Identifying Leaders and Followers in Online Social Networks

M. Zubair Shafiq, *Student Member, IEEE*, Muhammad U. Ilyas, *Member, IEEE*, Alex X. Liu, *Member, IEEE*, and Hayder Radha, *Fellow, IEEE*

***Abstract*—Identifying leaders and followers in online social networks is important for various applications in many domains such as advertisement, community health campaigns, administrative science, and even politics. In this paper, we study the problem of identifying leaders and followers in online social networks using user interaction information. We propose a new model, called the Longitudinal User Centered Influence (LUCI) model, that takes as input user interaction information and clusters users into four categories: introvert leaders, extrovert leaders, followers, and neutrals. To validate our model, we first apply it to a data set collected from an online social network called Everything2. Our experimental results show that our LUCI model achieves an average classification accuracy of up to 90.3% in classifying users as leaders and followers, where the ground truth is based on the labeled roles of users. Second, we apply our LUCI model on a data set collected from Facebook consisting of interactions among more than 3 million users over the duration of one year. However, we do not have ground truth data for Facebook users. Therefore, we analyze several important topological properties of the friendship graph for different user categories. Our experimental results show that different user categories exhibit different topological characteristics in the friendship graph and these observed characteristics are in accordance with the expected ones based on the general definition of the four roles.**

***Index Terms*—Online social networks, leaders, followers**

I. Introduction

A. Motivation

THIS PAPER represents the first study of identifying leaders and followers in online social networks *using only user interaction information*, to the best of our knowledge. Leaders in social networks are users whose opinions are highly influential on those of others. Followers in social networks are users whose opinions are highly influenced by those of leaders. Our study is motivated by a wide variety of applications in many domains such as advertisement, community health campaigns, administrative science, and even politics. For advertising, to improve the effectiveness of word of mouth advertising and increase the recommendation based product adoption, advertising companies want to give free samples to leaders instead of a random population [12]. For community health campaigns, such as HIV prevention programs [2] and school-level anti-smoking campaigns [22], targeting interventions at community leaders have been shown to be more effective than applying them to random individuals. Introducing new ideas to leaders of communities maximizes their spread to the rest of the population and is an effective way of indirectly reaching individuals that shy away from authorities. For administrative science, knowing who are leaders and who are followers is helpful in assembling effective product development teams that have a greater potential for delivering better work performance [18]. For politics, community leaders have been used to mobilize voters and to increase turnout in U.S. elections [4]. For the pharmaceutical industry, a significant portion of marketing budgets for new treatments is spent on identifying key opinion leaders, physicians with specialized knowledge [21]. Because the larger community of medical practitioners often resorts to leaders for trusted expert advice, pharmaceutical companies target these professional leaders to introduce new treatments.

B. Problem Statement

In this paper, we aim to develop a model that takes as input user interactions in an online social network to group users into the following four categories: Introvert Leaders (IL), Extrovert Leaders (EL), Followers (F), and Neutrals (N). User interactions are directed (inward and outward) and include but are not limited to wall posts, comments, and messages. The interaction information used in our study contains only the timestamp of interaction, the sender's identifier, and the receiver's identifier. Our method does not require the text content of user interactions because they are typically not available for analysis due to privacy concerns. *Introvert Leaders* are highly sought after individuals even though they interact very little with their friends. *Extrovert Leaders* frequently interact with their friends regardless of the level of reciprocation. *Followers* tend to interact with their friends as long as they reciprocate. Specifically, the number of outward interactions from followers tend to remain small when the number of inward interactions is small. *Neutrals* exhibit interaction levels that are independent of interaction levels of their friends.

Manuscript received February 15, 2012; revised July 20, 2012 and October 23, 2012. This work was supported in part by the National Science Foundation under Grant IIS-0968495.

M. Z. Shafiq and A. X. Liu are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, U.S.A. (e-mail: {shafiqmu, alexliu}@cse.msu.edu, Tel: +1-517-353-5152, Fax: +1-517-432-1061). Alex X. Liu is the corresponding author of this paper.

M. U. Ilyas is with the Department of Electrical Engineering, School of Electrical Engineering & Computer Science, National University of Sciences & Technology, Islamabad 44000, Pakistan (e-mail: usman.ilyas@seecs.edu.pk). Most of this work was conducted while he was a post-doctoral researcher in both the Department of Computer Science and Engineering and the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, U.S.A.

H. Radha is with Department of Electrical & Computer Engineering, Michigan State University, East Lansing, MI 48824, U.S.A. (e-mail: radha@egr.msu.edu).

Digital Object Identifier 10.1109/JSAC.2013.SUP.0513054

0733-8716/13/$31.00 © 2013 IEEE

C. Proposed Approach

In this paper, we propose a new model called the *Longitudinal User Centered Influence (LUCI)* model for grouping users in an online social network into the following four categories: introvert leaders, extrovert leaders, followers, and neutrals. This model only requires interaction data among users in an online social network as its input. The LUCI model allows us to observe four categories of users that exhibit distinct communication behavior. Introvert leaders rarely initiate interactions while extrovert leaders frequently initiate interactions. The communication habits of both introvert leaders and extrovert leaders changed little regardless of changes in the frequency they were approached by their friends. For followers, the number of interactions initiated by them with their friends depend on how many interactions they received. Neutrals rarely initiate interactions and use the social network very inconsistently; they may use social networks as a data collection tool and browse other people's profiles, but do not use it to communicate much.

The LUCI model essentially captures the effects of inherent user behavior from the interaction activities of users in their respective neighborhoods. Using linear regression formulation on interaction data, the LUCI model computes two coefficients, which we call *ego coefficient* and *network coefficient*, to characterize the behavior of each user. The ego coefficient quantifies the correlation between a user's past interaction information and future outward interactions. The network coefficient quantifies the correlation between a user's past inward interaction information and future outward interactions. The LUCI model then clusters users based on these two coefficients using the kernel k-means algorithm [5]. Grouping users based on the LUCI model can be done efficiently. First, for a given user, the number of the equations to be solved using linear regression grows only linearly with the number of friends of the user and the time duration over which the interactions of this user with their friends is collected. Second, the computation of ego coefficient and network coefficient can be easily parallelized because the computation for each user can be done independently.

To validate the LUCI model, we first apply it to a data set obtained from Everything2 consisting of interactions among approximately 8 thousand users over the duration of six years. Each user in this data set is labeled as a leader or follower based on whether the role in Everything2 is an administrator. Our experimental results show that the LUCI model achieves an average classification accuracy of up to 90.3% in classifying users as leaders and followers. Second, we apply our LUCI model on a data set collected from Facebook consisting of interactions among more than 3 million users over the duration of one year. Since we do not have ground truth data for Facebook users, we analyze several important topological properties of the friendship graph for different user categories. Our experimental results show that different user categories exhibit different topological characteristics in the friendship graph and these observed characteristics are in accordance with the expected ones based on the general definition of the four roles of introvert leaders, extrovert leaders, followers, and neutrals.

