

A Unified Framework for Community Structure Analysis in Dynamic Social Networks

Sajid Yousuf Bhat and Muhammad Abulaish

Abstract One of the major tasks related to the structural social network analysis is the detection and analysis of community structures, which is highly challenging due to consideration of various constraints while defining a community. For example, community structures to be detected may be disjoint, overlapping, or hierarchical, whereas on the other hand, community detection methods may vary depending on whether links are weighted or unweighted, directed or undirected, and whether the network is static or dynamic. Though a number of community detection methods exists in literature, most of them address only a particular aspect of the community structures, and generally community structures and their evolution analysis are studied separately. However, analyzing community structures and their evolution under a unified framework could be more useful, where community structures provide evidence about community evolution. Moreover, not many researchers have dealt with the issue of the utilization of detected communities and they have simply proposed methods to detect communities without emphasizing their utilities. This chapter presents a unified social network analysis framework which mainly aims to address the problem of community analysis, including overlapping community detection, community evolution tracking, and hierarchical community structure identification in a unified manner. At the end of this chapter, we present some important application areas wherein the knowledge of community structures facilitates the generation of important analytical results and processes to help solving other social network analysis related problems. The application areas mainly include to deal with spammers wherein we present the importance of various community-based features to learn predictive models for spammer detection in online social networks. In addition, we address the issue of detecting deceptions in online social networks which

Sajid Yousuf Bhat
Department of Computer Sciences, University of Kashmir, Srinagar, India.
e-mail: s.yousuf.jmi@gmail.com

Muhammad Abulaish (*corresponding author*)
Department of Computer Science, South Asian University, New Delhi, India
e-mail: abulaish@ieee.org

mainly includes dealing with cloning attacks, and the importance of community detection for facilitating the process of viral marketing in online social networks.

Key words: Social network analysis, Community detection, Hierarchical and overlapping community, Community evolution tracking, Community analysis applications.

1 Introduction

Social networks resemble as graph structures and aim to model relationships/ties reflecting associations, friendship, hyperlinks, financial exchange, co-authorship, citations, interactions and so on between social actors like persons, web pages, research articles, financial accounts, airports and so on [28]. Social networks have been an interest of study since mid 1930s starting with the introduction of *Sociometry* by Moreno [22], and later in 1950s, the applications of graph theory started becoming popular in sociological community [27]. Some of the most important characteristics of social networks that motivated for significant interest and developments in the field of social network analysis are listed below:

Power law and preferential attachment: Researchers like Barabasi and Albert [2] revealed that unlike random networks, the degree distribution of nodes in a large-scale real-world networks follows a scale-free power law, making some nodes to have relatively higher degree (forming hubs) with majority of node having lesser connections. Moreover, the probability of newly joined nodes to attach (form links) to nodes of higher degree is more than to nodes of lower degree, thus making the growth of such networks to follow a rich-get-richer scheme (aka preferential attachment) [3].

Assortativity and network clustering: Newman and Park [23] analyzed that identical nodes in terms of degrees in a social network tend to show more connectivity with each other than with others, thus highlighting positive correlations among the degrees of adjacent vertices (assortativity). Moreover, social networks have non-trivial clustering of nodes or network transitivity, making them to exhibit communities, i.e., nodes within a group have high similarity or connectivity than the remaining nodes of the network.

Small-world behavior and the strength of weak ties: Milgram [21] popularized the concept of the *six-degrees-of-separation* based on the analysis that the average path length between any two individuals in social world is six hops, thus reflecting the small-world behavior of the social networks. *Weak ties* hypothesis proposed by Granovetter [10] is another famous concept of sociology, which reflects that the degree of neighborhood overlap between two individuals increases in proportion to the strength of their tie. This reveals that strong ties can play an important role to form tight clusters, whereas weak ties play an important role for the flow of information and innovation in the network, which is often termed as the *strength of weak ties*.

Based on the characteristics mentioned above, Social Network Analysis (SNA) has its root in various disciplines, including sociology, anthropology, economics, biology, communication, and social computing. The process of SNA mainly focuses on understanding the behavior of individuals and their ties in social networks that translate them into large-scale social systems. Due to advent of Web2.0 and existence of many online social networks, the application of SNA has recently gained much popularity due to its ability to model and analyze various processes taking place in society, such as recommendations, spread of cultural fads or diseases, community formation and so on. Moreover, existence of huge amount of both structural and non-structural data (aka user-generated contents) provides an opportunity to analyze them for varied purposes, including open-source intelligence, target marketing, business intelligence, and so on. Besides traditional online social networks, a number of complex networks, such as co-authorship networks, citation networks, protein interactions networks, metabolic pathways, and gene regulatory networks also posse similar patterns that can be analyzed using SNA techniques. Keeping in view the huge amount of data associated with social networks, information extraction, data mining, and NLP techniques have emerged as key techniques for analyzing social networks at different levels of granularity. As a result, a significant increase has been seen in the research literatures at the intersection of computational techniques and social sciences. The exponential growth of online social networks and the large-scale complex networks induced by them provide unique challenges and opportunities in the area of social network analysis that can be broadly classified into *structural*, *non-structural*, and *hybrid* categories, depending on the nature of data being used for analysis. Figure 1 presents a categorization of SNA tasks that are briefly described in the following sub-sections.

