

Overlapping Social Network Communities and Viral Marketing

Sajid Yousuf Bhat

Department of Computer Science
Jamia Millia Islamia (A Central University)
Jamia Nagar, New Delhi-25, India
E-mail: s.yousuf.jmi@gmail.com

Muhammad Abulaish[#], SMIEEE

Department of Computer Science
Jamia Millia Islamia (A Central University)
Jamia Nagar, New Delhi-25, India
E-mail: abulaish@ieee.org

Abstract – Social networks have highly been used to understand the behavior and activities of individuals in nature and society. They are being used as a means to communicate, diffuse information, and to control the spread of diseases and computer viruses, in addition to many other tasks. Business organizations look upon social networks as an opportunity to spread the word-of-mouth for viral marketing and this task has gained significance with the popularity of Online Social Networks (OSNs). However, an important characteristic of social networks, including OSNs, which is the existence of overlapping communities of users, has not been exploited yet for the task of viral marketing even though it seems promising. This paper aims to present the importance of identifying overlapping communities for the task of viral marketing in social networks and also provides some experimental results on an email network to back the claims.

Keywords – Social network analysis, Business intelligence, Overlapping community mining, Viral marketing.

I. INTRODUCTION

Social Network Analysis (SNA) is one of the important techniques used in the field of sociology and is gaining significance in anthropology, biology, communication, economics, and social computing. The increasing popularity of social networks is largely because they help to model the various processes that take place in society, such as spread of cultural fads or diseases, formation of groups and communities, and so on. The application of SNA and social network concepts to a wide domain of research interests has recently gained huge popularity in various domains, including (but not limited to) transport planning, organized crime in adversary networks, analysis of biological networks, and business applications.

Among the most important tasks related to the field of social network analysis are community detection and viral marketing. The field of community detection deals with the problem of identifying densely connected groups within social networks, which are important because they highlight closely related functional units of a networked system. Detecting community structures from social networks is often challenging as it 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.

Viral Marketing (VM) on the other hand refers 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. The main aim of VM is to identify a set of seed nodes in a social network that are expected to have high influence in it. This ultimately requires developing product promotion strategies for only a small set of influential nodes in a social network. There exists a large amount of literature related to the developments in the two related fields but very few aim to incorporate them together. Recent developments in the area of community detection, which include methods for identifying overlapping communities, tracking evolution of communities in dynamic networks, identifying community structures at varying resolution levels and the roles (core and boundaries) of the nodes, but still they remain unexploited for the task of viral marketing.

In this paper, we aim to highlight the significance of social network community structure for the task of viral marketing. We present how the recent developments in the analysis of community structure can be exploited to support the task of viral marketing in social networks. The rest of the paper is organized as follows. Sections II and III provide a brief overview of the literature related to community detection and viral marketing in social networks, respectively. Section IV provides the insights on how community structures and related properties can be exploited for viral marketing. Section V presents some experimental results, and finally section VI concludes the paper.

II. SOCIAL NETWORK COMMUNITIES

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. Community detection from social networks has received a lot of attention in the recent past, and the field is still rapidly evolving. An open challenge related to community detection is the identification of overlapping communities that occur when a particular node in a network simultaneously belongs to several communities. Another challenge related to community detection is identifying hierarchical communities wherein a

[#] To whom correspondence should be addressed

community may be a part of a larger community. 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 community structures. In this regard, Bhat and Abulaish [1] proposed a density-based overlapping community detection method to track the evolution of overlapping communities in dynamic social networks. Moreover, tuning the resolution parameter of their method allows finding communities at varying resolutions/sizes, and thereby also finding a hierarchical community structure. Their method uses a local approach to find overlapping community structures and also identifies hubs and outliers from the underlying social network.

III. VIRAL MARKETING

Viral marketing involves identifying individuals with high *social networking potential* (size of an individual's social network and their ability to influence the network). A viral marketing strategy involves encouraging the word-of-mouth by distributing discounted or free products to targeted consumers with an assumption that the product is discussed further with friends and friends-of-friends which encourages 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 and identifying them is the main issue related to viral marketing.

