEODORIDIS, Y., AND SCHULT, R. 2006. MONIC: modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Philadelphia, PA, USA, August 20 - 23, 2006). KDD '06. ACM, New York, NY, 706-711.
- SPILOPOULOU, M. 2011. Evolution in social networks: A survey. In *Social Network Data Analytics*, Springer US, 149-175.
- STAUFFER, D. AND AHARONY, A. 1991. Introduction to Percolation Theory, *Taylor&Francis*, London.
- STEWART, A., CHEN, L., PAIU, R. AND NEJDL, W. 2007. Discovering information diffusion paths from blogosphere for online advertising. In *ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, NY, USA, 46-54.

- STONEDAHL, F., RAND, W., AND WILENSKY, U. 2010. Evolving viral marketing strategies. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation* (Portland, Oregon, USA, July 07 - 11, 2010). GECCO '10. ACM, New York, NY, 1195-1202.
- STUMME, G., HOTH, A. AND BERENDT, B. 2006. Semantic Web Mining - State of the art and future directions. *Journal of Web Semantics* 4, 2, 124-143.
- SUN, H., HUANG, J., HAN, J., DENG, H., ZHAO, P. AND FENG, B. 2010. gskeletonclu: Density-based network clustering via structure-connected tree division or agglomeration. In *Proceedings of the 2010 IEEE International Conference on Data Mining, ser. ICDM '10*. Washington, DC, USA: IEEE Computer Society, 2010, pp. 481-490.
- SUN, J. AND TANG, J. 2011. A survey of models and algorithms for social influence analysis. *Social Network Data Analytics*, 177-214.
- SWAN, J., NEWELL S., SCARBOROUGH, H. AND HISLOP D. 1999. Knowledge management and innovation; networks and networking. *Journal of Knowledge Management* 3, 4, 262-275
- TANG, L. AND LIU, H. 2010. Understanding Group Structures and Properties in Social Media. In *Link Mining: Models, Algorithms, and Applications*, P. S. YU, et al. Eds. Part 2, 163-185, DOI: 10.1007/978-1-4419-6515-8_6.
- TANTIPATHANANANDH, C., BERGER-WOLF, T., AND KEMPE, D. 2007. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (San Jose, California, USA, August 12 - 15, 2007). KDD '07. ACM, New York, NY, 717-726.
- TANTIPATHANANANDH, C. AND BERGER-WOLF, T. 2009. Constant-factor approximation algorithms for identifying dynamic communities. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, NY, USA: ACM, 827-836.
- TASKAR, B., WONG, M. F., ABBEEL, P. AND KOLLER, D. 2003. Link prediction in relational data. In *NIPS '03 Advances in Neural Information Processing Systems 16*. S. THRUN, L. SAUL AND B. SCHÖLKOPF, Cambridge, MA: MIT Press.
- TERO, A., TAKAGI, S., SAIGUSA, T., ITO, K., BEBBER, D. P., FRICKER, M. D., YUMIKI, K., KOBAYASHI, R. AND NAKAGAKI, T. 2010. Rules for biologically inspired adaptive network design. *Science Signaling* 327, 5964, 439.
- TOIVONEN, R., KOVANEN, L., KIVELÄ, M., ONNELA, J. P., SARAMÄKI, J. AND KASKI, K. 2009. A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks* 31, 240-254.
- TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. AND PORTER, M. A. 2008. Community structure in online collegiate social networks. *arXiv:0809.0960*.
- TRUSOV, M., BODAPATI, A. V. AND BUCKLIN, R. E. 2010. Determining influential users in internet social networks. *Journal of Marketing Research* 47, 643-658.
- TRUSOV, M., BUCKLIN, R.E. AND PAUWELS, K. 2009. Effects of Word-of-Mouth Versus Traditional Marketing: Findings From an Internet Social Networking Site. *Journal of Marketing* 73, 5, 90-102
- VAN-DUIJN, M. A. J., SNIJDERS, T. A. B. AND ZIJLSTRA, B. J. H. 2004. p2: A random effects model with covariates for directed graphs. *Statistica Neerlandica* 58, 234-254.
- VALLAM, R. D., SUBRAMANIAN, C. A., NARAYANAM, R., NARAHARI, Y., AND NARASIMHA, S. 2011. Topologies and Price of Stability of Complex Strategic Networks with Localized Payoffs: Analytical and Simulation Studies. *arXiv preprint arXiv:1201.0067*.
- WANG, D, PEDRESCHI, D., SONG, C., GIANNOTTI, F. AND BARABASI., A. L. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*. ACM, New York, NY, USA, 1100-1108.