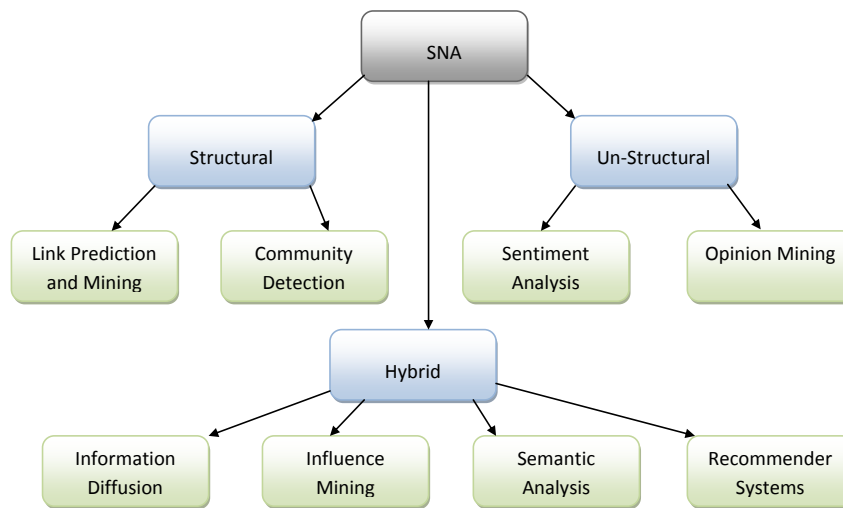


Fig. 1 Classification of social network analysis tasks.

1.1 Structural Social Network Analysis

Structural social network analysis mainly aims to analyze and mine structural (aka topological) data, such as link structure of the WWW, social network users interaction and friendship networks, citation networks and so on. Numerous graph mining techniques have been used for the analysis of such networks that reveal important patterns and characteristics of human social behavior that are often highly surprising. One of the most primitive and the challenging tasks in this domain is *link mining*. Introduced by Getoor [9], link mining is a data mining task mainly dealing with data related to social networks which tends to be heterogeneous, multi-relational, or semi-structured. Link mining is an emerging area that explicitly models links between data instances to find non-trivial interesting patterns in the data. This field of research encompasses a range of tasks including descriptive and predictive modeling, e.g., estimating the number of links in a network, predicting the relation type between two entities, estimating the likelihood of a link's existence and so on. Motivated by the dynamics and evolving nature of social networks, a sub-challenge sprouting from the domain of link mining is the problem of *link prediction*, which mainly aims to predict possible future links, given the present state of the nodes and links in a social network. Heterogeneity of real-world social networks which involves different node types, link attributes, weights and directions, friend and foe relations and so on pose a major challenge to link prediction. Another big challenge is the prediction of links that could be induced in a social network when new nodes are added. This is referred to as cold start link prediction problem, and it has received less attention till now.

One of the big challenges related to structural SNA that has received much attention is the problem of community detection, which represents the existence of groups of densely connected nodes. For example, employees working on same projects in an organization may communicate frequently; people calling each other frequently in a telecommunication network are usually friends, and so on. Discovery of such groups, clusters, or communities in a social network forms the basis for many other social network analysis tasks. Community detection in social networks is a challenging task and it is often viewed from different aspects as listed below.

- Local and global community detection
- Detection of overlapping communities
- Detection of hierarchical community structures
- Utilization of link weights and directions
- Community evolution tracking in dynamic networks

1.2 Non-Structural Social Network Analysis

Non-structural social network analysis is the domain which involves analytical deductions and pattern recognition primarily from the *user-generated contents* (in-

cluding texts, images, videos and so on wherein the structural or topological notions are vague) in online social networks. Traditionally content analysis refers to the systematic review of a body of texts, images and other symbolic matters with an aim to make valid inferences regarding their contextual usability [17]. Content analysis has been used as a research method for both qualitative and quantitative analyses of information in many fields wherein the main task is to extract new facts, patterns, information and also to determine their application domain [17]. The OSNs can be viewed as unstructured-content oceans wherein the process of generating new user content in the form of blogs, posts, comments, tags, messages, chats and so on never stops. We have seen how sometimes user initiated content-storms occur in OSNs and cause revolutions (e.g., the 2011 Egyptian revolution). This calls for developing methods and techniques for analyzing user-generated textual contents, which can also aid in dealing with the challenges discussed earlier, i.e., link prediction (users who generate or consume similar or related contents concerning specific topics are more probable to qualify for a missing link or a link in future), community detection (group of content/textual objects that are related to the same topic subject domain or group of individuals generating and consuming/propagating similar/related content).

With the popularity of the Web and its services, opinions regarding products and services are often expressed by people through formal or informal textual representations using e-mails, online reviews, and blogs. Consequently, the area of opinion mining and sentiment analysis, which is considered as a mutual collaboration of natural language processing, computational linguistics and text mining has emerged as a result for automatic identification and extraction of attitudes, opinions and sentiments from the voluminous amounts of user-generated contents. The motivation here is mainly driven by the commercial demand of cheap, detailed, and timely customer feedback to business organizations. Recent interest in developing such automated systems has swelled in order to assist the information analysts of various organizations in decision making and answering questions like: “how do people feel about the latest camera phone?”. In a broader view of Liu et al. [19], the task of opinion mining can be stated as determining the orientation of a sentiment related to a set of objects, expressed in terms of opinions on a set of attributes defining the objects, by a set of opinion holders. According to Liu et al. [19] some important technical challenges related to opinion mining in online social networks are:

- Object identification: The basic requirement is to identify the relevant object which forms a subject for an opinion holder to express opinion in a statement.
- Feature extraction: This involves identifying the object attributes about which opinion holders express their opinions. Sometimes the task is more challenging as an opinion holder may use alternative words or phrases to refer to a particular feature of an object in a subjective statement. This requires finding the groups of synonyms describing a particular feature.
- Polarity determination: This involves subjectivity classification of a statement as discussed earlier to determine if a statement contains an opinion and if an opinion exists, then also determining its polarity.

- **Integration:** This highlights the challenge of mapping the challenges mentioned above with each other, i.e., mapping together the orientation of an opinion, projected by the features extracted about a particular object, expressed by a particular opinion holder. It is challenging as not all information is always explicit in a statement.

