Final version of the accepted paper. Cite as: *M. Wani, M. A. Alrubaian, and M. Abulaish, A User-Centric Feature Identification and Modeling Approach to Infer Social Ties in OSNs, In Proceedings of the 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS'13), Vienna, Austria, ACM, pp. 107-114, Dec. 2-4, 2013..*

A User-Centric Feature Identification and Modeling Approach to Infer Social Ties in OSNs

Mudassir Wani
Center of Excellence in
Information Assurance
King Saud University
Riyadh, Saudi Arabia
mudasirwani7@gmail.com

Majed A. Alrubaian
College of Computer and
Information Science
King Saud University
Riyadh, Saudi Arabia
first_majed@yahoo.com

Muhammad Abulaish*
Department of Computer
Science
Jamia Millia Islamia
New Delhi–110025, India
mAbulaish@jmi.ac.in

ABSTRACT

This paper aims to identify user-centric features to calculate the strength of social ties between Online Social Network (OSN) users, and models the same using Latent Space Model (LSM). The modeling approach processes a socio-centric user-set as the users are directly (friend) or indirectly (friend-of-friend) related to a seed (target) user, which makes it easier to identify social ties between users as compared to random sampling from a set of diverse OSN users. For a given user, interaction data up to two levels is modeled and analyzed to generate a user-centric social network. Eleven different features related to Facebook have been identified to calculate the strength of social ties between users. LSM is used to visualize relationships in user-centric historical data and to estimate the probability of social ties between OSN users. The users are plotted using LSM in a three-dimensional (3D) social space around a seed user, and a link probability function is devised to calculate the probability of link between any two users with respect to the persona of the seed user. A sphere of influence around each user demarcating its active influence area is also identified and discussed in this paper.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*Statistical, Structural*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Pattern Analysis*

General Terms

Design, Experimentation

Keywords

Social network analysis, feature identification, social network modeling, Link prediction.

*To whom correspondence should be made

1. INTRODUCTION

Online Social Networks (OSNs) are growing exponentially with a highly dynamic nature and very complex structure. Depending on the dynamicity of an OSN, it changes over different granularity of time. A less popular OSN may change in days, an average OSN may change in hours, and a very popular OSN may change within a fraction of second. How we perceive an OSN is just a snapshot of that OSN at a particular point of time. An OSN can change while its information is being processed; however, this time-lag can be neglected [3]. A change is induced into an OSN through various activities of the participating users. The huge size of OSNs requires intense computational power and often results into the under-utilization of available data.

Considering the factors of exponential growth, highly dynamic nature, inherent complex structure, and intense computational power requirement, a full-scale analysis of an OSN is not feasible as the associated computational cost is very high, which increases with increasing dimensionality of data. Consequently, it becomes harder and harder to understand graphs showing thousands of nodes or edges and overlapping elements, and a meaningful graphical representation of large OSNs is not trivial because of the number of elements to display [1]. Therefore, instead of modeling the whole network, this paper introduces a user-centric social network modeling approach, which builds the social network around a seed (target) user of interest and models her social interactions up to a certain expanse. A set of eleven different features related to Facebook network have been identified to model the strength of social ties between the users. Latent Space Model (LSM) is used to visualize relationships in user-centric historical data and to estimate the probability of social ties between the users. Users are plotted in a three-dimensional (3D) social space using LSM, and a link probability function is used to calculate the link probability between a pair of users in the underlying network. The link probability asserts the authority on a link, based on the interactions of the users and their closeness with respect to the seed user.

Since link probabilities are calculated based on the spatial position of users in a 3D socio-centric space, users who tend to exhibit similar behaviour/personality as that of the seed user can be identified and considered to possibly form a link either with the seed user directly or through other users of the same network. The links identified so do not necessarily imply that a physical link would be formed in near future (a friendship link in case of Facebook). However, the person(s) of interest within a certain link probability threshold value can be identified.

The rest of the paper is organized as follows. Section 2 presents a brief review of the related works on feature identification and modeling online social networks. Section 3 presents a brief description and formulation of identified features. Section 4 presents the proposed social network modeling approach. Section 5 describes the experimental setup and results. Finally, section 6 concludes the paper with possible future directions of work.

2. RELATED WORK

Traditional OSN modeling techniques involved processing large quantity of data. Catanese et al. [2] acquired information only from profiles in friendship relation with the seed and from publicly accessible profiles. However, they simply presented a visualization of extracted Facebook data and developed a fast algorithm for data cleaning, running in $O(n \log n)$, exploiting the hash property of the Java HashSet, which removes all duplicate nodes, fixes all edges to link the unique instance of source and target nodes, and finally deletes the parallel ones. A similar approach is followed in this paper, where only friendship relations among real users are acquired, and fan pages and companies are discarded.