Domingos and Richardson [2] tackled the viral marketing problem as the process of *influence maximization*. They model the problem as Markov random fields and discuss heuristic approaches for determining marketing strategies which tend to provide an approximate maximization of the overall expected lift in profit. Sun and Tang [3] modeled node influence in terms of some important social networking primitives which include *node degree*, *edge betweenness*, *structural holes*, and *homophily*. They present some models for maximizing the influence spread in social networks which include *high-degree heuristic*, *low-distance heuristic*, and *degree-discount heuristic*.

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.

IV. COMMUNITIES FOR VIRAL MARKETING

The areas of community detection and viral marketing are well studied and have significantly emerged with the recent developments in the respective fields. However, very few studies have attempted to tie them together. The significance

of their merger can be reasoned by the insights provided by a few researchers recently which we summarize below.

The analysis of Katona et al. [4] on an OSN dataset shows 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. The empirical analysis of Lerman et al. [5] on some online social networking sites also highlights the significance of community structures for viral marketing. They stress on the observation that dense community structure found in social networks results in a lower epidemic threshold. Hinz et al. [6] pointed out that seeding the hubs (high-degree seeding) for viral marketing results in higher number of referrals because hubs are more actively involved in diffusion process due to higher number of links.

In the following sub-sections, we describe some of the community-related concepts and node properties in social networks that can be used to aid the task of viral marketing. The main issue addressed here is the identification of significant nodes that possibly have a higher level of local or global influence in a social network.

A. Overlapping Nodes and Hubs

A challenge being addressed in the task of community detection is the identification of overlapping communities wherein a node can belong to multiple communities. The more the number of communities a node belongs to, the more significant seed it qualifies for viral marketing as it is expected to diffuse information between multiple groups in a social network. Besides overlapping nodes, density-based community detection methods like [1] identify hubs which are nodes which do not belong to any community but connect multiple communities. Similar to overlapping nodes, hubs which connect more communities show similar significance for being considered as seed nodes. Nodes which occupy structural holes in the network, i.e., hubs and overlapping nodes prove to be more influential and in this regard, community detection methods which identify hubs and overlapping nodes from social networks are significant for the task of viral marketing.

B. Core Nodes and Boundary Nodes

A few community detection methods like [1] categorize nodes within a community as core-nodes and boundary nodes. Core-nodes of a community form central, important nodes which hold the community together while as boundary nodes represent less significant nodes of a community and are mainly local neighbors around the core-nodes. In the context of viral marketing, a small set of connected core-nodes within a community represent significant seed nodes that can influence the whole community. Moreover, community detection methods like [1] also identify outliers from social networks, i.e., nodes which can be considered as noise since they do not belong to any community nor they qualify as hubs. Filtering out such nodes also helps in reducing the domain of nodes from where the top- k influential nodes need to be identified for viral marketing.

C. Local Communities

Unlike global community detection methods, which require the complete network, local community detection methods allow communities to be identified around individual node(s). They require information only related to the local neighborhood of the node around which a community needs to be detected. For the task of viral marketing, local community detection methods can simplify and speedup the task of identifying the potential community of nodes that can be influenced by a seed node.

D. Hierarchical Communities

Besides identifying local communities, it is desirable to identify community structures at varying resolutions of density/size characteristics and visualize them as a hierarchical structure wherein smaller communities (at lower levels) are contained in larger communities (at higher levels). Community detection methods like [1] allow identifying such structure by varying a resolution parameter to set the minimum size of a community. For the task of viral marketing, hierarchical community structures can be used to identify nodes or communities from the lowest hierarchical level around which important communities accumulate at the higher levels. Such nodes or communities represent important seed nodes that can influence a large portion of the social network.