1.3 Hybrid Social Network Analysis

With hybrid social network analysis, we aim to address all those processes and frameworks that aim to utilize both structural and content related information of online social networks to extract useful patterns and make meaningful deductions. Some of the most popular hybrid SNA tasks that have received much of the attention are explained in the following paragraphs:

Information Diffusion: The area of information diffusion deals with the analysis and modeling of information flow in social networks with an aim of explaining the spreading processes of various entities like news, diseases, computer viruses and so on in the real social world. With the advent of modern communication tools like instant messaging apps (WhatsApp, Skype), online social networks (Facebook, Twitter), etc., efficient measurement of social contagion is proposed by many researchers, resulting in effective models for information diffusion. The works of Kato et al. [15] and Bakshy et al. [1] have significantly thrown some light on the importance of the topological social network properties, discussed earlier in this section, and the user-generated content for modeling and analyzing the information diffusion in online social networks.

Influence Analysis: Certain business processes, mainly marketing, aim to analyze the influence of an individual over a population with an aim of utilizing highly influential people to endorse products. For example, in line with the spread of computer or pathological viruses through self-replicating process, viral marketing exploits social contagion and influence ranking to promote brand awareness or to do advertisements. In literature, influence analysis has been mainly performed following a hybrid model, i.e., by incorporating both the content and network structure for analysis.

Semantic Analysis: The amalgamation of the Semantic Web framework and social media platform is looked upon as the next step towards constructing collective intelligence and knowledge systems for effective information retrieval and efficient content search [12]. One of the most popular ontologies used to represent the network of people is the *friend-of-a-friend* (FOAF) that describes people, their relationships and actions. Current state-of-the-art related to Semantic Web attempts automatic construction of vocabulary models from knowledge sources existing in the form of natural texts, based on the observation that online social network users create contextual ontologies through collective intelligence [20]. Such ontologies can aid in the annotation of user-generated contents within their communities based

on the fact that online social network services also allow users to annotate their published resources through short descriptive tags.

Recommender Systems: Recommender systems are special information filtering systems that aim to reduce the domain of object instances (books, videos, friends, news, events, research papers and so on) based on the level of interest (automatically determined or explicitly mentioned) shown by users towards them. Introduced in the mid-1990's and based on the concept of collaborative filtering, the recommendation problem is formulated as the problem of estimating ratings (predicting the level of interest a user can show on a particular item) for the items that have not been seen by the users [14]. The problem of designing recommender systems aims to develop systems and techniques based on the concept of social networks that assist and augment the natural social process of recommendations along with determining resources (previously unseen) that may match an individual's interests. Various problems related to exploiting OSNs for effective recommendations have been identified by the researchers in [13]. Some of them are briefly described in the following paragraphs:

- How can the merger of social networks and recommendations help to increase consistent user participation and contribution? Is it possible to encourage like-minded users through user matching via collaborative filtering [8]?
- Is it possible to increase trust in recommender systems through utilizing the various aspects of social networks such as published friend relations in OSNs (preferred recommenders) [8]? Do friends share preferences, and if they do, how can such information be exploited by recommender systems [13]?
- Is it possible to tackle the cold-start problem through explicitly specifying the users' closest neighbors [8]?
- How weak and strong ties among users can be useful for search and recommendation purposes?

The field of SNA is very vast and we would like to stress that a single book or paper cannot span it justly. In this regard, this chapter stresses on community detection and presents a unified framework that aims to facilitate other SNA-related tasks.

2 Community Detection

The task of community detection involves identifying groups (communities) through the study of network structures and topologies. As discussed in Section 1.1, community detection is highly challenging as it involves numerous aspects that are briefly summarized in the following paragraphs.

Overlapping and hierarchical communities: Real-world community structures show overlapping behavior, i.e., nodes may belong to multiple communities. For example, in a social network, a person may be a member of multiple work groups,

and thus it may be significant to identify all such overlapping memberships. In literature, however, most of the community detection approaches tend to find mutually exclusive community structures. It is thus highly desirable to devise approaches for identifying overlapping community structures. Besides showing overlapping behavior, community structures may possess a hierarchical structure at different resolutions wherein a larger community contains multiple smaller communities. Very few methods in literature address these two issues collectively, leading to the need of further research to fill this gap.

Dynamic communities and their evolution: An important property of the real-world social networks lies in their tendency to change dynamically, i.e., addition/deletion of new nodes and links, change in intensity of interactions. These changes directly affect the structure of underlying communities synchronously resulting in *birth, growth, contraction, merge, split, and death* of communities with time. Some of the methods that have been proposed to identify the evolution of communities with time include [11], [26], and [16]. However, Lin et al. [18] pointed out that these methods have a common limitation wherein communities and their evolution have been studied separately.

Utilities: Not many research works deal with the issue of utilization of detected communities and simply propose a method to detect them without emphasizing on what to do with them later. Communities in social networks map to functional modules like functionally coherent groups of proteins in protein interaction networks, discussion groups in online social network, and so on. Community structure of a network can also be used to determine the roles of individual members, wherein boundary nodes can be used to facilitate interactions between two or more community members, whereas the central nodes may play an important role for the control and stability of the community. Knowing an underlying community structure of a communication network can be used for spammer identification in online social networks by characterizing the interaction behavior of legitimate users within the identified communities. Spammer interaction can then be filtered out or blocked based on the source and target communities of the interaction.