Authors in [11] proposed a dairy-like measure to calculate Facebook usage by users of different age groups and the motivation behind the same. However, it is a survey without actually traversing the network of a user on Facebook and hence lacks any proper modeling approach. In [7], the authors constructed a model to describe Facebook adoption by different users, with emphasis on its contribution in educational context. A total number of 11 observed and 3 latent variables provided by the model were used for analysis. This is again a survey-based analysis which models the system based on functionality. The authors in [6] analyzed network behavior to show how it is influenced by 5 user-related features, including gender, ethnicity, demographic traits, and cultural preferences by creating an incremental dataset collected over a time span of 3 years starting from the year 2006. They used four network parameters namely size, density, heterogeneity, and betweenness centrality to infer their association with race/ethnicity, class, and gender.

The authors in [8] discussed the amount of personal information that is accessible to other users and the implications of its explicit exposure by users. A set of 400 Facebook profiles were used to conduct the experiment from the networks of varying sizes. Sixty one variables were calculated from these profiles which were classified into default/standard information, sensitive personal information, and potentially stigmatizing information as per their content. In [5], the authors have examined the interplay between positive and negative links in social media to see how it affects the structure of online social networks.

In [9], the authors analyzed 133 Facebook profiles to calculate the relation between the virtual and real personalities of users. Sebastian et al. [12] conducted a web-based survey of 2603 random Facebook profiles to study its relation with the attitudes and behaviors of users in real-life scenario. Gjoka et al. [4] tried to obtain an unbiased sample of the Facebook network. A random sample was calculated using Metropolis-Hasting random walk and a re-weighted random walk, and the efficiency of these graph-crawling/sampling techniques and algorithms was analyzed. It however lacked the functionality of user-centric or user-specific social graph representation. They represented the collected sample as an undirected graph with users as set of nodes and set of edges depicting mutual friendship relationships between them.

3. FEATURE IDENTIFICATION

In this paper, Facebook is used as a case study to identify features pertaining to its users. As of September 2012, Facebook has over one billion active users (The Wall Street Journal, October 4, 2012). On analysis, a set of 11 features that can define the strength of a tie between two users on Facebook has been identified. The classification and a brief description of the identified features are presented in the following sub-sections.

3.1 Interaction-Based Features

These features are based on the content of the data shared between two users. Commonly used data on Facebook includes comments, likes, posts, shares, and tags. On analysis, following four features are identified in this category: *rate of comments* (ρ^c), *rate of user likes* (ρ^l), *rate of posts* (ρ^p), and *rate of content exchanged* (ρ^w).

3.1.1 Rate of comments

Comments serve as the main source of active communication on Facebook. Any two users with strong friendship ties will comment on each other posts more frequently and hence will have a higher rate of comments with respect to each other. A user can generate comments on a time-line post, URL or any other visible activity on the time-line of other users. Active users with strong ties have a higher rate of comments with each other.

The rate of comments, ρ^c_{xy}, of a user x on user y can be defined using equation 1, where S^c_{xy} is the set of comments made by user x on y's time-line and S^c_x is the set of comments made by user x.

$$\rho^c_{xy} = \frac{|S^c_{xy}|}{|S^c_x|} \tag{1}$$

3.1.2 Rate of user likes

The Facebook *like* button is a feature that allows users to show their support for specific comments, pictures, time-line posts, and statuses. Added in February 2009, the *like* button allows users to show their appreciation for content without having to make a written comment[1]. *Like* is a way to give positive feedback or to connect with things a user cares about on Facebook. A user can like content that her friends post to give them feedback on Facebook. A story about a users like will appear on her time-line and may also appear in her news feed. A user can like the visible activity of another users time-line on Facebook. Active users with strong ties have a higher rate of *likes* with respect to each other as they are exhibiting a positive feedback (acknowledgement) to each other's activity.

The rate of user likes, ρ^l_{xy}, of a user x for user y can be defined using equation 2, where S^l_{xy} is the set of likes made by user x on y's time-line and S^l_x is the set of likes made by user x.