D. Key Findings

We now provide some key findings about different user categories from our experiments.

- **Followers:** The outward interactions of followers are driven largely by the level of inward interactions, which are essentially controlled by their neighbors. We have found that followers are part of closely connected communities and have the highest average clustering coefficient compared to other user categories.
- **Leaders:** The number of outward interactions of extrovert leaders is determined more by the number of outward interactions in the past and less by the number of interactions they receive from their neighbors. The communication of extrovert leaders is self-driven and they have the highest average degree in the friendship graph amongst all user classes. Extrovert leaders are seen to have more friends than all other user types, followed closely by introvert leaders. Introvert leaders consistently have little outward interactions but receive a lot of inward interactions from their neighbors. Our results show that these highly sought-after introvert leaders have the smallest average shortest path length to the rest of the users in the social network. We note that the properties of leaders both introvert and extrovert are very similar to those of hubs in small-world networks [23]. Our analysis of clustering coefficient revealed that extrovert and introvert leaders are more likely to be connected to different internally well-knit communities. Their average clustering coefficients are lower than those of followers. The connections to diverse communities makes extrovert and introvert leaders have lower average shortest path lengths than other user categories. We observed that introvert and extrovert leaders tend to be surrounded mostly by followers. Due to this, the distribution of eigenvector centralities of introvert and extrovert leaders is similar to that of followers.
- **Neutrals:** The number of outward interactions of neutrals is not related to either inward or outward interactions in the previous time period. Neutral users have the lowest average degree and are mostly connected to followers or other neutrals in the friendship graph. The average eigenvector centrality of neutrals is two orders of magnitude lower than other user categories.

E. Major Contributions

We make three key contributions in this paper.

1) We propose a new model called the Longitudinal User Centered Influence model for identifying leaders and followers in online social networks.
2) We validated the classification accuracy of the LUCI model using a data set collected from Everything2.
3) We applied our model on a large real-world data set from Facebook. Since we do not have ground truth information for Facebook, we performed detailed analysis on the key topological properties of the friendship graph for different user categories that yielded interesting insights.

The rest of the paper proceeds as follows. We provide an overview of related work in Section II. In Section III, we provide the details of social network data sets used in our study. We then provide details of our proposed LUCI model for user categorization in Section IV. In Section V we validate the LUCI model by applying it to data set obtained from the Everything2 for which we have ground truth available. In Section VI, we also applied our LUCI model on a large data set from Facebook and further conducted detailed analysis on the key topological properties for different classes of users in the friendship graph. We finally conclude the paper in Section VII.

II. Related Work

We classify related work into the following four categories. To the best of our knowledge, this is the first work that identifies leaders and followers in online social networks using only interaction information.

The first category of related work rely on manual collection of data from research subjects. For example, Reagans *et al.* [18] used fixed-roster and free-recall approaches to collect survey data from 1518 project teams in a contract research and development firm. The goal was to study the impact of social network structure and social capital on team performance. Likewise, Amirkhanian *et al.* [2] collected survey data from 14 social networks in Russia and Bulgaria to identify leaders within them.

The second category of related work rely on social network graph heuristics such as PageRank and degree centrality. For example, Java *et al.* use PageRank to quantify the influence of a blog in the blogosphere [15]. The goal of their study is to filter out spam blogs whose PageRank values are typically lower than legitimate blogs. Goldenberg *et al.* use a degree centrality based heuristic to identify hubs in an online social network [12]. They further classify hubs, with large degree centrality, as either innovators or followers. We later show in our experimental results that both PageRank and degree centrality are ill-suited to identify leaders from interaction data.

The third category of related work rely on a hybrid of social network graph and interaction information. Hajian *et al.* propose an index called Magnitude Of Influence (MOI) to quantify users' influence on their neighbors in online social networks [14]. MOI is further weighted by the influence rank of neighbors using the PageRank algorithm to determine the final influence ranking of a user. The authors used a data set from FriendFeed to test their proposed approach; however, they did not evaluate its accuracy with respect to the ground truth. Khrabrov *et al.* combine PageRank scores of users with other dynamic information to quantify influence of users in online social networks [16]. By computing their proposed metric for users daily, they study how the influence of users evolves over time. They tested their approach using a Twitter data set; however, they also did not evaluate its accuracy with respect to the ground truth.

Finally, the fourth category of related work mine content of blogs to identify opinion leaders in the blogosphere. Bodendorf *et al.* [3] mine blog contents to generate text features using linguistic and statistical analysis. They apply support vector machines to assign opinion labels to all users. Song *et al.* [19] identify opinion leaders in the blogosphere using both link and content information. They first use Latent Dirichlet Allocation to generate a topic space and then use cosine similarity to measure information novelty of a blog with respect to other blogs. Note that content information is openly available in blogosphere; however, it is not typically available in online social networks due to privacy concerns.

III. Data Set

In this section, we provide details of the social network data sets used in our study. A major challenge we faced in this study was the unavailability of social network data sets with labeled "ground truth" information about leaders or followers. To the best of our knowledge, no publicly available social network data set has labeled ground truth information. For instance, none of the prior related work used ground truth information for validation [3], [12], [14]–[16], [19]. Fortunately, we were able to get hold of labeled data from an online social network called Everything2. We also used unlabeled data from Facebook for manual analysis and characterization.

A. Everything2

We examined the network of messages sent within a wiki-like online peer-production social network called Everything2. It is an online, content generation social network (similar to Wikipedia) and is more than ten years old. Everything2 also has tools that allow users to send asynchronous private messages to each other. This pattern of messages (*i.e.* interactions) among users creates a social network, with users as the nodes and the messages as directed edges. Everything2 also has an active community of users who maintain the site entirely through volunteer efforts. These users are called *administrators* or *leaders*. This labeling allows us to obtain a ground truth in detecting leader and follower roles using only interaction patterns of users. The leaders in Everything2 are allowed to perform "super actions", such as deleting pages, rating pages, *etc.* Note that normal users cannot perform these "super actions". We use the interaction information among Everything2 users to detect who might be a leader and then compare that prediction against the actual list of leaders to compute classification accuracy. The data set collected for this study spans the time period of $2002-2008$. Each record in this data set contains time stamped user interaction information along with both user identifiers. For each user, the information whether he is also a leader is provided by a binary flag. This data set contains a total of approximately 8 thousand users and contains a total of 3.9 million interactions among these users. Among the 8 thousand total user population, approximately $1,500$ users are flagged as leaders.

B. Facebook

We use a publicly available data set collected from Facebook by Wilson *et al.* [24]. This data set consists of two types of graphs. First, we have an undirected friendship graph in which vertices represent users and edges represent the friendship between two users. Second, we have a directed pair-wise user interaction graph in which the vertices represent users and the directed edges represent the interaction from one user to another. The interaction data spans a time duration of one year. The collected data set contains more than 3 million users and 23 million friendship edges. The ratio of the total number of friendship edges to the total number of users in the data set is ≈ 7.6. This highlights that a significant fraction of users in the collected data have only a few friendships. Note that we do not have ground truth data for Facebook users; therefore, we only analyze several important topological properties of the friendship graph for different user categories.