3 A Unified Community Analysis Framework

We have presented a unified community detection and evolution analysis framework, HOCTracker, in [7], which is derived from density-based clustering approach, wherein a cluster is searched by detecting the neighborhood of each object in the underlying database. For any node, its neighborhood is defined in terms of its distance with other nodes in the database. A node q is assigned to the neighborhood of another node p only if the distance between them is less than or equal to a threshold ϵ . In network context, this distance constraint (which is usually a structural metric) is checked only for those nodes that have a link/edge between them. Now, depending upon whether the neighborhood of a node contains more than μ nodes, a new cluster with p as a core object is formed. Building upon the notion of these

core nodes, a density-connected cluster is identified as a maximal set of density-connected nodes. Density-based clustering and community detection methods are relatively fast and easily scalable to large databases/networks with simplified parallel implementations. However, density-based methods require two input parameters which include the global neighborhood threshold, ϵ , and the minimum cluster size, μ towards which they are highly sensitive. Estimating an optimal value for the ϵ parameter automatically is a long-standing challenge [24]. HOCTracker on the other hand automatically estimates the value for ϵ based on the local neighborhood of each node. A similar approach followed for μ uses a resolution parameter η tuned as required or estimated using a heuristic approach given in [7].

Figure 2 presents a conceptual overview of the proposed social network analysis framework, which consists of three overlapping but inter-related functioning modules, each addressing one particular aspect of community analysis – *overlapping community detection*, *community evolution tracking*, and *hierarchical community structure identification*.

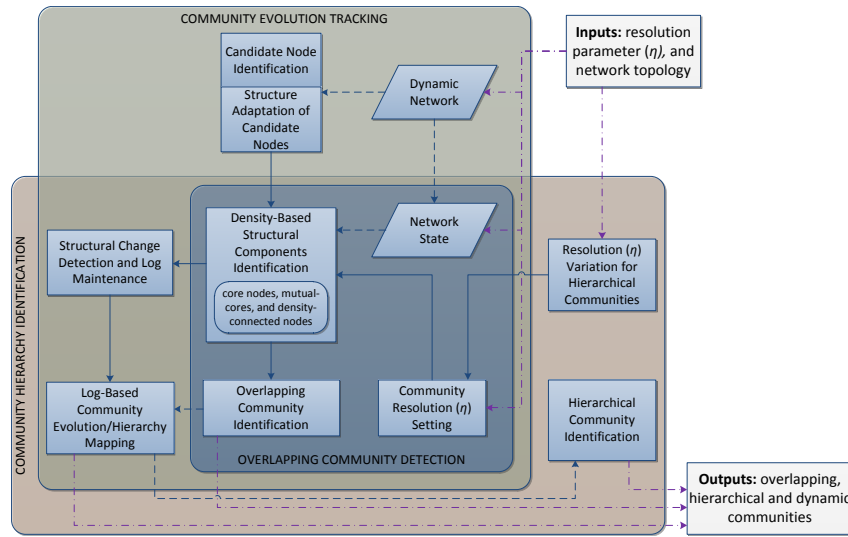


Fig. 2 A unified community analysis framework

The core constituent of all these modules is the identification of density-based structural components. For a given value of the resolution parameter (η), the basic overlapping community detection process involves identifying overlapping density-based components from the given state of a network. It follows an iterative and local expansion approach to directly expand a community from a randomly selected seed node until all nodes in the network are visited. It allows a node to be a member of multiple communities and thus identifies overlapping communities for a social network.

The evolution detection and tracking module adapts a reference community structure, identified at any initial stage of a dynamic network, to the changes occurring in the underlying network. This involves re-processing of only active node neighborhoods in the network which is different from other methods that require re-processing neighborhoods of all nodes. The mapping of evolutionary relationships between communities across consecutive timesteps is simplified by using an efficient log-based approach.

The hierarchical community detection module aims to generalize the overlapping community detection and evolution tracking process to identify and map the hierarchical relations between overlapping communities identified at varying resolutions by finding density-based structural components at different values of the resolution parameter (η).

3.1 Distance Function

Although the complete details of the method can be seen in our research paper [7], here we briefly present the novel dual-layer distance function used by HOCTracker, incorporating both directed and weighted attributes of the edges, if available. A brief description of the distance function is given in the following paragraphs:

Layer-1: For two directly connected nodes p and q , let V_p and V_q be the sets of nodes to which nodes p and q have out-links respectively, and let V_{pq} be the set of nodes to which both p and q have out-links. The *layer-1* distance is defined as shown in equation 1. It can be seen in equation 1 that the function ensures that the distance is further calculated by *layer-2*, i.e., $d(p, q)$ only if the number of common recipients of nodes p and q is more than some fraction (specified by η) of the minimum of the two. Otherwise, it is taken as 1 (maximum). At this point we can note that the fraction η (in the range $0 < \eta \leq 1$) is the required input parameter to specify the resolution of communities to be identified.

$$dist(p, q) = \begin{cases} d(p, q) & \text{if } |V_{pq}| > (\eta \times \min(|V_p|, |V_q|)) - 1 \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

Layer-2: The second layer of the distance function is based on the reciprocity of interactions between the two nodes including their common neighborhoods. The *layer-2* distance function is given by equation 2 wherein $I_{\overleftrightarrow{pq}}$ is the amount of reciprocated interactions between nodes p and q taken as the minimum of the interactions from p to q and vice-versa.

$$\delta(p, q) = \begin{cases} \left(\frac{\sum_{s \in V_{pq}} (I_{\overleftrightarrow{ps}}) + I_{\overleftrightarrow{pq}}}{|V_{pq}| + 1} \right) & \text{if } I_{\overleftrightarrow{pq}} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

In order to get a symmetric distance, $d(p, q)$ is taken as the maximum of their mutual directed-response (or *minimum of the reciprocals of their mutual directed-response*) values normalized by their respective total weight of outgoing interactions (represented by $I_{\vec{p}}$ and $I_{\vec{q}}$ respectively) in the interaction graph, as given in equation 3.

$$d(p, q) = \begin{cases} \min\left(\frac{\delta(p, q)^{-1}}{I_{\vec{p}}}, \frac{\delta(q, p)^{-1}}{I_{\vec{q}}}\right) & \text{if } \delta(p, q) > 0 \wedge \delta(q, p) > 0 \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

The dual layer distance function thus measures the maximum average reciprocity between two nodes and their common neighbors only if the overlap of their neighbors is significant (specified by η). Smaller values for $d(p, q)$ represent higher response between nodes p and q and translates to more closeness between p and q (less closeness for higher values).

