2012 32nd International Conference on Distributed Computing Systems Workshops

Discovering Shared Interests in Online Social Networks

Feng Wang, Kuai Xu, and Haiyan Wang
Arizona State University
Email: {fwang25, kuai.xu, haiyan.wang}@asu.edu

***Abstract*—The capacity of rapidly disseminating information such as latest news headlines has made online social networks a popular and disruptive venue for spreading influence and distributing contents. Given the importance of online social networks, it becomes increasingly imperative to understand the shared interests of users on the popular information or contents that circulate through these networks. This paper proposes a novel graphical approach based on bipartite graphs and one-mode projection graphs to model the interactions of users and information and to capture the shared interests of users on the information. The experiments based on data-sets collected from `Digg`, a popular social news aggregation site, have demonstrated the proposed approach is able to discover inherent clusters of users and information within online social networks. The evaluation results also show that these clusters exhibit distinct characteristics. To the best of our knowledge, this paper is the first attempt to apply bipartite graphs and one-mode projections to shed light on the interactions of people and information in online social networks and to discover the clustered nature of users and contents.**

I. Introduction

As online social networks continue to grow in users, traffic and information, it has become increasingly important to understand the interactions of people and information on these networks. For example, the discovery of shared interests among users could improve the quality of recommendation systems [1], [2], [3] or detect community of interests in social networks. Although there exists an extensive body of research work on network topology, traffic characteristics and information spreading in online social networks [4], [5], [6], [7], [8], [9], [10], little attempt has been made to discover the clusters of users with shared interests or the clusters of information interacted by similar sets of social network users.

In this paper, we propose a novel graphical approach to model the interactions in online social networks using bipartite graphs which represent users and information of online social networks with two disjoint vertices sets. Further, we exploit one-mode projections of bipartite graphs to capture the similarity of information and shared interests of users [11], [12] through creating graph edges among information and users, respectively. Specifically, the one-mode projection of bipartite graphs on users connects users that interact with the same one or more information, e.g., news stories, Web links, photos, videos. The weighted edges of this graph represent the degree of such similarity, e.g., number of shared news stories between two users. Similarly, the one-mode projection of bipartite graphs on the information connects two information that are interacted by the same users. The weighted edge between two information represents the number of users that interact with both of them.

The edges in one-mode projection graphs serve as the *similarity* between the information or the *shared interests* between users. Thus, we further build the *similarity matrix* of the one-mode project graphs based on the connections among the information or users. The similarity matrix in turn leads us to explore clustering algorithms to divide the users and information into distinctive and meaningful clusters such that each cluster groups users with similar interests or information with similar contents or topics.

Our experiment results with real data-sets collected from `Digg`, a popular social news aggregation site, show that the clustering step indeed divides the users and information into clusters with distinct characteristics. More importantly, the availability of the *users* or *information* clusters significantly improves our understanding on the interactions of users and information in online social networks. We believe that the insights gained from the discovery of shared interests among users will facilitate information spreading and the classifications on contents circulating in these networks, which will be very valuable for filtering unwanted contents such as spams in online social networks.

The contributions of this paper are as follows:

- We introduce a novel graphical approach to model the interactions of users and information in online social networks using bipartite graphs and one-mode projection graphs;
- We apply a simple yet efficient clustering algorithm to divide users and information into meaningful clusters with distinct characteristics;
- We evaluate the proposed approach using real data-sets collected from a popular social news aggregation site, and find that the approach indeed helps discover the shared interests of users and the similarity of the information.

The remainder of this paper is organized as follow. Section II describes the proposed method of modeling interactions of users and information in online social networks using bipartite graphs and one-mode projection graphs. Section III introduces a simple yet efficient clustering algorithm

1545-0678/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDCSW.2012.15

IEEE computer society

to divide users and information into distinctive clusters based on the similarity matrix built from one-mode projection graphs. Section IV presents the experiment results based on real data-sets collected from a popular social news aggregation site. Section V briefly discusses related work, and Section VI finally concludes this paper and outlines our future work.

II. Methodology

In this section we first describe how we model the interactions of users and information in online social networks with bipartite graphs. Subsequently, we explore one-mode projections of bipartite graphs to capture the shared interests of users and the similarity of information.

A. Modeling Interactions of Users and Information with Bipartite Graphs

Online social networks have recently become a popular venue for users to create, spread, and discuss information such as news stories, blogs, pictures and videos. The interactions of users and information could be naturally modeled with *bipartite graphs*, where users and information form two disjoint vertex sets [13], [14], [15], [16]. Bipartite graphs have widely used in analyzing collaboration networks such as the co-authorship of authors on Wikipedia's articles and the collaborations of actors on movies [17], [18], since the graph structure of these collaboration networks exhibit bipartite patterns.