IV. Proposed Approach

In this section, we provide details of our proposed model for identifying leaders and followers. Before we introduce its details, we first define some notation and terminology. Then, we provide an overview of the Friedkin-Johnsen (FJ) influence model that provides a basis of our proposed LUCI model.

A. Notation and Terminology

The relationships and interactions among a group of N people in social networks are captured by friendship and interaction graphs. The friendship graph of the users is denoted by an undirected and unweighted graph $G(V, E)$. Here V denotes the set of vertices with elements $\{v_1, v_2, \ldots, v_N\}$, where each vertex represents a user in a social network. Users connect to other users in the social network by friendship ties, which are denoted by the set of edges E. Each edge is a tuple of the form (v_i, v_j) that represents a friendship tie between users v_i and v_j. We also represent the friendship graph $G(V, E)$ as an adjacency matrix $\mathbf{W}$ of size $N \times N$. The friendship ties considered in our study are bidirectional; therefore, $\mathbf{W}$ is an unweighted, symmetric matrix. This means if users v_i and v_j are friends then the entry in the i-th row and j-th column (denoted $w_{i,j}$) as well as the entry in the j-th row and i-th column (denoted $w_{j,i}$) of $\mathbf{W}$ are both set to 1. If v_i and v_j are not friends then $w_{i,j}$ and $w_{j,i}$ will both be 0.

The interaction graph of the users is denoted by a directed and weighted graph $I(V, U)$. Each element of U is a 3-tuple of the form $(v_i, v_j, m_{i,j})$, which represents that user v_i has generated $m_{i,j}$ interactions with user v_j. Furthermore, we define a vector $\mathbf{y}(t)$ of length N. Its i-th entry $y_i(t)$ denotes the aggregate number of interactions generated by v_i to all single hop neighbors during time period t, *i.e.* $y_i(t) = \sum_{i'=1, i' \neq i}^{N} m_{i,i'}(t)$. We also represent the interaction graph $I(V, U)$ as an adjacency matrix $\mathbf{M}$ of size $N \times N$. Since an interaction between two users is initiated by one of them, the adjacency matrix $\mathbf{M}$ representing the interaction graph is a weighted and non-symmetric matrix. This means that if a user v_i initiates $m_{i,j}$ number of interactions with another user v_j during a time period t then the entry in the i-th row and j-th column (denoted $m_{i,j}(t)$) is set to $m_{i,j}$. Naturally, if two users v_i and v_j are not friends (*i.e.* $w_{i,j} = w_{j,i} = 0$) they cannot have interactions and $m_{i,j}(t) = m_{j,i}(t) = 0$.

We assume time to be divided into discrete and non-overlapping periods of equal duration, indexed by t. In our experiments, the duration of time periods is set to be one month for Facebook data and six months for Everything2 data. The duration selected for Everything2 is larger than Facebook because interaction data in Everything2 is more sparse in time. Although the time variable t is discrete, we do not use square bracket notation to avoid confusion with matrices.

B. The FJ Influence Model

The FJ influence model [10] relates the interaction behavior $y_i(t)$ of a user v_i during time period t to its interaction behavior $y_i(t-1)$ in the preceding time period $t-1$ by Equation 1.

$$y_i(t) = \alpha(t) \sum_{i'=1}^{N} w_{i,i'} y_{i'}(t-1) + \beta(t) y_i(t-1) + e_i \qquad (1)$$

According to this model, the interaction behavior of a user at time t is a linear function of its interaction behavior in the previous time period $t-1$ and the combined influence exerted on it by its neighbors. The FJ influence model makes the following assumptions about the evolution of users' interaction behaviors [11]:

- *Cognitive Weighted Averaging:* Users' interaction behaviors evolve by a process of weighted averaging of their neighbors' and their own previous interaction behaviors.
- *Fixed Social Structure:* The social network that exists between users does not change.
- *Decomposability:* The process by which users' interaction behaviors evolve can be decomposed into discrete non-overlapping periods of time.
- *Simultaneity:* The simultaneous equations accurately predict the transmission of influence during corresponding time period.

The model assigns a weight to the combined external influences from neighbors on a user in the previous time period $t-1$. $\alpha(t)$ is a scalar weight of the endogenous influence. The model also assigns a weight to the influence exerted by users on their neighbors in the previous time period $t-1$. $\beta(t)$ is a scalar weight of the exogenous influence. Note that because the FJ model is cross sectional, both scalars $\alpha(t)$ and $\beta(t)$ in Equation 1 are functions of time t. The last term on the right hand side is the model error e_i, which denotes the deviation of model prediction from the interaction behavior $y_i(t)$ of the user v_i at time t. Error terms $e_i, \in \mathbb{N}, 1 \leq i \leq N$ ($\mathbb{N}$ is the set of natural numbers) are assumed to be independent identically distributed (iid) Gaussian with zero mean and variance σ^2 (see Frank and Fahrbach [9]). The influence model in Equation 1 can be expressed in the matrix form for all users simultaneously as follows.

$$\mathbf{y}(t) = \alpha(t)\mathbf{W}\mathbf{y}(t-1) + \beta(t)\mathbf{y}(t-1) + \mathbf{e}, \qquad (2)$$

where $\mathbf{e} = [e_1, e_2, \cdots e_N]^t$ is the vector of model errors of length N. Often times the two influence terms, $\mathbf{W}\mathbf{y}(t-1)$ and $\mathbf{y}(t-1)$, on the right hand side of Equation 2 are each normalized over all users to have unit variance. The selection of the optimum weights $\alpha(t)$ and $\beta(t)$ that minimize the mean square error of the model becomes an optimization problem which is solvable using linear regression. The optimization of two weights $\alpha(t)$ and $\beta(t)$ is based on a system of N simultaneous equations. The relative values of $\alpha(t)$ and $\beta(t)$ provide insights into the role of extrinsic and intrinsic influence on the evolution of user interaction behavior.