As the proposed method adapts a density-based approach, a neighborhood threshold needs to be specified for marking the boundary of the smallest cluster possible. Unlike other density-based methods, HOCTracker automatically determines (for any node) a local version of the neighborhood threshold called the *local-neighborhood threshold* represented by ε . For a node p the value for ε_p is calculated as the average of reciprocated interactions of p with all its out-linked neighbors and defined using equation 4.

$$\varepsilon_p = \begin{cases} \frac{\left(\frac{I_{\leftrightarrow p}}{|V_p|}\right)^{-1}}{I_{\vec{p}}} & \text{if } |V_p| > 0 \wedge I_{\leftrightarrow p} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

In the above equation $I_{\leftrightarrow p}$ is the amount of reciprocated interactions for node p (i.e., $\sum_{\forall q \in V_p} \min(I_{\vec{p}q}, I_{\vec{q}p})$), and $\frac{I_{\leftrightarrow p}}{|V_p|}$ is the average reciprocated interactions between node p and all other nodes in V to which p has out-links. The denominator $I_{\vec{p}}$ normalizes the value of ε_p in the range $[0, 1]$ and represents the total out-link weights of node p in the interaction graph of the social network.

Based on the above definitions of distance function $dist(p, q)$ and threshold ε_p , the *local ε_p -neighborhood* of a node p , represented by N_p , consists of a subset of V_p with which p 's distance is less than or equal to ε_p , and it is defined formally by equation 5.

$$N_p = \{q : q \in V_p \wedge dist(p, q) \leq \varepsilon_p\} \quad (5)$$

Now, in order to define the notion of a density-based community in a social network at a given resolution fraction (η), the proposed approach uses the following key definition.

Definition 1 (Core node). A node p with non-zero reciprocated interactions with any of its neighbor(s) in V_p is designated as a core node if its local ε_p -neighborhood contains at least μ_p (local minimum-number-of-points threshold for p) of nodes in V_p , as given in equation 6, where $\mu_p = \eta \times |V_p|$.

$$CORE_\eta(p) \Leftrightarrow |N_p| \geq \mu_p \quad (6)$$

3.2 Community Detection and Tracking

To start the process of community detection, HOCTracker selects a random node p and computes the local ε_p threshold using equation 4. If the value of ε_p is greater than zero, the dual-layer distance function $dist(p, q)$ mentioned in Section 3.1 is used to identify the N_p of p . Thereafter, depending upon the result of the core node test given in equation 6, the following procedure is used to find the communities:

1. If node p forms a core-node, the set of visited nodes V with which p has mutual-core relations and the set U of un-visited nodes in N_p are determined.
2. If V is empty, a new community is formed by including p and all nodes in N_p . Moreover, the community memberships of all the nodes in this newly formed community are appended with a new community-ID C , and node p forms a primary core of community C . If N_p includes some core nodes which are not in the set V , they form the secondary core nodes of C .
3. If V is non-empty, and all the core nodes in V are primary cores of a single community C , then p joins and also forms a primary core of C . The community memberships of all the nodes in N_p including p are appended with label C .
4. If V is non-empty and some core nodes in V are primary cores of different communities, then the communities are merged to form a single community. A new community ID replaces the community membership of the merged communities for all visited density-reachable nodes of p , and the primary nodes of merged communities including p form the primary nodes of the new community.
5. Mark node p as visited.
6. For each node q in U , if q is not marked as *waiting* then mark q as *waiting* and add q to the queue.
7. Repeat these steps for each node removed from the queue until the queue is empty.

If the randomly selected node p does not qualify as a core, it is marked visited and may be added as a non-core to the neighborhood of some other core. Otherwise, it is treated as an outlier. The above process is iteratively repeated for each new randomly selected node until all nodes in the network are visited. This process finally identifies an overlapping community structure from an underlying network.

In order to track and map the evolution of communities in a dynamic network at two successive time-steps, HOCTracker uses a log called *intermediate evolution log(IEL)* to record the intermediate transitions occurring in the communities after processing the neighborhoods of active nodes. All intermediate community transitions caused due to re-computing the local ε -neighborhood of candidate nodes at a new timestep result in entry into the IEL. The information thus stored in the log is then used to deduce a bipartite mapping between the communities at different time-steps efficiently.

We have already mentioned that HOCTracker requires a single parameter η to be specified for identifying communities at a particular resolution. The value of η , although inversely proportional to the size of communities identified, can be easily

tuned to detect communities at different levels of size characteristic forming a hierarchical representation of overlapping community structures. For establishing hierarchical relationships between communities identified at different levels (different values of η) HOCTracker uses the same log-based evolutionary mapping technique discussed earlier. Communities at a higher level can be viewed as resulting from the *merge* and *growth* of communities at lower levels. Similarly, community structure at a lower level can be considered to result from the *death*, *split*, and *shrinkage* of communities at a higher level. This forms a unified framework for detecting and tracking overlapping and hierarchical communities in dynamic social networks.

For a dynamic network, three values of η are estimated for each new state of a network by following a heuristic. Before community structure is identified for a new level, community structure and state of evolution log for the previous level are saved. If the modularity score for next level is less than previous, the saved state is reloaded to represent communities for current state of the network and used for mapping evolutionary relations. A simple illustration of community detection and tracking by HOCTracker on a dynamic network involving 3 time-steps is shown in figure 3.