[1] http://whatis.techtarget.com/definition/Facebook-Like-button

$$\rho^l_{xy} = \frac{|S^l_{xy}|}{|S^l_x|} \tag{2}$$

3.1.3 Rate of posts

It represents the rate of posts between users. A user can generate posts on the time-line of another Facebook users. At a more abstract level, most of the activities of a user is presented as a post on her time-line. Therefore, when a user writes something on her own or on another users' time-line, shares a post of another users on her time-line, or is tagged in a post by some other users, it is presented as a post on her time-line. Therefore, in this paper, the rate of posts encompasses rate of shares and tags as well. On analysis, it is found that active users with strong ties have higher rate of posts with respect to each other, i.e., the users interact more frequently.

The rate of posts, ρ^p_{xy}, of a user x on user y can be defined using equation 3, where S^p_{xy} is the set of posts made by user x on y's time-line, and S^p_x is the set of posts made by user x.

$$\rho^p_{xy} = \frac{|S^p_{xy}|}{|S^p_x|} \tag{3}$$

3.1.4 Rate of content exchanged

Users exchange information over social networks. If there is a higher rate of information sharing between two users, it strengthens their relationship and thereby increases their tie strength. The content exchanged (word count), ρ^w_{xy}, of a user x for user y can be defined using equation 4, where S^w_{xy} is the set of words written by user x on $y's$ time-line and S^w_x is the set of words written by user x.

$$\rho^w_{xy} = \frac{|S^w_{xy}|}{|S^w_x|} \tag{4}$$

Considering the interaction set, $I = \{\rho^c, \rho^l, \rho^p, \rho^w\}$, for every node the interaction value, Γ_{xy}, of users x and y is defined using equation 5.

$$\mathit{\Gamma}_{xy} = \frac{\sum\limits_{f^i \in I} min(f^i_{xy}, f^i_{yx})}{|I|} \tag{5}$$

3.2 Interest-Based Features

Certain activities of a particular user depicts her interests in a social network. In case of Facebook, these include *common pages*, *URL sharing*, and *event participation*. Based on the activities of a user, following three features fall in this category: *rate of common events* (ρ^e), *common pages* (η^{pl}), and *common urls* (η^{url}).

3.2.1 Rate of common events

A Facebook event is a calendar-based resource, which can be used to notify users of upcoming occasions. Events can be created by anyone and can be open to anyone or to specific users. Facebook events are a great way to spread the words on upcoming events or occasions, as they are able to reach thousands of people in a short period of time. The event also provides an "RSVP" list, which displays lists of invitees grouped by their response. Invitees are either placed in "attending," "not attending," "may be attending," or "hasn't responded" categories. If an invitee RSVPs that they are "attending" the event, it appears in their news feed to notify their friends. When the date of the event approaches, it is displayed on the invitees' home pages to remind them[2]. When two users are attending the same event, it shows there common interest in some cause/occasion. Being in a same event increases the likelihood of communication and sharing of ideas pertaining to the event, and hence results into better social ties.

The common event participation rate, ρ^e_{xy}, for users x and y can be defined using equation 6, where S^e_x and S^e_y are the set of events attended by users x and y, respectively.

$$\rho^e_{xy} = \frac{|S^e_x \bigcap S^e_y|}{|S^e_x \bigcup S^e_y|} \tag{6}$$

3.2.2 Rate of common pages

Pages are for businesses, organizations and celebrities to share their stories and connect with people. Like time-line, pages can be customized by adding apps, posting stories, and hosting events. Pages engage and grow its audience by posting regularly. People who like a page get updates in their news feeds. It represents the maximum number of common pages that any two users like. Common page like shows common interest and hence depicts a strong tie between two users.

The common pages, η^{pl}_{xy}, of users x and y can be defined using equation 7, where S^{pl}_x and S^{pl}_y are the set of page likes by users x and y, respectively.

$$\eta^{pl}_{xy} = \frac{|S^{pl}_x \bigcap S^{pl}_y|}{|S^{pl}_x \bigcup S^{pl}_y|} \tag{7}$$

3.2.3 Rate of common URLs

A user on Facebook can post any valid URL or link on her time-line. When two different users post same link on their respective time-lines, they share common interest and hence have a strong tie.

The common URL posts, η^{url}_{xy}, of users x and y can be defined using equation 8, where S^{url}_x and S^{url}_x are the set of URL posts by users x and y, respectively.

$$\eta^{url}_{xy} = \frac{|S^{url}_x \bigcap S^{url}_y|}{|S^{url}_x \bigcup S^{url}_y|} \tag{8}$$

3.3 Spatial Features

Every user has some associated personality traits based on her spatial location. Some basic properties provide an insight into the personality of a user on an OSN like place of origin, religion, political views, age, languages spoken, etc. On analysis, following four features are found in this category: *geographical similarity* (η^g), *belief similarity* (η^b), *age similarity* (δ^a), and *language similarity* (η^{lg}).