C. Proposed Model

We use the FJ influence model as our basis of extension because it is a generalization of a large body of theoretical work on *social network influence theory*. An interested reader is referred to the summary of prior theoretical work on social network influence theory and further analytical analysis reported by Friedkin *et al.* in [10], [11]. However, a major limitation the FJ influence model is that $\alpha(t)$ and $\beta(t)$ only provide a summary statistic for the whole social network, *i.e.* they cannot be used to infer the relative behavior of individual users. Conceptually, we generalize the FJ influence model to allow all users to have different extrinsic and intrinsic influence factors in the LUCI model. Mathematically, we introduce the following changes in its formulation to accommodate the aforementioned generalization. Every user v_i has two weights terms ρ_i and γ_i, equivalent to $\alpha(t)$ and $\beta(t)$. In order to be able to solve for ρ_i and γ_i for each user v_i, we partition the interaction data into t_{max} time periods, indexed by t and $t \in \mathbb{N}$, each of equal duration. The information for time period t is stored in a *messaging matrix* of size $N \times N$ denoted by $\mathbf{M}(t)$. The entry in the i-th row and j-th column of $\mathbf{M}(t)$ is $m_{i,j}(t)$ denoting the number of interactions from user v_i towards user v_j during the time period t. We obtain the following system of equations for every user v_i if the flow of influence in the formal influence model in Equations 1 and 2 is replaced by the user interactions mentioned above.

$$\sum_{i'=1}^{N} m_{i,i'}(t) = \rho_i \sum_{i'=1}^{N} m_{i',i}(t-1) + \gamma_i \sum_{i'=1}^{N} m_{i,i'}(t-1) + e_i \qquad \forall t \in \mathbb{N}, 1 \leq t \leq t_{max} \tag{3}$$

According to this model, the interaction behavior of a particular user v_i at time t is a linear function of its interaction behavior and the combined interactions of its neighbors in the previous time period $t-1$. The model assigns a weight to the degree to which a user's neighbors' interaction behaviors $m_{i',i}(t-1)$ in the previous time period $t-1$ influence its interaction behavior $m_{i,i'}(t)$ in the current time period t. This assigned weight is called the *network coefficient*. The network coefficient of a user v_i is denoted ρ_i. The model also assigns a weight to the degree to which a user's interaction behavior $m_{i,i'}(t-1)$ in the previous time period $t-1$ influences its interaction behavior $m_{i,i'}(t)$ in the current time period t. This assigned weight is called the *ego coefficient*. The ego coefficient of a user v_i is denoted γ_i.

Comparing the terms in Equation 3 with those in Equation 1 we can draw some parallels that are described below. The leader/follower categorization model equates the interaction behavior of users with the number of interactions directed towards their neighbors. Similarly, the combined influence from neighbors is equated with the total number of interactions generated by friends on their profile. But the most important departure is the assignment of a separate (ρ_i, γ_i) tuple to each user v_i.

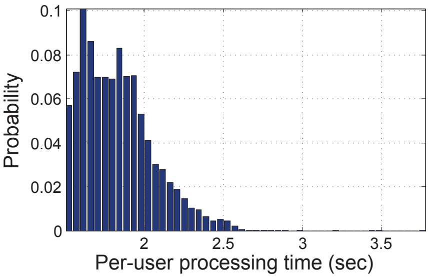


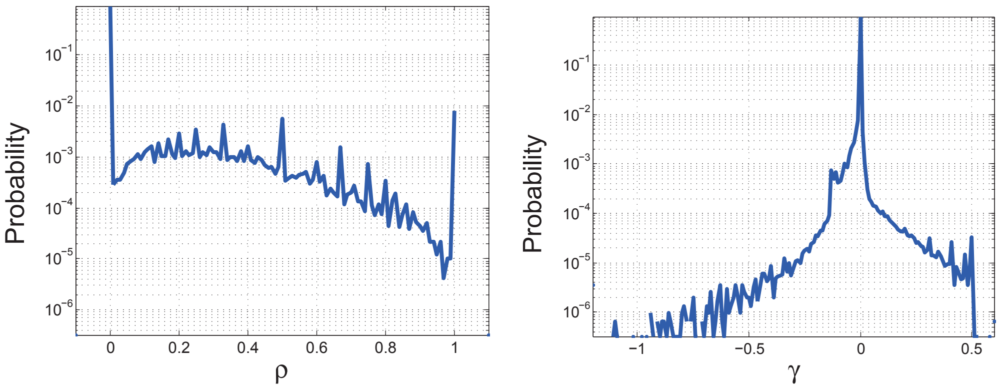
Fig. 1. Distribution of execution time for individual users

A data set that gives us t_{max} messaging matrices can be used to set up at most $t_{max}-1$ equations for each user to compute the (ρ_i, γ_i) tuple. Once a system of equations like the one in Equation 3 has been set up for a user v_i, the only remaining unknowns are (ρ_i, γ_i), which are optimized for the least square error $\mathbf{e}_i$ over at most $t_{max}-1$ number of equations. It is noteworthy that the LUCI model has an advantage over the FJ influence model in terms of performance and complexity. Recall that the FJ influence requires to solve a system of N simultaneous linear equations. Instead, the LUCI model requires to solve a system of $t_{max}-1$ simultaneous linear equations for all users. Since $t_{max} \ll N$ for most practical cases; therefore, the dimensionality of the underlying problem is significantly reduced. In fact, as shown in Figure 1, our preliminary study showed that it takes less than 2 seconds on average to solve Equation 3 for a user. Furthermore, N systems of $t_{max}-1$ simultaneous linear equations can be solved in parallel because they are not interdependent.

D. Node Clustering

(ρ, γ) tuples quantify the behaviors of all users. The marginal histograms of ρ and γ are plotted in Figure 2. Figure 2(a) is the marginal pdf of ρ, the network coefficient, plotted on semi-log axes. It is one-tailed resembling the exponential distribution. In Figure 2(b), we plot the marginal pdf of γ, the ego coefficient. It is a two tailed distribution that resembles the Laplacian distribution.

To systematically analyze the behavior of all users, we aim to cluster them based on the values of two random variables ρ and γ. However, there is no obvious structure observable from their scatter plot shown in Figure 3. A well-known method used to cluster such nonlinearly separable data is kernel k-means [5]. In kernel k-means, the data points are mapped from the input space to a higher dimensional feature space through a nonlinear transformation. A well-known approach is to define a set of reference points in the original space and the distances to the reference points constitute the transform. The simple k-means clustering is performed after applying the transformation. The simple k-means clustering is preferred over other clustering methods, such as hierarchical clustering, because it can deal with large data sets [17]. To cluster the users into distinct groups, we apply the unsupervised kernel k-means algorithm to the pair of ρ and γ values. To apply k-means, we have to select the suitable value of k, which is an open research problem. We use a well-known heuristic called the gap statistic, which is based on the change in intra-cluster dissimilarity for increasing values of k. The value of k is selected to be the elbow of intra-cluster dissimilarity as the value of k is varied [20]. Using this methodology, we clustered all users



(a) Network coefficient

(b) Ego coefficient

Fig. 2. Marginal distributions of network and ego coefficients of users in Facebook data set plotted on semi-log scale

into four clusters ($k = 4$) that occupy distinct regions in the ρ-γ plane. These regions are also marked in Figure 3. We label the four clusters based on our understanding of ρ and γ values, mentioned in Section IV-C. • High ρ, Zero γ: Recall that a high value of ρ reflects the strong impact of the network effect on a user's tendency to have outward interactions with neighbors. The values of γ close to 0 indicate a weak or non-existent relationship between activity levels in consecutive time periods. A user with a high ρ and negligible γ is the one whose outgoing interaction activity is determined solely by the received interaction activity. Such users' activities are driven, and in a sense controlled, by those of their neighbors. Based on this understanding we label such users as *Followers*, denoted as F.