For the case of online social networks, we could use a bipartite graph $\mathcal{G} = \{\mathcal{U}, \mathcal{S}, \mathcal{E}\}$ to represent the interactions between users and information. The vertex sets $\mathcal{U}$ and $\mathcal{S}$ denote all users and information in online social networks, respectively, while $\mathcal{E}$ denote the interactions between users and information. For example, $e_{u,s} \in E$ suggests that the user u interacts with the information s through creating, posting, commenting, digging, tweeting or other activities in online social networks. Hence the $e_{u,s}$ reflects the interest of the user u on the information s. In other words, all the edges in the bipartite graph characterize the interactions between users and information in online social networks, and provide a unique perspective for understanding *why users interact with a certain set of information*.

B. One-Mode Projections of Bipartite Graphs

Projecting bipartite graph onto the unipartite space produces one-mode projection graph, in which an edge is generated between two nodes in the same vertex set if both of them connect to one or more same nodes in another vertex set of the bipartite graph, e.g., two users in online social networks tweeting on the same news story. As a result, one-mode projection graphs are often used to capture the hidden information or structure from the nodes in the same vertex set [13], [19].

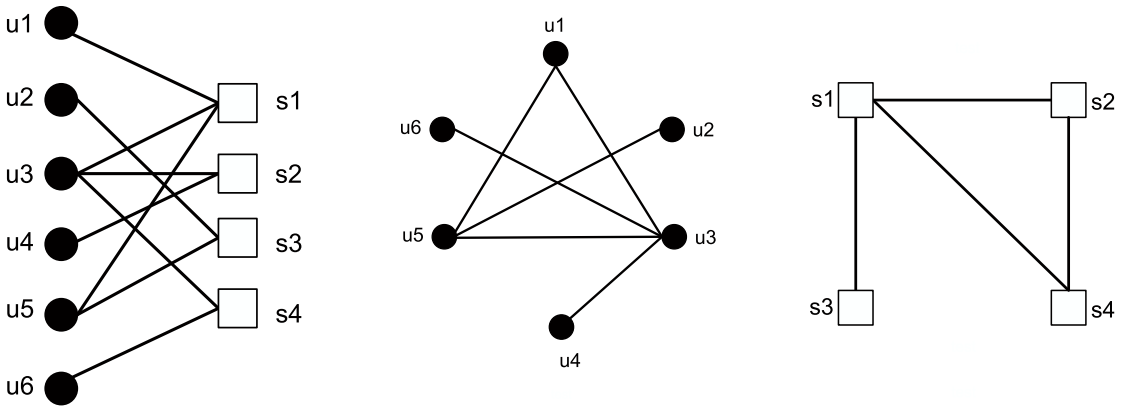
Figures 1[a-c] illustrate an example of bipartite graphs and its one-mode projection graphs on both vertices sets. Figure 1[a] shows six users in the left vertices set and four news stories in the right vertices set, while Figures 1[b][c] show the one-mode projection graphs on the left and right vertices sets, respectively. The clique formed by nodes u_1, u_3, u_5 in Figure 1[b] is due to their common connections to the same story s_1, while the clique of s_1, s_2, s_4 in Figure 1[c] is due to the user u_3 interacting with these news stories in the bipartite graph, as shown in Figure 1[a].

In this paper we leverage one-mode projection graphs to discover the shared interests of users in online social networks as well as to study the similarity of information spreading on these networks. Based on the bipartite interaction graph of users and information in online social networks, we obtain two one-mode projection graphs: $\mathcal{G}_\mathcal{U} = \{\mathcal{U}, \mathcal{E}_\mathcal{U}\}$ and $\mathcal{G}_\mathcal{S} = \{\mathcal{S}, \mathcal{E}_\mathcal{S}\}$. An edge $e_{i,j}$ forms between two users u_i and u_j in $\mathcal{G}_\mathcal{U}$ if and only if both of them interacts with one or more same information. The weighted edges of the graph $\mathcal{G}_\mathcal{U}$ represent the degree of common interests. Similarly, two information s_i and s_j are connected in $\mathcal{G}_\mathcal{S}$ if and only if one or more users interact with both of them in the bipartite graph. The weighted edge between two information represents the number of users that interact with both of them. Therefore, the one-mode projection graph $\mathcal{G}_\mathcal{U}$ essentially captures the shared interests of users, while the graph $\mathcal{G}_\mathcal{S}$ characterizes the similarity of information based on the user interaction patterns.