3.3.1 Geographical similarity

Users originating from same geographical location tend to have same ideology and interests. The higher value of geographical similarity indicates a possibility of stronger ties

[2] http://whatis.techtarget.com/definition/Facebook-event

between two users. It represents the geographical closeness of any two users based on the place of origin. In this paper, only state and country are considered for calculating this parameter. Based on the information available on the profile of a particular user, the geographical location is divided into two groups namely ‘Current’ and ‘Home’ location. The current location depicts the present location of a user, which can be same as the home location of a user or different in case the user has displaced temporally or permanently for whatever reasons. Considering the state and country of a user, 16 possibilities arise for current state (C_s), current country (C_c), home state (H_s), and home country (H_c), which are valued between 0 & 1 for users x and y and shown in table 1

Table 1: Geographical similarity values

No.	Case	Value
1	$C_s^x = C_s^y$ & $C_c^x = C_c^y$	1.000
2	$C_s^x = H_s^y$ & $C_c^x = H_c^y$	0.750
3	$H_s^x = C_s^y$ & $H_c^x = C_c^y$	0.625
4	$H_s^x = H_s^y$ & $H_c^x = H_c^y$	0.500
5	$C_s^x \neq C_s^y$ & $C_c^x = C_c^y$	0.375
6	$C_s^x \neq H_s^y$ & $C_c^x = H_c^y$	0.250
7	$H_s^x \neq C_s^y$ & $H_c^x = C_c^y$	0.250
8	$H_s^x \neq H_s^y$ & $H_c^x = H_c^y$	0.125
9	Otherwise	0.000

The geographical similarity, η_{xy}^g, between users x and y can be defined using table 1, where C_s^x , C_s^y, and H_s^x, H_s^y represents the current and home states of x and y, respectively; C_c^x , C_c^y and H_c^x , H_c^y represents the current and home countries of x and y, respectively. As the model presented in this paper is based on users’ interactions, current geographical location has a major impact rather than home or passive geographical location. Apparently higher weighage is assigned to the cases where current locations, i.e., both state and country of first user x are involved (C_s^x and C_c^x; entries 1 and 2 in table 1); it is followed by current locations of user y (C_s^y and C_c^y; entry 3 in table 1); and finally the home locations of both users (entry 4 in table 1). The rest of the cases involve one inequality, i.e., only countries of the users match (C_c and H_c; entries 5 to 8 in table 1). For rest of the scenario, a value of 0.0 is considered as a geographical similarity value.

3.3.2 Belief similarity

Like geographical similarity, users with same religious and political belief tend to share common ideas and views, thereby increasing the friendship tie between the users. It represents the religious and political similarity between two users and has a value between 0 and 1 depending upon the respective religions and political beliefs of users under consideration.

The belief similarity, η_{xy}^b, between users x and y can be defined using equation 9, where R_x and R_y are the religions of users x and y, respectively, and P_x and P_y are the political views of users x and y, respectively.

$$\eta_{xy}^b = \begin{cases} 0.0 & \text{if } (R_x \neq R_y \wedge P_x \neq P_y), \\ 0.5 & \text{if } (R_x \neq R_y \wedge P_x = P_y) \vee (R_x = R_y \wedge P_x \neq P_y), \\ 1.0 & \text{otherwise} \end{cases} \tag{9}$$

3.3.3 Age similarity

Users belonging to same age group have a natural tendency to interact more and exhibit stronger relationships. The lesser age difference the users have, the more stronger their ties tend to be. In this paper, only the year of birth is considered for calculating the age difference. The age difference, δ_{xy}^a, of users x and y is defined using equation 10, where A_x and A_y are the ages of users x and y in years, respectively.

$$\delta_{xy}^a = 1 - \frac{|A_x - A_y|}{A_x + A_y} \tag{10}$$

3.3.4 Language similarity

Language has an important effect on the interactions between two users. On analysis, it is found that knowing similar languages results in better understanding of the posted content. The language similarity, η_{xy}^{lg}, for users x and y is defined using equation 11, where S_x^{lg} and S_y^{lg} are the set of languages of users x and y, respectively.

$$\eta_{xy}^{lg} = \frac{|S_x^{lg} \bigcap S_y^{lg}|}{|S_x^{lg} \bigcup S_y^{lg}|} \tag{11}$$