E. Dynamic Communities

As mentioned earlier, social networks are dynamic in nature and change their structure due to the addition/removal of nodes and addition/breakage of links between nodes. This in turn also causes the community structures in these network to change with time. The various events related to the evolution of communities in dynamic networks include formation of new communities (birth), dissolution of existing communities (death), joining of multiple communities into a single community (merge), division of a community into multiple sub-communities (split) and joining and leaving of nodes to and from communities respectively (growth and shrinkage). Based on the rate of change in the membership, a community can be identified as static (no change to the membership of the community occurs with time), stable (changes occur to the membership of a community at a slow rate and a set of stable nodes maintain their membership with the community throughout) and volatile (the community shows high rate of membership changes including frequent split and merge events). For a static community any central node can be chosen to significantly influence the whole community. Similarly, for a stable community, a subset of the stable nodes can be chosen to influence the whole community including the nodes that join and leave the community with time. Volatile communities can be expected to be influenced by the members of the stable communities and the nodes that leave these communities with time through the process of cascading influence.

In a community with a static membership, each node has a strong influence on a fixed set of nodes. However, for a community which suffers many membership changes, a regular/faithful node of the community has a higher chance

to influence more nodes with time, although, with lesser degree of influence.

V. EVALUATING THE SIGNIFICANCE

In this section, we aim to provide some experimental basis to back our claims mentioned earlier. Using a weighted online social network of email communications from Enron [7] consisting of 87,273 nodes (reduced to 13,750 nodes based on non-zero out-degree), we analyze the community-based roles of the nodes (overlapping nodes and outliers in particular) in light of influence-ranking of nodes based on their betweenness centrality.

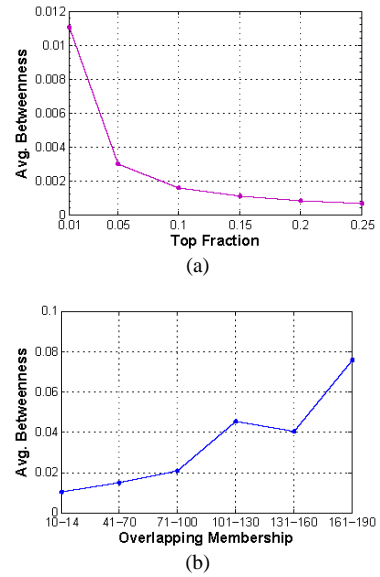


Figure 1. Average betweenness centrality of nodes (a) for top influential fractions ranging from 1% to 25% of the total nodes (b) for overlapping nodes arranged in bins of membership.

We have compared the average betweenness scores of the nodes appearing in the top fractions (based on betweenness) and the overlapping nodes identified by the community detection method proposed in [1]. The overlapping nodes are arranged into various bins based on the number of communities a node belongs to (membership of an overlapping node) ranging from 10-40 to 161-190. The results are shown in Fig. 1 wherein Fig. 1a plots the average betweenness (y-axis) of the nodes in top fractions ranging from 1% to 25% (x-axis) and Fig. 1b plots the average betweenness (y-axis) of the overlapping nodes arranged into bins of membership (x-axis).

From Fig. 1, we can see that overlapping nodes show significant betweenness centralities wherein nodes having higher membership show higher influence. It also indicates that in order to identify a minimal set of nodes to maximize influence, it is better to consider highly overlapping nodes than to consider top-ranked nodes based on betweenness centrality. This statement can be backed by the results shown in Fig. 2 which highlights the fact that more than

90% of the nodes in the top 1% influentials are overlapping nodes and their occurrence decreases as we go down the ranking.

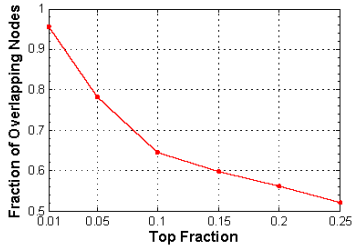


Figure 2. Occurrence of overlapping nodes in the top influential fractions ranging from 1% to 25% of total nodes.