• Low ρ, Positive γ: Recall that a high and positive value of γ reflects the strong impact of users' interaction activity in the preceding time period on their interaction activity in the following time period. Positive values of γ indicate a strong relationship between activity levels in consecutive time periods. A user with a high value of γ and negligible value of ρ is the one whose outgoing traffic volume in a particular time period is more strongly influenced by that in the previous time period, rather than network effects. To understand the implications of a low ρ and positive γ refer back to Equation 3. A positive γ implies that for that particular user the number of outgoing interactions is similar to the same number in the previous time period. However, the low value of ρ means that this number is not significantly affected by how many interactions it receives in the preceding period. Such users' activities are driven, and in a sense controlled, by themselves and not by their neighbors. These users are very communicative and are actively involved with their neighbors. Based on this understanding we label such users as *Extrovert Leaders*, denoted as EL.

• Low ρ, Negative γ: Recall that a low and negative value of γ reflects a strong inverse impact of users' messaging activity in the preceding time period on their messaging activity in the following time period. Negative values of γ indicate a strong relationship between a user's activity levels in consecutive time periods. A user with a low γ and ρ values is the one whose outgoing interaction activity in a particular time slot is more strongly influenced by that in the previous time period, rather than network effects. However, unlike extrovert leaders, this type of user generates fewer outgoing messages. Such users' activities are driven, and in a sense controlled, by themselves and not by their neighbors. These users are actively involved with their neighbors in the sense that they regularly receive a significant number of inward interactions from neighbors but send out very little in return. We label such users as *Introvert Leaders*, denoted as IL, based on our understanding.

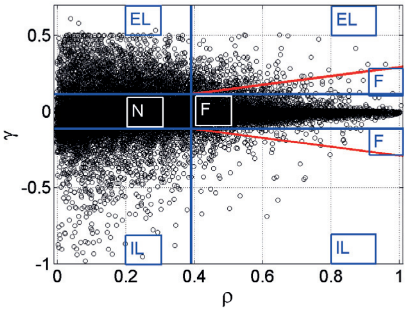
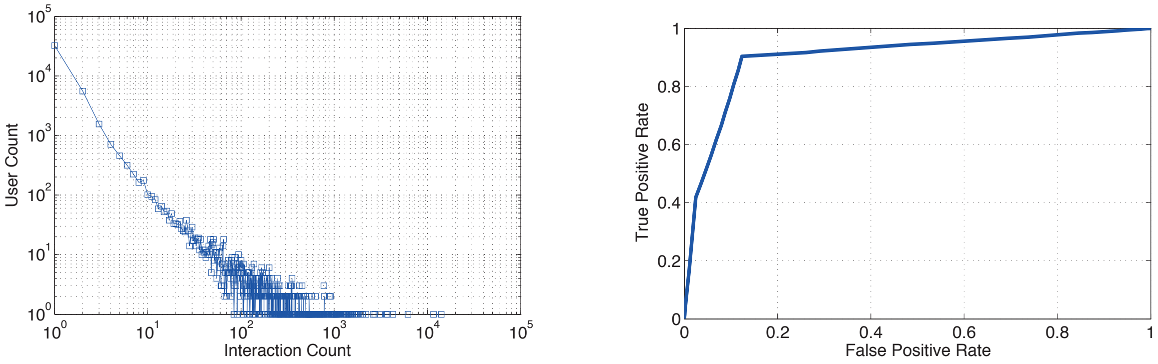


Fig. 3. Scatter plot of ρ and γ for all users in Facebook data set

• Low ρ, Zero γ: Users whose ρ and γ coefficients are 0 or approach it are either inactive, or have incoming and outgoing interaction activities that have little or no correlation. Their traffic flows may be characterized as random or independent. Based on this understanding we label such users as *Neutrals*, denoted as N.

V. Everything2 Validation Results

In this section, we use a data set collected from Everything2 to validate the accuracy of LUCI model. Figure 4(a) plots the distribution of the number of interactions for all users in this data set. This distribution is highly skewed, *i.e.* a small fraction of users have a large number of interactions while others have very few interactions. We convert the x-axis and the y-axis to logarithmic scale to emphasize this observation. The distribution curve roughly follows a straight line on log-log scale, thereby providing evidence of a heavy-tailed distribution. To compute the parameters of the LUCI model



(a) User activity distribution

(b) ROC plot

Fig. 4. Results for Everything2 data set

(ego and network coefficients), the interactions are divided into equal slices of six months each. For each slice, users and messages are modeled as nodes and directed interactions, respectively.

A. Evaluation

We now compute the ego and network coefficients of all users from the interaction information in the Everything2 data set. Similar to the unsupervised classification scheme in Section IV-D, we use a well-known kernel-based supervised classification algorithm to accurately detect leaders in Everything2. Specifically, we use a support vector classifier with the radial basis kernel function to identify leaders and followers in Everything2. We use the standard stratified 10-fold cross validation to ensure that the classification results of LUCI model are generalizable.

We report the results of our support vector classifier in terms of standard Receiver Operating Characteristics (ROC) [8]. The most comprehensive ROC based accuracy measure is called Area Under the Curve (AUC), which denotes the area under the ROC curve. AUC varies in the range of $[0, 1]$, where 1 represents the perfect classification accuracy. Note that we only feed two features, called ego and network coefficients, for each user in the support vector classifier. The experimental evaluation of this classifier using Everything2 data set provide AUC $= 90.3\%$. Figure 4(b) shows the standard ROC curve for our support vector classifier. At the optimal operating point, it achieves an average accuracy of 89.0%, an average true positive rate of 87.7%, with a false positive rate of 9.5%. These results highlight the effectiveness of LUCI model to accurately identify leaders and followers in online social networks.

VI. Facebook Analysis Results

To further validate our LUCI model, we apply it to a data set collected from Facebook. However, due to the lack of ground truth data, we cannot conduct the analysis that we do for the Everything2 data set. Therefore, for different user categories, we analyze the following important topological properties of the friendship graph: vertex degree, number of triangles, clustering coefficient, eigenvector centrality, and average shortest path length. Our experimental results show that different user categories indeed exhibit different topological characteristics in the friendship graph and these observed characteristics are in accordance with the expected ones based on the general definition of the four roles of introvert leaders, extrovert leaders, followers, and neutrals.

We now define some notations used in this section. An undirected friendship graph is denoted by $G(V, E)$, where V and E represent the sets of vertices and edges of the graph G. Each vertex represents a user and each edge represents friendship or connection between two users. Here $|V|$ and $|E|$ denote the number of vertices and edges in the graph G, respectively.