- WANG, X., MOHANTY, N. AND MCCALLUM, A. 2006. Group and topic discovery from relations and their attributes. In *NIPS'06: Proceedings of 12th Annual Conference on Neural Information Processing Systems*, Whistler, BC, Canada, 1449-1456.
- WANG, X., MOHANTY, N., AND MCCALLUM, A. 2005. Group and topic discovery from relations and text. In *Proceedings of the 3rd international Workshop on Link Discovery* (Chicago, Illinois, August 21 - 25, 2005). LinkKDD '05. ACM, New York, NY, 28-35.

- WANG, Y., WU, B. AND DU, N. 2008b. Community Evolution of Social Network: Feature, Algorithm and Model, *arXiv:0804.4356*.
- WASSERMAN, S. AND FAUST, K. 1994. *Social Network Analysis: Methods and Applications*, Cambridge University Press.
- WASSERMAN, S. AND PATTISON, P. 1996. Logit models and logistic regression for social networks: I. an introduction to markov graphs and p . *Psychometrika* 61, 3, 401-425.
- WASSERMAN, S. S. AND ROBINS, G. L. 2005. An Introduction to Random Graphs, Dependence Graphs, and p^* . In *Models and Methods in Social Network Analysis*, P.J. CARRINGTON, J. SCOTT, AND S. WASSERMAN, Eds. Cambridge University Press, New York, 148–161.
- WATTS, D. J. AND DODDS, P. S. 2007. Influentials, networks, and public opinion formation. *Journal of Consumer Research* 34, 4 (December), 441-458.
- WATTS, D. J. AND STROGATZ, S. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393, 440-442.
- WEI, F., WANG, C., MA, L., AND ZHOU, A. 2008. Detecting overlapping community structures in networks with global partition and local expansion. In *Proceedings of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development* (Shenyang, China, April 26 - 28, 2008). Y. ZHANG, G. XU, G. YU, AND E. BERTINO, Eds. Lecture Notes In Computer Science. Springer-Verlag, Berlin, Heidelberg, 43-55.
- WHITE, H. C., BOORMAN, S. A. AND BREIGER, R. L. 1976. Social structure from multiple networks. I. blockmodels of roles and positions. *American Journal of Sociology* 81, 4, 730.
- WOLFE, A. P. AND JENSEN, D. 2004. Playing multiple roles: Discovering overlapping roles in social networks. In *ICML'04: Workshop on Statistical Relational Learning and its Connections to Other Fields*, Banff, Alberta, Canada.
- WONG, L. H., PATTISON, P. AND ROBINS, G. 2006. A spatial model for social networks. *Physica A* 360, 1, 99.
- WU, L., WABER, B., ARAL, S., BRYNJOLFSSON, E. AND PENTLAND, A. 2008. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. In *Proceedings of the International Conference on Information Systems*, Paris, France.
- WU, X., ZHANG, L. AND YU, Y. 2006. Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web (WWW '06)*. ACM, New York, NY, USA, 417-426.
- WUYTS, S., DEKIMPE, M. G., GIJSBRECHTS, E. AND PIETERS, R. 2010. *The Connected Customer: The Changing Nature of Consumer and Business Markets*. Routledge Academic, New York.
- XU, X., YURUK, N., FENG, Z. AND SCHWEIGER, T. A. J. 2007. SCAN: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07)*. ACM, 2007, pp. 824–833.
- YADATI, N. AND NARAYANAM, R. 2011. Game theoretic models for social network analysis. In *Proceedings of the 20th international conference companion on World wide web (WWW '11)*. ACM, New York, NY, USA, 291-292.
- YANG, J. AND LESKOVEC, J. 2010. Modeling Information Diffusion in Implicit Networks. In *ICDM'10 Proceedings of IEEE 10th International Conference on Data Mining*, 599-608.
- ZACHARY, W. W. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33, 452-473.
- ZHAO, J., WU, J. AND XU, K. 2010. Weak Ties: A Subtle Role in the Information Diffusion of Online Social Networks. *Physical Review E* 82, 1, 016105, DOI: 10.1103/PhysRevE.82.016105
- ZHANG, S., WANG, R. AND ZHANG, X. 2007. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A* 374, 1, 483-490.
- ZHELEVA, E., GETOOR, L., GOLBECK, J. AND KUTER, U. 2010. Using Friendship Ties and Family Circles for Link Prediction, In *Advances in Social Network Mining and Analysis*, Lecture Notes in Computer Science, 5498/2010, 97-113, DOI: 10.1007/978-3-642-14929-0_6