We also determine the chance of an overlapping node to appear in the top influential list of a social network. In this regard, we calculate the probabilities according to which overlapping nodes (according to bins of membership) appear in the top 1% and 5% of the influential nodes in the network. The probability for a particular membership bin to appear in a top fraction is calculated as the ratio of the number of its candidates occurring in a top fraction to the total number of its candidates in the whole network. The probabilities of the various overlapping membership bins to appear in the top 1% and 5% influential nodes are plotted in Fig. 3.

The results shown in Fig. 3 further strengthen the claim that highly overlapping nodes directly represent highly influential nodes in a social network as they are more likely to appear in the top 1-5% of the influential nodes in the network.

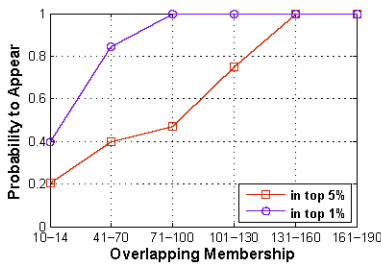


Figure 3. Probability of an overlapping node (according to membership bins) to appear in the top influential 1% and 5% of the total nodes.

Now besides the overlapping nodes, our proposed community detection method [1] also identifies outliers, i.e., nodes which are not assigned to any community and are considered as noise. Our analysis of the betweenness centrality ranking of the outliers reveals that they mainly occur towards the tail of the diminishing influence and have very low betweenness centrality score. In Fig. 4 we present some results related to the memberships of the various bottom node fractions. Figure 4 indicates that on an average 60% of the least-influential-nodes fractions of the network ranging from 10% to 60% (x -axis) are outliers. In addition to

this, on an average almost 100% of the least-influential-nodes fractions ranging from 10% to 60% (x -axis) have their community membership less than 2 (i.e., non-overlapping nodes including outliers). It means that the non-overlapping nodes (including outliers) identified from a network can be directly considered as least influential and can be excluded for the task of viral marketing.

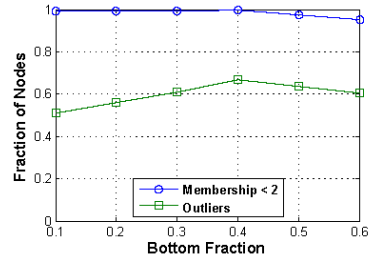


Figure 4. Occurrence of outliers and non-overlapping nodes in the bottom least-influential nodes ranging from 10% to 60% of the total nodes.

VI. CONCLUSION

In this paper, we have presented some key points wherein overlapping community structures in social networks can be used to augment the process of viral marketing. We have also presented some preliminary results indicating the higher significance of overlapping nodes and lesser significance of non-overlapping and outlier nodes based on their betweenness centrality scores.

ACKNOWLEDGMENT

Sajid Yousuf Bhat is a Senior Research Fellow under the UGC-BSR fellowship for meritorious students and this work was partially supported by the grant.

REFERENCES

- [1] S. Y. Bhat and M. Abulaish, "OCTracker: A Density-Based Framework for Tracking the Evolution of Overlapping Communities in OSNs," Proc. IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining, 2012, pp. 501-505.
- [2] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," Proc. of ACM KDD'02, NY, pp. 61-70, 2002.
- [3] J. Sun, and J. Tang, "A survey of models and algorithms for social influence analysis," Social Network Data Analytics, Springer US, pp. 177-214, 2011.
- [4] Z. Katona, P. P. Zubcsek, and M. Sarvary, "Network effects and personal influences: The diffusion of an online social network," Journal of Marketing Research, vol. 48(3), pp. 425-443, 2011.
- [5] K. Lerman, R. Ghosh, and T. Surachawala, "Social contagion: An empirical study of information spread on digg and twitter follower graphs", arXiv preprint arXiv:1202.3162, 2012.
- [6] O. Hinz, B. Skiera, C. Barrot, and J. U. Becker, "Seeding strategies for viral marketing: An empirical comparison," Journal of Marketing, vol. 75(6), pp. 55-71, 2011.
- [7] B. Klimt and Y. Yang. "The Enron corpus: A new dataset for email classification research." In Proc. European Conf. on Machine Learning, pp. 217-226, 2004.