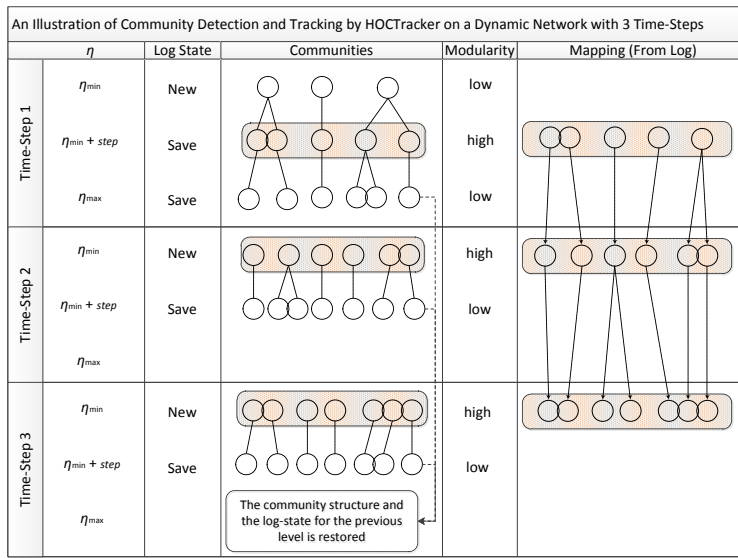


Fig. 3 An illustration of community detection and tracking by HOCTracker on a dynamic network with 3 time-steps

4 Applications of Community Structures

As mentioned earlier, community structure within a network represents the existence of groups of densely connected nodes with only sparse connections between the groups. For example, employees that work on related projects in an organization may communicate frequently via email; people that call each other more often in a telecommunication network are usually friends, and so on. Discovery of such groups, clusters, or communities in a social network forms the basis for many other social network analysis tasks and finds its application in numerous domains. For example, communities in social networks map to functional modules like functionally coherent groups of proteins in protein interaction networks, discussion groups in online social network, and so on. However, not many research works deal with the issue of utilization of detected communities and simply propose a method to detect them without emphasizing on what to do with them later. In this section, we present some novel application domains and methods for utilization of communities detected by community detection methods. These application domains include spammer detection in online social networks, detection of copy profiling attacks, and identifying influential nodes for viral marketing.

4.1 *Online Spammer Detection*

One of the most spiteful activities performed in online social networks is spamming, where a spammer broadcasts his/her desired content to as many benign users as possible through e-mails, instant messages, comments, posts and so on. The main aim of such an activity is the promotion of products, viral marketing, spreading fads and fake news, and sometimes to harass other users. In order to tackle the nuance of spamming, devising methods for automatic detection of spammers and their spamming activity is a highly desirable task. Detected spammers can be removed from social networks, and future spamming activities can be significantly controlled by analyzing spammers' behavior and taking precautionary measures.

In literature, numerous spammer detection methods have been proposed using content analysis, mainly involving keywords-based filtering. However, spammers have developed many counter-filtering strategies incorporating non-dictionary words and images in the spam content. The most common limitations of content-based spam filtering systems include need of high computations and the issue of users' privacy. The need of access to users' private messages, posts, profile details, etc. is often considered as a negative aspect for the content-based spam filtering systems. Recently, spammer detection methods have utilized classification models for learning rules from social network based topological features, including in-degree, out-degree, reciprocity, clustering coefficient and so on [25]. Along these lines, Bhat and Abulaish [4] proposed a community-based spammer detection method for OSNs by proposing novel community-based node features that can be extracted

from the overlapping community structure of the underlying OSN. For each node, the community-based features that they aim to extract include the following:

- Whether or not the node is a *core-node* (as discussed in the previous section).
- The number of participating communities for a node.
- The number of nodes belonging to other communities to which it has outlinks.
- The ratio between the number of its in-links from outside of its community to that of its out-links.
- The probability of a node to have an out-link outside of its community.
- The reciprocity ration with the nodes outside of its community.
- The probability that the non-member out-linked nodes belong to a common foreign community.

Based on these features, Bhat and Abulaish [4] proposed a classification framework to identify the spammers in OSNs, as shown in figure 4. The evaluation of the proposed spammer classification system over a Facebook and Enron datasets reveal that the novel community-based features are significant for identifying spammers in OSNs.

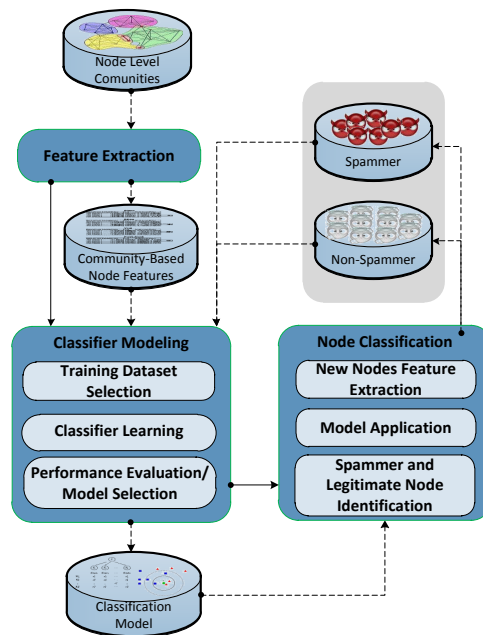


Fig. 4 Spammer classification model learning and application

4.2 Detecting Copy-Profiling/Cloning Attacks

One of the most deceptive forms of malicious attacks seen over online social networking platforms now a days is the *cloning attack*, which is different from the traditional spammer attacks and is often difficult to detect and track. A cloning attack (aka copy-profiling attack) involves an attacker to create clones (similar accounts) of other benign accounts (targets) by copying and replicating their profile details. Then the attacker sends friend requests to some of the friend accounts of the target account with a hope that some non-suspicious account users may get deceived and accept the friend request from the clone. The deception becomes possible because some recipients consider the friend request to be coming from the actual benign user cloned by the attacker, without realizing that a similar account is already befriended by them. This initial befriending with some benign accounts forms the first stage of the cloning attack. After this stage, it becomes easier to breach the social circle and trust of users who now share friends with the clone [6]. Figure 5 illustrates a two-stage clone. Here, 5(a) shows the first stage wherein an attacker clones a benign user (Ben) and multicasts friend requests to some of the Ben's friends, out of which a few accept.