A. Vertex Degree

Metrics: The number of edges connected to a given vertex v_i in a graph is called the degree of the vertex and is denoted d_i. In the context of friendship graphs in online social networks, a vertex corresponds to a user and the vertex degree represents the number of friends or connections of the user. In typical online social networks it is not uncommon for a single user to have hundreds of friends. Several studies have pointed out that it is not possible to regularly communicate with such a large circle of friends [6]. Other recent studies have also highlighted that only a small subset of these edges in online social networks truly represent the actual friendships or connections based on communication practices from overlaid interaction information [7]. However, in this study we do not take into account such differences and do not distinguish between friendships based on interaction information, so all edges have equal weight. Though, we expect such differences to be reflected in our analysis of topological properties of the four user categories because users are grouped using overlaid interaction information.

Results: Figure 5 shows the distribution of vertex degree for the four different user categories. We first note that users belonging to neutral class have the least average degree as compared to users in the other three categories. For the neutral category, 99% of the users have less than 100 friends. The degree distribution of followers most closely resembles with that of neutrals but the average degree of followers is significantly larger.[1] It is evident that the shape of both

[1] Statistical significance is verified using two-sample t-test to compare sample means, test statistic: -296.4, p-value < 0.001.

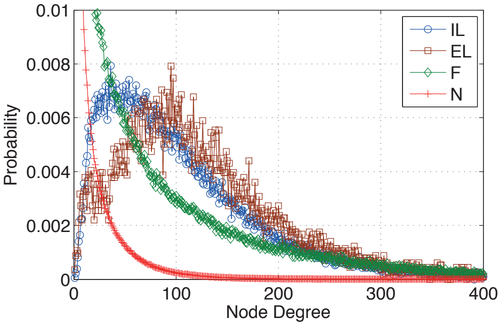


Fig. 5. Distribution of vertex degree

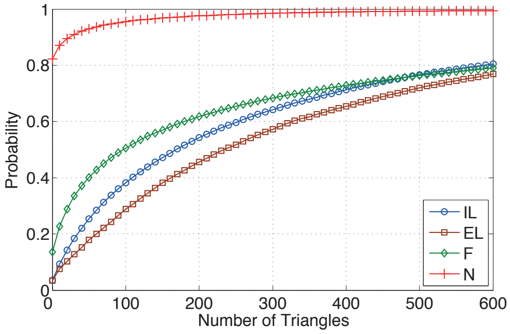


Fig. 6. Distribution of the number of triangles

distributions matches that of a geometric distribution. On the other hand, the average degrees of both introvert and extrovert leaders are significantly larger than those of followers and neutrals.[2] In contrast to followers and neutrals, the shape of the degree distributions for both types of leaders closely matches a Poisson distribution. It is interesting to note that the average degree of extrovert leaders is significantly larger than that of introvert leaders.[3]

Insights: The observed differences are closely linked to our intuition about the definitions of different user classes. As previously mentioned in Section IV, extrovert leaders have low values of ρ and highly positive values of γ, indicating that they reach out and have a lot of outwards activity. Salespersons are a typical example of such types of users. These users are most well-connected as compared to all other user classes. Introvert leaders have low values of ρ and highly negative values of γ, indicating that they have little outwards interactions and tend to be at the receiving end of most interactions. Such users are highly sought-after individuals such as heads of different organizations. Introvert leaders have less friends than extrovert leaders but are still reasonably well-connected. Followers, whose activity is primarily driven by the activity of their neighbors, are seen having fewer connections and have a degree distribution that is more concentrated in the low degree range. Neutrals, who have little interaction with others, exhibit greater disengagement from other users and have even fewer connections than followers. It is interesting to note that a large number of introvert and extrovert leaders have low vertex degrees. This contradicts the definition of leaders proposed by Goldenberg *et al.* that leaders have larger vertex degree than the rest of the user population in an online social network [12].

B. Number of Triangles

Metrics: Triangles in a friendship graph capture the notion that *friend of friend is a friend*. For each vertex, we count the number of triangles that includes the given vertex as one of its three vertices.

Results: It is interesting to analyze the results for the number of triangles that a user belong to with respect to vertex degrees. For a vertex with degree d, the maximum possible number of unique triangles that this vertex belongs to is $\binom{d}{2}$. A vertex belongs to $\binom{d}{2}$ triangles if its neighbors form a clique (*i.e.*, any two of its neighbors have an edge between them). We observed in Figure 5 that introvert and extrovert leaders have approximately 50 and 100 friends, which is significantly larger than followers or neutrals. Therefore, the maximum possible number of triangles for leaders is typically much larger than that for followers. Figure 6 plots the cumulative distribution of number of triangles for the four user categories. We note that neutrals are part of fewer number of triangles than all other classes. The other three user categories have roughly similar distribution in terms of the number of triangles that a user belong to. Clearly, the difference in vertex degree does not translate into number of triangles for different classes of users.

Insights: We can explain this observation by looking at interaction behavior of different user categories. Extrovert leaders reach out and connect to their neighbors even when receiving lesser inwards interactions. Furthermore, they have the highest degree as compared to other user classes. We can speculate that they connect to diverse set of people that have fewer common links. Therefore, the number of triangles for extrovert leaders is lower than expected. A similar reasoning holds true for introvert leaders with the difference that they mostly receive interactions from their neighbors. Followers, on the other hand, have fewer number of neighbors and are part of more tightly connected communities with a lot of common neighbors. Therefore, followers have a large number of triangles even though their average degree is much smaller than introvert and extrovert leaders. A measure that captures the ratio of the number of triangles to the number of edges is called clustering coefficient and is discussed next.

C. Clustering Coefficient

Metrics: The clustering coefficient c_i of a vertex v_i is defined as the ratio of the number of existing edges among v_i and v_i's neighbors and the number of all possible edges among them. Using Δ_i to denote the number of triangles containing vertex v_i and d_i to denote the degree of vertex v_i, we define the clustering coefficient of vertex v_i as:

$$c_i = \frac{\Delta_i}{\binom{d_i}{2}} = \frac{2\Delta_i}{d_i(d_i-1)} \tag{4}$$

[2] Statistical significance is verified using two-sample t-test to compare sample means. Introvert leaders-followers test statistic: -26.5, p-value $<$ 0.001, Introvert leaders-neutrals test statistic: -252.6, p-value $<$ 0.001, Extrovert leaders-followers test statistic: -34.7, p-value $<$ 0.001, Extrovert leaders-neutrals test statistic: -134.2, p-value $<$ 0.001.

[3] Statistical significance is verified using two-sample t-test to compare sample means, test statistic: -19.9, p-value $<$ 0.001.