III. Discovering Shared Interests Via Clustering Algorithms

The shared interests of two users u_i and u_j could be numerically represented with the number of the shared information they have interacted with, i.e., $s_{i,j} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$, where S_i and S_j denote the set of information the users u_i and u_j have interacted with, respectively. Thus, the shared interests among all users could be represented with a similarity matrix. The availability of the similarity matrix leads us to the next step of finding clustering algorithms to discover clusters of users and information that share similar characteristics [20], [21], [22], because clustering algorithms have been widely used in clustering communication patterns of Internet end systems in recent years [23], [24], [25]. The goal of the clustering step is to divide the users and information into different groups based on their interaction patterns.

In this paper, we adopt the agglomerative clustering algorithm on the similarity matrix of one-mode projection graphs $G_\mathcal{P}$ and $G_\mathcal{S}$, since agglomerative algorithms optimize the clustering results by maximizing the internal similarity within the same clusters as well as minimizing the external similarity between nodes in different clusters [20]. We evaluate the quality of clustering results with the default



(a) An example of bipartite graphs (b) One-mode projection graph on users (c) One-mode projection graph on information, e.g., news stories

Figure 1. Modeling the connections between users and information, e.g., news stories in online social networks with bipartite graphs and one-mode projection graphs

$\mathcal{I}_2$ criterion function [20], i.e., maximizing

$$\sum_{k}^{i=1} \sqrt{\sum_{v,u \in C_i} similarity(v,u)}, \qquad (1)$$

where C_i denotes the i-th cluster, $i = 1, 2, \ldots, k$ and the $similarity(v,u)$ between information or users in the same clusters is computed with the *cosine* function.

IV. Experiment Results

The data used in this study were collected in a previous study [26], in which Lerman et al. collected the news stories that were promoted to the front page of `digg.com`, a major social news aggregate site, due to its popularity during June 2009. For each news story, the data set includes a list of Digg users who had *voted* or *digged* the story and the associated time-stamp. In total, there are over 3 million votes (also called `diggs`) from `139,409` users on the most popular `3,553` news stories during that month. Figure 2 shows the distribution of the votes for all the news stories in the data-set. In average, each news stories receives votes from nearly 850 users (0.6% of all the users in the data-set).

Figure 3 illustrates a *heavy-tail* distribution of news stories `digged` by each user. As shown in this figure, there exists a few users who have actively voted on a large number of news stories, while over 70% of users have voted 10 or less news stories.

To gain an in-depth understanding of the interactions of users and information, we first study the shared interests of users on the news stories. A simple `digg` activity or vote behavior $d_{u,s}$ suggests the interest of the user u on the news story s. Hence a pairwise of `digg` activities, ($d_{u,s}$, $d_{v,s}$) indicate the *shared* or *common* interests of two users u and v on the same news stories s. Discovering such shared

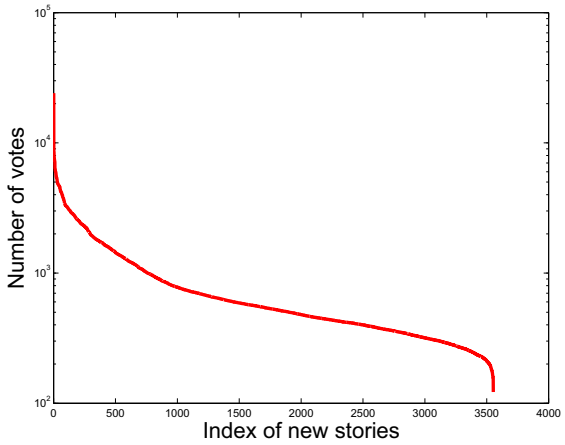


Figure 2. Distribution of votes received by all the news studies in the data-sets

interests of all users in social news sites not only helps understand the interactions of users and information, but also provides valuable insights for the classifications of users and information into certain categories.

Based on the proposed methodology in the previous sections, we use bipartite graphs to characterize the interactions of users and news stories, and then build one-mode projection graphs on users and news stories, respectively. Further we run the clustering algorithm discussed in Section III on the similarity matrix of the one-mode projection graphs and obtain clusters of users and news stories. The *users clusters* group users with shared interests on news stories, while the *news story clusters* group news stories that are voted by a similar set of users.