4. PROPOSED USER-CENTRIC SOCIAL NETWORK MODELING APPROACH

This section presents the proposed user-centric social network modeling approach to identify and predict strength of social ties between OSN users. The experiment starts with a single seed user and the 11 features mentioned above are calculated with respect to all of its visible interactions with other users in the same network. Same procedure is repeated up to two levels of interactions starting from the seed user, i.e., the analysis extends up to the users who interact with users having interaction with the seed user. Multi-Dimensional Scaling (MDS) is used to transform the given user-feature collection into a 3D social space. Since the feature values are calculated with respect to a pair of users x and y interacting with each other in a network (Facebook), all features cannot be used as dimensions for latent space model. On analysis, along with the limitations of publicly available data, three generic features are identified that exist between any pair of users irrespective of their interactions. These features include geographical, age, and language similarities represented using η^g, η^a, and η^{lg}, respectively. These are used as three dimensions in our model. Thus, the dimension set D for LSM is defined as $D = \{\eta^g, \eta^a, \eta^{lg}\}$ and the feature set F as $F = \{\rho^c, \rho^l, \rho^p, \rho^e, \rho^w, \eta^{pl}, \eta^{url}, \eta^b\}$.

Since a single user is used as a seed in this model, the features in D are calculated between the seed user and all other users, i.e., for each feature $f_{xy} \in D$, the second user y is always the seed user. This results into a friendship graph of n entities in k-dimensional space. Considering M as an $n \times k$ matrix, in which a row r_i represents the latent

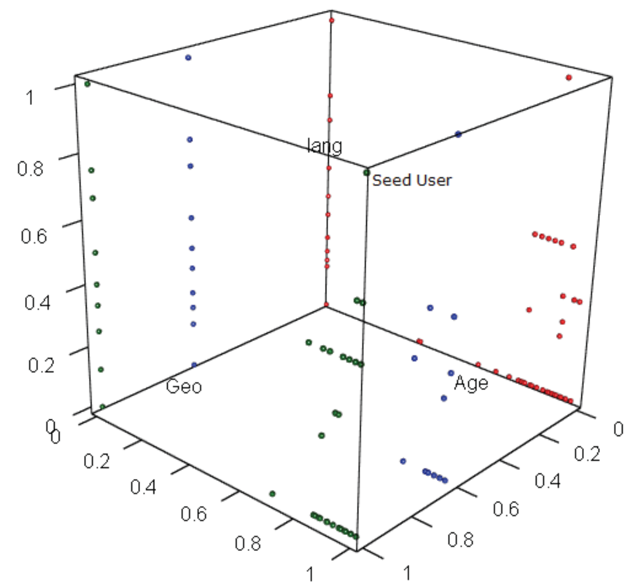


Figure 1: Three-dimensional (3D) social space simulation

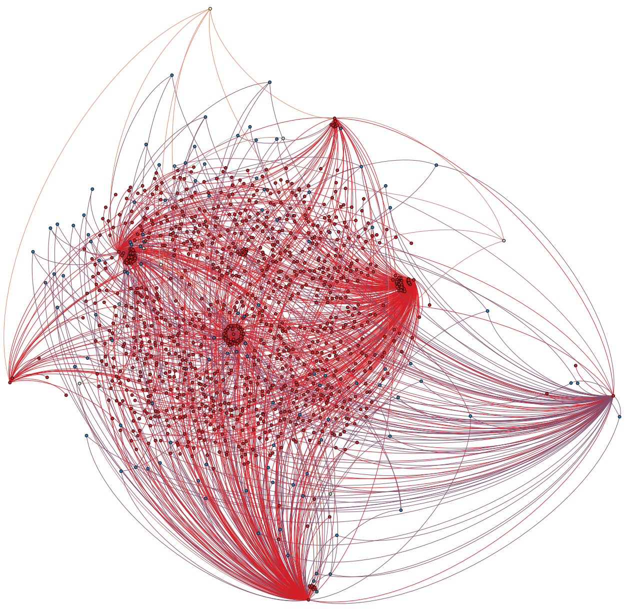


Figure 2: Real interaction network

position of user i in the latent space, the Euclidean distance, d_{ij}, between two nodes i and j is defined using equation 12.

$$d_{ij} = \sqrt[k]{\sum_{p=1}^{k} (x_{ip} - x_{jp})^2} \qquad (12)$$

The friendship graph $G = (V, E, \Omega)$ generated in this 3D space is a weighted undirected graph, with users as nodes and their social ties as edges. The weight of an edge represents the link probability between two nodes. Since there exists a feature set F for every node, the similarity value γ_{xy} for users x and y is defined using equation 13.