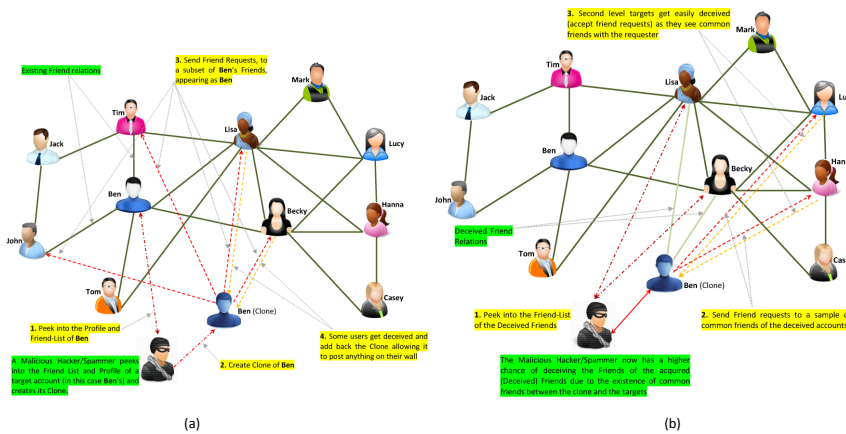


Fig. 5 Illustration of a cloning attack in 2 stages

The second stage of the attack is illustrated in figure 5(b), wherein the attacker selects some of those accounts with which the clone has common friends. In order to maintain stealth and increase the level of infiltration, a cloning attack in an advanced form may involve creating multiple clones of different accounts at each subsequent infiltration stage. Once an attacker has infiltrated to a significant level of its requirements, it has multiple options to exploit the privileges by performing malicious activities like posting spam messages, launching phishing attacks and so on.

For each cloned account an attacker follows a greedy strategy whereby it tries to maximize the number of benign accounts befriended while minimizing the number of friend requests broadcasted as illustrated in figure 5. In a bid to deceive any detection, the attacker also tries to fetch a moderate number of friends for each clone and links these clones with each other to mimic legitimate behavior. However, a general behavior observed for legitimate users in online social networks is the formation of communities whereby similar users tend to form separate overlapping clusters. In case of dynamic networks, the frequency of merging and splitting of communities is not frequent, whereas the birth, death, and growth of communities are observed more frequently. This behavior is in contrast to that of the cloned accounts of an attacker which result in dense regions in the network due to addition of clones and cross-links. A robust cloning behavior of an attacker results in frequent merging of communities in the underlying social network, as demonstrated in figure 6.

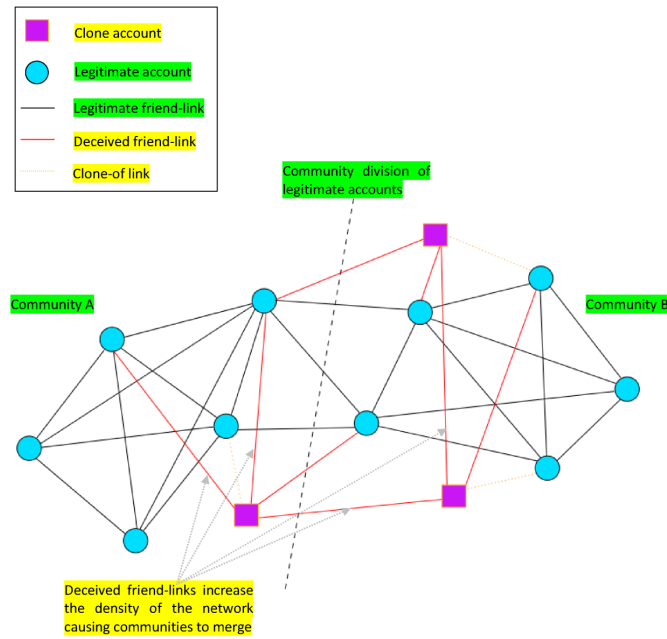


Fig. 6 Illustration of the robust cloning behavior of an attacker

Based on the above observation, it can be argued that a community detection and tracking method like the one discussed in Section 3 can detect and track cloning attacks by detecting regions facing frequent community merge events in the underlying network. Such regions can be explored by conducting profile similarity tests and neighborhood overlap analysis and those with high similarity can be considered as a clone.

4.3 Influential Node Detection for Viral Marketing

Viral Marketing (VM) is a strategy of exploiting a self-replicating process like that of viruses with an aim of increasing product/brand visibility and product sales [5]. The main process involves identifying few nodes in a social network that have high influence and developing promotion strategies for that smaller set. So, one of the preliminary tasks for a viral marketing campaign is the identification of influential nodes, and it is generally a challenging task. In this regard, Bhat and Abulaish [5] suggested that the nodes that are shared by overlapping communities are good candidates to be considered as influential. In addition, density-based community detection methods, as discussed in Section 3, identify hubs that have similar properties to that of overlapping nodes and can also form good seed nodes for a VM campaign. Based on these assumptions, Bhat and Abulaish [5] tested their hypothesis on an email dataset and argued that overlapping nodes are among the best candidates to be considered as influential nodes. As shown in figure 7, their analysis reveals that an overlapping node has a higher chance to appear in the top 1% and 5% of influential nodes in a social network, ranked according to their betweenness centrality. Thus, strengthening the argument that overlapping nodes are highly influential in social networks.