To evaluate the clustering results, we propose a simple metric, *voting consistency*, to denote the similarity of voting

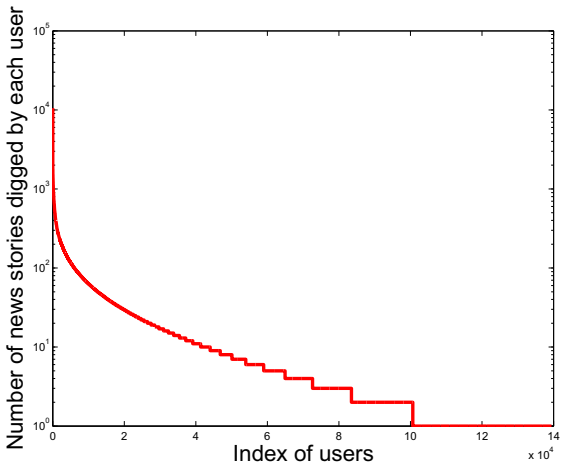


Figure 3. Distribution of news stories digged by users

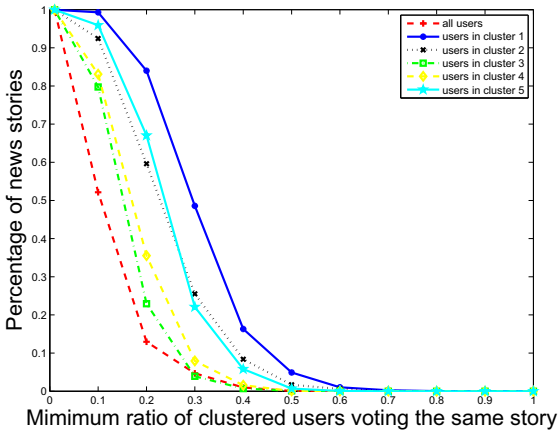


Figure 4. Percentage of news stories that are voted by users in the same *user clusters*

patterns in *users clusters* and *news story clusters*. Specifically, the voting consistency of a *user cluster* is measured by the percentage of news stories that are voted by a certain ratio of users in the same cluster. Figure 4 shows the percentage of news stories that are voted by users in the same clusters, where x axis is the ratio of users in a given cluster and y axis denotes the percentage of news stories that are voted by the ratio of users exceeding x in the same cluster. For example, in the *user cluster* 1 there are nearly 50% of news stories that are voted by at least 30% users in the same cluster. Such high voting consistency is significant and interesting, since comparing with all the news stories, there are only 4.7% news stories that receive votes from 30% of all users. For another instance, in the *user cluster* 5 there are over 65% of news stories that are voted by at least 20% of all users in this cluster. However, there are only 12.9% news stories that receive votes from over 20% of all users. Similar observations on the high voting consistency hold in other *user clusters* as well.

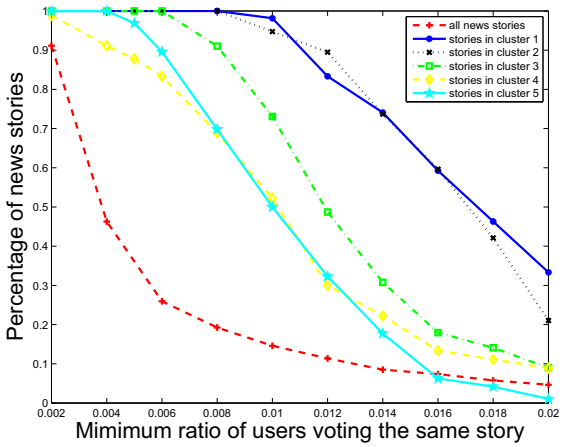


Figure 5. Percentage of news stories that are voted by users in the same *news story clusters*

Next we evaluate the clustering results on the news stories with the same *voting consistency* metric. Similar to Figure 4, Figure 5 illustrates the percentage of news stories that are voted by a certain ratio of users for *news story clusters*. The overall observation is that the news stories in individual clusters receive consistent votes from similar groups of users compared with all the news stories (the red bottom line in the figure) in the entire data-set. These results suggest that the discovery of *news story clusters* successfully divide news stories into distinctive groups. One of our future work is to explore these clusters to classify news stories and users for detecting anomalous user voting behavior and unwanted spam contents in social news sites.