$$\gamma_{xy} = \frac{\sum_{f^i \in F} min(f^i_{xy}, f^i_{yx})}{|F|} \qquad (13)$$

Sarkar and Moore [10] defined the link probability p^L_{ij} between two entities or nodes i and j as given in equation 14, where $K()$ is the kernel function, such that two entities have high probability of linkage only if their latent coordinates are within radius r_{ij} of one another. Beyond this range there is a constant noise probability ρ of linkage. The link probability is inversely proportional to α; two users having higher link probability will have high similarity value γ. Moreover if a node is having high degree (i.e., large sphere of influence), it is active on a OSN and is more likely to form links as compared to a dead or dormant node with less sociability. Based on these observations, α is replaced by $(\gamma + r_{ij})$ in equation 14.

$$p_{ij} = \frac{1}{1 + e^{(d_{ij} - \alpha)}} K(d_{ij}) + \rho(1 - K(d_{ij})) \qquad (14)$$

Since the model is build on the interactions between the users, the interaction value Γ from equation 5 is also introduced into equation 14 to get equation 15, where the kernal function $K(d_{ij})$ is defined using equation 16.

$$p_{ij} = \frac{1}{1 + e^{\{d_{ij} - (\gamma + r_{ij})\}}} K(d_{ij}) + \Gamma_{ij} \qquad (15)$$

$$K(d_{ij}) = \begin{cases} (1 - (d_{ij}/r_{ij})^2)^2 & \text{if}(d_{ij} \leq r_{ij}), \\ 0 & \text{otherwise} \end{cases} \qquad (16)$$

Here, each entity i has a radius r_i, which is used as a sphere of interaction within latent space. An entity with higher degree will have a larger radius. The authors in [10] defined the radius of entity i with degree δ_i as $C(\delta_i + 1)$, r_{ij} as $C \times (max(\delta_i, \delta_j) + 1)$, and C is estimated by a simple linear search on the score function. The constant 1 ensures a non-zero radius. This modifies equation 15 as equation 17, which is used to calculate link probability between a pair of users in the dataset.

$$p_{ij} = \begin{cases} p^L_{ij} K(d_{ij}) + \Gamma_{ij} & \text{if}(d_{ij} \leq r_{ij}), \\ \Gamma_{ij} & \text{otherwise} \end{cases} \qquad (17)$$

The close vicinity of two users in LSM alone does not ensure their link. Instead, if the users are in close proximity of each other but have a small radius of influence, they will have a less link probability as compared to users that have high radius of influence determined by r_{ij} in equation 17.

5. EXPERIMENTAL SETUP AND RESULTS

In this section, the data collection and processing methods are discussed, followed by the discussion on experimental results. While extracting publicly available data from Facebook, the user privacy concerns and policies were followed and no personal information or their social behavior was either stored or analyzed. At no place in the entire experiment, the actual user IDs or details are used. The