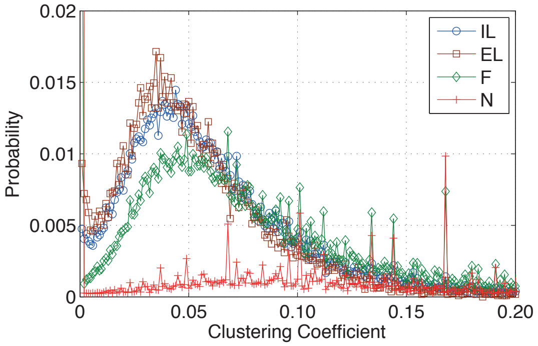


Fig. 7. Distribution of clustering coefficient

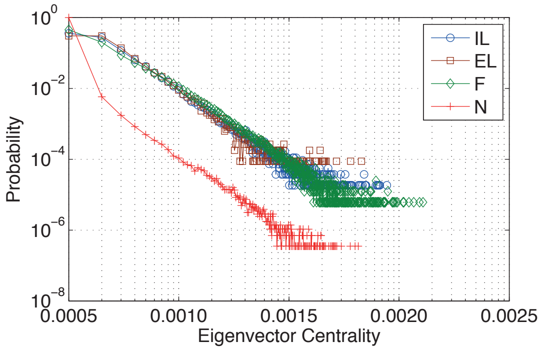


Fig. 8. Distribution of eigenvector centrality

Results: Figure 7 shows the distribution of clustering coefficient for all four categories. The clustering coefficient of most neutral nodes is zero indicating that they are not part of any triangles. On the other hand, followers have the largest average value of clustering coefficient. The average clustering coefficient of introvert and extrovert leaders is larger than followers. Between introvert and extrovert leaders, introvert leaders have larger average clustering coefficient.

Insights: To explain this observation, we have to jointly understand the number of triangles and the degrees for different categories of users. We have previously observed that the distribution of the number of triangles is similar for followers and both types of leaders. Also, the average degree of followers is significantly lesser than that of both types of leaders. Since clustering coefficient is merely a ratio of the number of triangles and the number of possible triangles (determined by vertex degree), this translates into larger average values of clustering coefficient for followers as compared to both types of leaders. Both types of leaders have similar degree distribution but extrovert leaders have a significantly larger average degree than introvert leaders. This translates into larger values of average clustering coefficient for introvert leaders than extrovert leaders.

D. Eigenvector Centrality

Metrics: Eigenvector centrality is a well-known measure to quantify the importance of a vertex in a network. Degree centrality counts the number of edges to neighbors and weights them equally. Eigenvector centrality not only counts the number of edges to neighbors but also weighs them by the neighbors' respective eigenvector centralities [13]. To define eigenvector centrality x_i of vertex v_i, let $\mathbf{W}$ denote the adjacency matrix of the graph where $w_{i,j}$ is 1 if an edge exists between vertices v_i and v_j, and 0 otherwise.

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{N} w_{i,j} x_j \tag{5}$$

where λ is the principal eigenvalue of matrix $\mathbf{W}$.

Results: Figure 8 shows the distribution of eigenvector centrality for different user classes. Note that we have plotted the distribution of eigenvector centrality on log-log scale. We observe straight lines for all user categories exhibiting the characteristic of power-law distributions. There is no significant difference among eigenvector centrality distributions of introvert leaders, extrovert leaders and followers. For neutrals, the slope of eigenvector centrality distribution is much steeper indicating more skewness.

Insights: We would have expected leaders to have higher values of eigenvector centrality than followers [15]. However, this is not the case in our results. To explain our results, we have to revisit the way eigenvector centrality is computed. Recall that neutrals and followers are the two largest categories in our data but have a lower average degree than leaders. We speculate that leaders, though with higher degree themselves, are less connected to other leaders. This implies that most edges of leaders are shared with followers and neutrals. Such edges are weighed lesser in computation of eigenvector centrality for leaders. This results in smaller than expected average values of eigenvector centrality. Likewise, our results indicate that followers, although having lower degree, are connected to several high degree vertices (mostly leaders). The links with leaders (with higher average degree) boosts the value of eigenvector centrality for followers.

E. PageRank

Metrics: Page Rank is a variant of eigenvector centrality. The two main differences between Page Rank and eigenvector centrality are different scaling factors and the use of left hand and right hand eigenvector, respectively.

Results: Figure 9 shows the distribution of PageRank for different user classes. Similar to the distribution of eigenvector centrality, we have plotted the distribution of PageRank on log-log scale. We again observe roughly straight lines for all user categories exhibiting the characteristic of power-law distributions. However, unlike eigenvector centrality, we see differences among the distributions of introvert leaders, extrovert leaders and followers. Specifically, the PageRank distribution of extrovert leaders is least skewed, followed by introvert leaders and followers. The slope of PageRank distribution is steepest, similar to eigenvector centrality, indicating most skewness.

Insights: Unlink eigenvector centrality, the average PageRank values of leaders are higher than followers. This indicates that the scaling factor and the use of left hand eigenvector helps emphasize the differences among the two leader categories and followers.

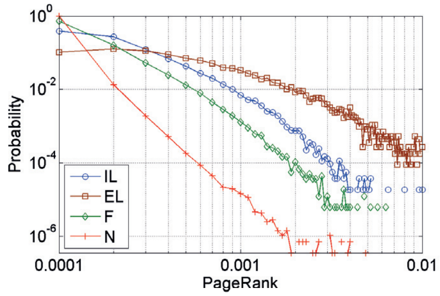


Fig. 9. Distribution of PageRank

F. Average Shortest Path Length

Metrics: The average shortest path length ($\bar{l_i}$) denotes the average of number of steps along the shortest paths from a given vertex v_i to the rest of the vertices in a graph. Using this notation, we define average (shortest) path length for a given vertex in a graph as:

$$\bar{l_i} = \frac{\sum_{\forall v_{i'} \in V, i \neq i'} l_{i,i'}}{|V| - 1} \quad (6)$$

Due to the small-world effect, social networks tend to have smaller average values of average shortest path length compared to random graphs with the same number of vertices and edges [1].

Results: Figure 10 shows the distribution of average distance for different classes of users. It is interesting to note that introvert leaders have the smallest values of average shortest path lengths to other vertices in the friendship graph. As expected, neutrals have the highest average shortest path length to other users. Extrovert leaders and followers closely follow introvert leaders in shorter average shortest path lengths.

Insights: The observation that introvert leaders have smaller average shortest path lengths to other vertices in a friendship graph is interesting because they have lower average degree than extrovert leaders. In random graphs, one would expect vertices with a higher degree to have smaller average shortest path lengths. We revisit the definition of introvert leaders to explain this finding. Recall that introvert leaders are people with whom users of other categories are eager to interact. Thus, introvert leaders are the most sought-after users who have more inwards interactions. This interaction behavior may lead to a network structure where the average shortest path length to introvert leaders is lower than expected. Or equivalently, introvert leaders can reach other users in a friendship network in the least number of steps on average. Higher average shortest path lengths of followers and neutrals compared to both types of leaders are a result of their low connectivity.