V. Related Work

Recent years have witnessed the success of online social networks in transforming the way people connecting with information such as news stories, pictures, and videos. The benefits of online social networks have gone beyond connecting people with information, as many recent studies have exploited online social networks for a variety of important applications, including mitigating multiple identities attacks (also known as Sybil attacks) [27], improving online marketplaces [28], and spreading positive influence [29].

To gain a deep understanding of online social networks, a rich body of prior work have been devoted to study network topology and traffic characteristics of online social networks [4], [5], [6], [7], [8], [9]. For example, a measurement study [4] on Renren, the largest online social network in China, characterizes users interactions and user popularity, while [5], [6] study the interactions of users in online social networks and user behaviors using HTTP

header information extracted from network traffic. In [8], [7], Nazir et al examine end-to-end performance of online social networks based applications and investigate the impact of network-level performance on user experience. Several studies [30], [26], [31] also have studied the voting patterns of users in online social networks using data-sets collected from the Digg news aggregation site. Different from these work, this paper focuses on the shared interests of users [11], [1], [2], [3], [12] and the similarity of information on social news aggregation sites.

Graphical models have been widely used for studying communication networks [32], [33], [34], [19], [35]. For example, [33], [34] use graphlets to visualize the social interactions of end systems on the Internet for understanding host traffic behaviors. Similarly, Yu at al. introduce traffic activity graphs to capture the communication patterns of end hosts engaging in the same network applications [32]. Unlike these studies, our study proposes a new approach of applying bipartite graphs to model the interactions of people and information and subsequently using one-mode projections of bipartite graphs to capture the shared interests of people on the information. Our recent work [19] has demonstrated the capability of bipartite graphs and one-mode projections in discovering similar behavior patterns of end hosts in the same network prefixes.

VI. Conclusions and Future Work

This paper presents a novel graphical approach to study the interactions of users and information in online social networks. We use bipartite graphs to model the interactions of users and information, and subsequently build one-mode projections of bipartite graphs to capture the shared interests of users and the similarity of information. Through a simple yet effective clustering algorithm, we find the inherent clusters of information and users with each cluster exhibiting distinctive characteristics. Using real data-sets collected from Digg, a popular social news aggregation site, our experiments results suggest that the proposed methodology is able to uncover the shared interests of users in voting news stories, and to discover the similarity of news stories that attract the same set of users. These findings will improve our understanding on the interactions of users and information in online social networks. We are currently in the process of extending the same methodology on other online social networks such as Twitter and Flickr. In addition, we are exploring the clustering results to classify the users and information and to detect anomalous user voting behavior and unwanted spam contents in online social networks

Acknowledgment

We would like to thank Kristina Lerman for making the Digg 2009 data set available to our research project. This work was supported in part by Arizona State University New College SRCA grants.

References

[1] X. Li, L. Guo, and Y. Zhao, "Tag-based social interest discovery," in *Proceedings of international conference on World Wide Web*, April 2008.

[2] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt, "Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis," in *Proceedings of International Conference on The Semantic Web*, October 2008.

[3] V. Zanardi, and L. Capra, "Social Ranking: Uncovering Relevant Content Using Tag-based Recommender Systems," in *Proceedings of ACM Conference on Recommender Systems*, October 2008.

[4] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Zhao, "Understanding Latent Interactions in Online Social Networks," in *Proceedings of ACM SIGCOMM International Measurement Conference*, November 2010.

[5] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding Online Social Network Usage from a Network Perspective," in *Proceedings of ACM SIGCOMM International Measurement Conference*, November 2009.

[6] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing User Behavior in Online Social Networks," in *Proceedings of ACM SIGCOMM International Measurement Conference*, November 2009.

[7] A. Nazir, S. Raza, D. Gupta, C.-N. Chuah, and B. Krishnamurthy, "Network Level Footprints of Facebook Applications," in *Proceedings of ACM SIGCOMM International Measurement Conference*, November 2009.

[8] A. Nazir, S. Raza, and C.-N. Chuah, "Unveiling Facebook: A Measurement Study of Social Network Based Applications," in *Proceedings of ACM SIGCOMM International Measurement Conference*, October 2008.

[9] H. Kwak, Y. Choi, Y.-H. Eom, H. Jeong, and S. Moon, "Mining Communiteis in Networks: A Solution for Consistency and its Evaluation," in *Proceedings of ACM SIGCOMM International Measurement Conference*, November 2009.

[10] F. Wang, H. Wang, and K. Xu, "Diffusive Logistic Model Towards Predicting Information Diffusion in Online Social Networks," in *Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking (HOTPOST)*, Macao, China, June 2012.