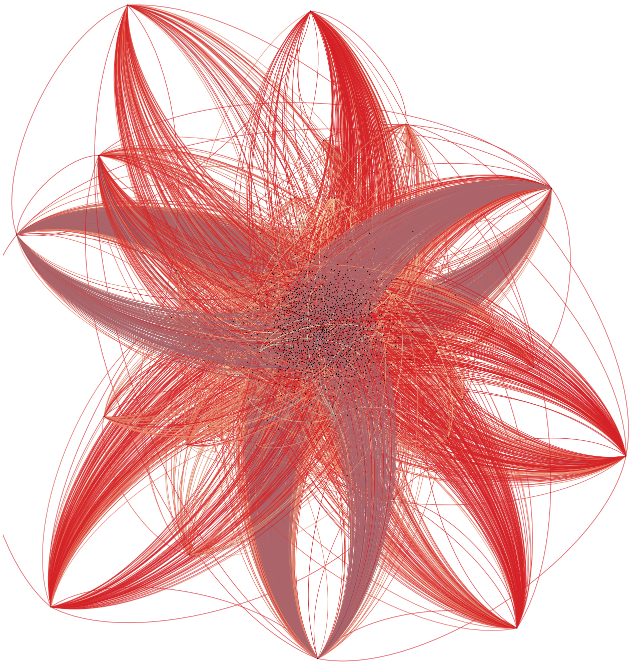


Figure 3: Interaction network for probability interval $0.00 < p \leq 0.25$

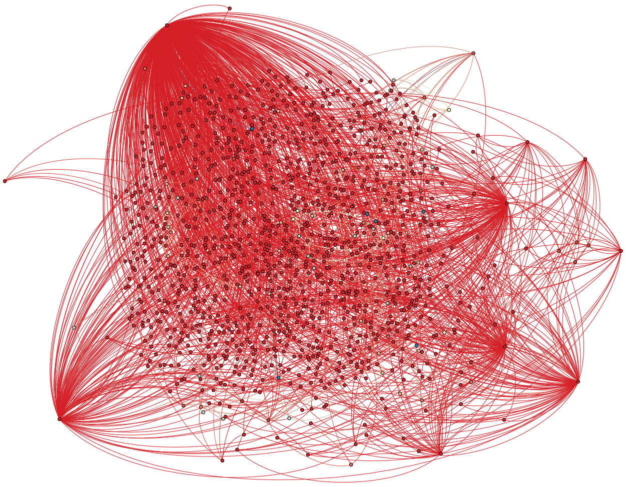


Figure 4: Interaction network for probability interval $0.25 < p \leq 0.50$

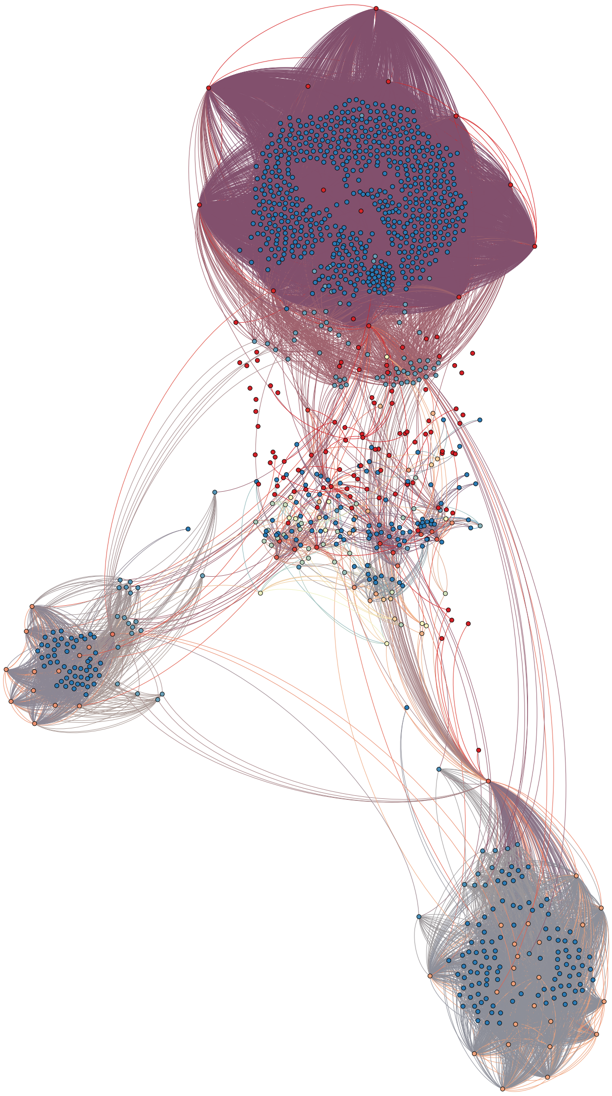
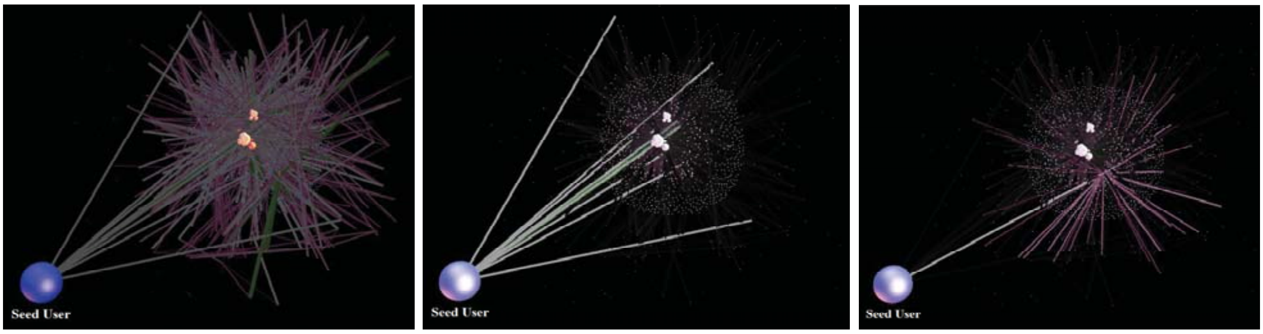


Figure 5: Interaction network for probability interval P^* ($p > 0.51$)

names of users along with their IDs were encrypted using a hash function. The only purpose behind the experiment is to construct an interaction-based network and estimate link probabilities between the users in the network.