VII. Conclusions

This paper presents a first step towards clustering users in online social networks into the four categories of introvert leaders, extrovert leaders, followers, and neutrals using user interaction information. We make three key contributions in this paper. First, we propose a new model, called the Longitudinal User Centered Influence (LUCI) model, for this categorization. Second, we validated our model on a Everything2 data set. Third, we further validated our model on a Facebook data set. Experimental results on both data sets show that our model is able to capture the characteristic differences of these four user categories. In future, we also plan to conduct a systematic analysis of sensitivity of our results to the size and number of windows in the available interaction data. We also plan to use the LUCI model to identify leaders and followers in other online social networks. Furthermore, we plan to compare the results of LUCI model using topic specific user interaction information in online social networks.

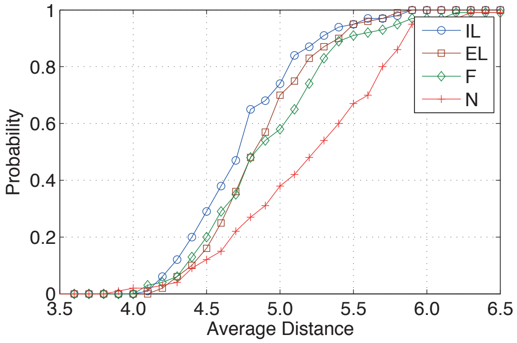


Fig. 10. Distribution of average shortest path length

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] Y. Amirkhanian, J. Kelly, E. Kabakchieva, T. McAuliffe, and S. Vassileva. Evaluation of a social network HIV prevention intervention program for young men who have sex with men in Russia and Bulgaria. *AIDS Education and Prevention*, 15(3):205–220, 2003.

[3] F. Bodendorf and C. Kaiser. Detecting opinion leaders and trends in online social networks. pages 65–68, 2009.

[4] G. Cox, F. Rosenbluth, and M. Thies. Mobilization, social networks, and turnout: Evidence from Japan. *World Politics*, 50(3):447–474, 1998.

[5] I. S. Dhillon, Y. Guan, and Y. Guan. Kernel kmeans, spectral clustering and normalized cuts. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[6] N. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook 'friends': Exploring the relationship between college students' use of online social networks and social capital. *J. Computer-Mediated Communication*, 12(3), 2007.

[7] N. Ellison, C. Steinfield, and C. Lampe. Connection strategies: Social capital implications of facebook-enabled communication practices. *New Media & Society*, 2011.

[8] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. *TR HPL-2003-4*, HP Labs, USA, 2004.

[9] K. Frank and K. Fahrbach. Organization culture as a complex system: balance and information in models of influence and selection. *Organization Science*, 10(3):253–277, 1999.

[10] N. Friedkin and E. Johnsen. Social influence and opinions. *The Journal of Mathematical Sociology*, 15(3):193–206, 1990.

[11] N. Friedkin and E. Johnsen. Social influence networks and opinion change. *Advances in Group Processes*, 16:1–29, 1999.

[12] J. Goldenberg, S. Han, D. R. Lehmann, and J. W. Hong. The role of hubs in the adoption processes. *J. Marketing, American Marketing Association*, 2008.

[13] J.L. Gross, J. Yellen. Handbook of Graph Theory. *CRC Press*, 2003.

[14] B. Hajian and T. White. Modelling influence in a social network: Metrics and evaluation *Proc. IEEE International Conference on Social Computing*, 2011.

[15] A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the spread of influence on the blogosphere. In *Proc. 15th International World Wide Web Conference*, 2006.

[16] A. Khrabrov and G. Cybenko. Discovering influence in communication networks using dynamic graph analysis. *Proc. IEEE International Conference on Social Computing*, 2010.

[17] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symposium on Math Statistics and Probability*, 1967.

[18] R. Reagans, E. Zuckerman, and B. McEvily. How to make the team: Social networks vs. demography as criteria for designing effective teams. *Administrative Science Quarterly*, 49(1):101–133, 2004.

[19] X. Song, Y. Chi, K. Hino, and B. Tseng. Identifying opinion leaders in the blogosphere. In *ACM conference on Conference on information and knowledge management*, 2007.

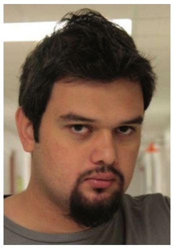
[20] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Statistical Society: Series B (Statistical Methodology)*, 63:411–423, 2001.

[21] P. Topham. Finding key opinion leaders using large scale social network analysis - a comparative analysis of methods for finding key opinion leaders. Technical report, Lnx Research, 2007.

[22] T. Valente, B. Hoffman, A. Ritt-Olson, K. Lichtman, and C. Johnson. Effects of a social-network method for group assignment strategies on peer-led tobacco prevention programs in schools. *American Journal of Public Health*, 93(11):1837, 2003.

[23] D. Watts. Networks, dynamics, and the small-world phenomenon. *American J. Sociology*, 105(2):493–527, 1999.

[24] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao. User interactions in social networks and their implications. In *Proc. ACM European Conference on Computer Systems*, 2009.



M. Zubair Shafiq received B.E. degree in Electrical Engineering from National University of Sciences and Technology (NUST), Pakistan, in 2008. He is currently pursuing the Ph.D. degree in computer science at Michigan State University. He was with Next Generation Intelligent Networks Research Center (nexGIN RC), Pakistan as a researcher from 2007 to 2009. His research interests include measurement and modeling of cellular networks and online social networks, and computer and network security.

Muhammad U. Ilyas did his Post-doctoral work at the Department of Computer Science & Engineering and later jointly with the Department of Electrical & Computer Engineering at Michigan State University (MSU) from 2009-2011. He received his Ph.D. and MS Electrical Engineering from MSU in 2009 and 2007, respectively. He received his MS Computer Engineering from the Lahore University of Management Sciences (LUMS), Lahore, Pakistan in 2005, and his BE Electrical Engineering from the National University of Sciences & Technology (NUST) in 1999. He is currently an Assistant Professor in the Department of Electrical Engineering at the School of Electrical Engineering & Computer Science (SEECS) of the National University of Sciences & Technology. His research interests include system modeling and measurement, social network analysis, networking, algorithms, and security.

Alex X. Liu received his Ph.D. degree in computer science from the University of Texas at Austin in 2006. He is currently an assistant professor in the Department of Computer Science and Engineering at Michigan State University. He received the IEEE & IFIP William C. Carter Award in 2004 and an NSF CAREER award in 2009. He received the MSU College of Engineering Withrow Distinguished Scholar Award in 2011. His research interests focus on networking, security, and dependable systems.

Hayder Radha received the Ph.M. and Ph.D. degrees from Columbia University (1991 and 1993). He is a Professor of Electrical and Computer Engineering (ECE) at Michigan State University (MSU). He was a Philips Research Fellow and a Distinguished Member of Technical Staff at Bell Laboratories. Dr. Radha is an IEEE Fellow. He is an elected member of the IEEE Technical Committee on Image, Video, and Multidimensional Signal Processing (IVMSP) and the IEEE Technical Committee on Multimedia Signal Processing (MMSP).