[11] M. Schwartz, and D. Wood, "Discovering shared interests using graph analysis," *Communications of the ACM*, vol. 36, no. 8, pp. 78 – 89, August 1993.

[12] L. Dietz, "Inferring Shared Interests from Social Networks," in *Proceedings of Neural Information Processing Systems Workshop on Computational Social Science and the Wisdom of Crowds*, December 2010.

[13] J.-L. Guillaume, and M. Latapy, "Bipartite graphs as models of complex networks," *Physica A: Statistical and Theoretical Physics*, vol. 371, no. 2, pp. 795 – 813, 2006.

[14] M. Barber, "Modularity and community detection in bipartite networks," *Physical Review E*, vol. 76, no. 6, 2007.

[15] E. Sawardecker, C. Amundsen, M. Sales-Pardo and L. Amaral, "Comparison of methods for the detection of node group membership in bipartite networks," *European Physical Journal B - Condensed Matter and Complex Systems*, vol. 72, no. 4, pp. 671–677, 2009.

[16] R. Guimera, M. Sales-Pardo, and L. Amaral, "Module identification in bipartite and directed networks," *Physical Review E*, vol. 76, no. 3, 2007.

[17] M. Newman, S. Strogatz, and D. Watts, "Random graphs with arbitrary degree distributions and their applications," *Phys. Rev. E*, vol. 64, no. 2, p. 026118, 2001.

[18] R. Jesus, M. Schwartz, and S. Lehmann, "Bipartite networks of Wikipedia's articles and authors: a meso-level approach," in *Proceedings of International Symposium on Wikis and Open Collaboration*, October 2009.

[19] K. Xu, F. Wang, and L. Gu, "Network-Aware Behavior Clustering of Internet End Hosts," in *Proceedings of IEEE INFOCOM*, April 2011.

[20] G. Karypis, "CLUTO: A Clustering Toolkit," http://glaros.dtc.umn.edu/gkhome/views/cluto.

[21] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in *Proceedings of Neural Information Processing Systems (NIPS) Conference*, 2001.

[22] Y. Weiss, "Segmentation Using Eigenvectors: A Unifying View," in *Proceedings of the International Conference on Computer Vision*, 1999.

[23] F. Hernandez-Campos, A. B. Nobel, F. D. Smith, and K. Jeffay, "Statistical Clustering of Internet Communication Patterns," in *Proceedings of Symposium on the Interface of Computing Science and Statistics*, March 2003.

[24] B. Krishnamurthy, and J. Wang, "On network-aware clustering of Web clients," in *Proceedings of ACM SIGCOMM*, August 2000.

[25] S. Wei, J. Mirkovic, and E. Kissel, "Profiling and Clustering Internet Hosts," in *Proceedings of the International Conference on Data Mining*, June 2006.

[26] K. Lerman, and R. Ghosh, "Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks," in *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*, May 2010.

[27] B. Viswanath, A. Post, K. Gummadi, and A. Mislove, "An Analysis of Social Network-Based Sybil Defenses," in *Proceedings of ACM SIGCOMM*, September 2011.

[28] G. Swamynathan, C. Wilson, B. Boe, K. Almeroth, and B. Zhao, "Do Social Networks Improve e-Commerce? A Study on Social Marketplaces," in *Proceedings of ACM SIGCOMM Workshop on Online Social Networks*, August 2008.

[29] F. Wang, H. Du, E. Camacho, K. Xu, W. Lee, Y. Shi, and S. Shan, "On Positive Influence Dominating Set in Social Networks," *Theoretical Computer Science*, vol. 412, no. 3, pp. 265–269, January 2011.

[30] K. Lerman and A. Galstyan, "Analysis of Social Voting Patterns on Digg," in *Proceedings of ACM SIGCOMM Workshop on Online Social Networks*, August 2008.

[31] S. Tang, N. Blenn, C. Doerr, and P. Mieghem, "Digging in the Digg Social News Website," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1163 – 1175, October 2011.

[32] Y. Jin, E. Sharafuddin, and Z.-L. Zhang, "Unveiling core network-wide communication patterns through application traffic activity graph decomposition," in *Proceedings of ACM SIGMETRICS*, June 2009.

[33] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, and G. Varghese, "Network monitoring using traffic dispersion graphs," in *Proceedings of ACM SIGCOMM Internet Measurement Conference*, October 2007.

[34] T. Karagiannis, K. Papagiannaki and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," in *Proc. of ACM SIGCOMM*, August 2005.

[35] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons, 1990.