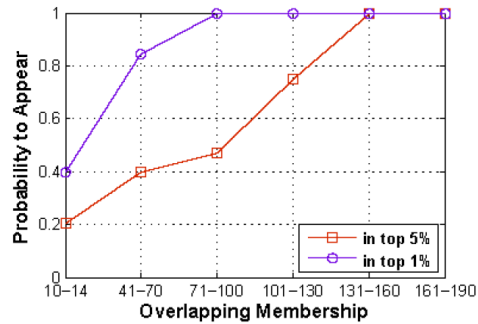


Fig. 7 Probability distribution of overlapping nodes to appear in the top influential 1% and 5% of the total nodes.

5 Conclusion and Future Work

In this chapter, we have presented a unified community analysis framework with an aim of addressing the major challenges, including the detection of overlapping and hierarchical communities and tracking their evolution in dynamic networks. Unlike many other research works in this field, we have also presented some of the important application areas for the identified communities which include detection of spammers and cloning attacks and leveraging community structures for the task of viral marketing. One of the future directions of research in this field can

be the realization of a multi-dimensional framework which utilizes both structural and non-structural information for social network analysis tasks. For example, to analyze the topological community structures and their evolution in the light of topical user sentiments on a common ground so as to generate greater insights related to information diffusion in a social network. Some important insights that may be gained by merging structural communities and user-centered topics/sentiments include identification of communities around discussion topics/sentiments and vice-versa, analyzing the correlation between the dynamics of communities (birth, merge, split, etc.) and the dynamics of discussion topics and sentiments, identifying opinion leaders within communities, diffusion of information (topics, rumors, fads, news) within and across communities, and modeling information diffusion process in a community-based perspective.

References

1. E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web, WWW'12*, pages 519–528, New York, NY, USA, 2012. ACM.
2. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
3. A.-L. Barabási and E. Bonabeau. Scale-free. *Scientific American*, 288:50–59, 2003.
4. S. Y. Bhat and M. Abulaish. Community-based features for identifying spammers in online social networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 100–107. ACM, 2013.
5. S. Y. Bhat and M. Abulaish. Overlapping social network communities and viral marketing. In *International Symposium on Computational and Business Intelligence*, 2013.
6. S. Y. Bhat and M. Abulaish. Using communities against deception in online social networks. *Computer Fraud and Security*, 2014(2):8–16, 2014.
7. S. Y. Bhat and M. Abulaish. Hocstracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks. *IEEE Transactions on Knowledge and Data engineering*, 27(4):1019–1013, 2015.
8. P. Bonhard. Who do trust? combining recommender systems and social networking for better advice. In *Proceedings of the Workshop on Beyond Personalization, in Conjunction with the International Conference on Intelligent User Interfaces*, pages 89–90. Citeseer, 2005.
9. L. Getoor. Link mining: A new data mining challenge. *SIGKDD Explorations*, pages 1–6, 2003.
10. M. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
11. D. Greene, D. Doyle, and P. Cunningham. Tracking the evolution of communities in dynamic social networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, ASONAM '10*, pages 176–183, Washington, DC, USA, 2010. IEEE Computer Society.
12. T. R. Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Semantic Web*, 6(1):4–13, 2008.
13. J. He and W. W. Chu. *A social network-based recommender system (SNRS)*. Springer, 2010.
14. W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 194–201. ACM Press/Addison-Wesley Publishing Co., 1995.
15. Z. Katona, P. P. Zubcsek, and M. Sarvary. Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research*, 48(3):425–443, 2011.

16. M.-S. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. *Proceedings of the VLDB Endowment*, 2(1):622–633, 2009.
17. K. Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage, 2004.
18. Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data*, 3:8:1–8:31, April 2009.
19. B. Liu and K. Chen-Chuan-Chang. Editorial: special issue on web content mining. *ACM SIGKDD Explorations*, 6(2):1–4, 2004.
20. P. Mika. Social Networks and the Semantic Web: The Next Challenge. *IEEE Intelligent Systems*, 20(1), 2005.
21. S. Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.
22. J. L. Moreno. Who shall survive?: A new approach to the problem of human interrelations. 1934.
23. M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.
24. H. Sun, J. Huang, J. Han, H. Deng, P. Zhao, and B. Feng. gskeletonclu: Density-based network clustering via structure-connected tree division or agglomeration. In *Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM'10*, pages 481–490, Washington, DC, USA, 2010. IEEE Computer Society.
25. A. H. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10, 2010.
26. Y. Wang, B. Wu, and N. Du. Community Evolution of Social Network: Feature, Algorithm and Model. *arxiv*, physics.soc-ph, 2008.
27. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, 1994.
28. H. White, S. Boorman, and R. Breiger. Social structure from multiple networks: I. blockmodels of roles and positions. *American Journal of Sociology*, 81(4):730–80, 1976.

Index

- Assortativity and network clustering, 2
- Cloning attacks detection, 16
- Community detection, 4
- Community evolution, 8
- Community evolution tracking, 10
- Content analysis, 5
- Copy-profiling detection, 16
- Density-based clustering, 9
- Dual-layer distance function, 10
- Dynamic communities, 8
- Hierarchical communities, 7
- Hierarchical community detection, 10
- HOCTracker, 9
- Hybrid social network analysis, 6
- Influence analysis, 6
- Influential node detection, 18
- Information diffusion, 6
- Link mining, 4
- Non-structural social network analysis, 4
- Online spammer detection, 14
- Opinion mining, 5
- Overlapping communities, 7
- Overlapping community detection, 10
- Polarity determination, 5
- Power law, 2
- Preferential attachment, 2
- Recommender systems, 7
- Semantic analysis, 6
- Sentiment analysis, 5
- Small-world behavior, 2
- Social network analysis, 3
- Strength of weak ties, 2
- Structural social network analysis, 4
- Viral marketing, 18