For initial processing, a total of 1210 Facebook profiles including the seed user were processed. Out of these, 33 users interacted directly with the seed user, and the rest 1176 interacted directly with these 33 users. From 1210 Facebook profiles, 11 features discussed earlier were calculated between every pair of users, i.e., a total of 1210×1210 values were calculated and stored. To plot this user-centric graph in 3D space, the dimension set D was calculated for each profile with respect to the seed user. This resulted in a 1210×3 dimension matrix. Using this dimension matrix, a 3D graph was plotted using `rgl package` in 'R'[3]. The resultant graph is shown in figure 1.

[3] http://www.r-project.org/



(a) A multi-level view of the P^* network

(b) Visualization of level-1 links projected from the seed user

(c) Visualization of level-2 links projected from a level-1 user

Figure 6: Visualizing network with link probability P^* using YifanHu's multi-level layout of Gephi

The seed user lies at the extreme end with all dimensions equal to 1. All other users are distributed between the origin $(0, 0, 0)$ and seed user $(1, 1, 1)$. The link probability increases as the user moves closer to each other in this 3D space. Since these dimensions were calculated w.r.t. the seed user, the link probability is defined in context of the seed user, i.e., link probability p_{xy} between two users x and y is the likelihood of user x to interact with user y based on how closely they behave in reference to the seed user. This modeling technique calculates the link probabilities between different users and identifies users of interest.

Using equation 17, the link probabilities between 1210 users are calculated, and based on these values, the interaction network shown in figure 2 is divided into 4 groups. The first group has 5382 edges with $0 < p_{xy} \leq 0.25$, the second group has 1745 edges with $0.25 < p_{xy} \leq 0.50$, the third group has 274283 edges with $0.50 < p_{xy} \leq 0.75$, and the last group has only 160 edges with $0.75 < p_{xy} \leq 1.0$ (table 2). A visualization of the first two groups is shown using `Gephi`[4] in figures 3 and 4. For these and other such visualizations, Force Atlas-2 layout is applied with an edge weight influence of 5.0.

As the network is generated based on the interactions, most of the edges (37.49% of Total Possible Links (TPL)) have a probability in the interval $]0.5, 0.75]$. This interval also has a very high average clustering coefficient value of 0.94 depicting a maximum number of triadic closures in the generated network and hence serves of minimum use.

If however, the case of even odds ($50-50$ chance) is eliminated by considering a link probability value P^* greater than 0.51, a total of 14093 edges (1.93% of the TPL) are formed which is greater than the percentage of actual interaction links (0.15% of the TPL) (figure 5). The average clustering coefficient for P^* is lesser than the Overall Probable Links (OPL), depicting less number of triadic closures in the generated network. With the network diameter greater than that of OPL, the probability interval P^* has a higher reach in the network. Also the average path length less than 3 means that the network is mostly limited around the second level of interactions (table 3). Figure 6 is a visualization of a network with link probability P^* using YifanHu's multi-level layout of Gephi.

Table 2: Link probability distribution

Link probability	Edges	% of TPL	ACC	APL	ND	AD
$0.00 < p \leq 0.25$	5382	0.73	0.409	2.301	5	8.896
$0.25 < p \leq 0.50$	1745	0.23	0.026	3.551	12	2.884
$0.50 < p \leq 0.75$	274283	37.49	0.942	2.367	9	226.68
$0.75 < p \leq 1.00$	160	0.02	0.001	2.561	5	0.264

ACC: Average Clustering Coefficient TPL: Total Possible Links ($\frac{n \times (n-1)}{2}$)
APL: Average Path Length ND: Network Diameter
AD: Average Degree

Table 3: Interaction network statistics

Network	Edges	% of TPL	ACC	APL	ND	AD
Interaction	1099	0.15	0.096	3.231	5	1.817
OPL ($P_l > 0$)	281570	38.49	0.947	1.687	4	465
P^* ($P_l > 0.51$)	14093	1.93	0.851	2.8	9	23.29

OPL: Overall Probable Links

6. CONCLUSION AND FUTURE WORK

In this paper, a user-centric social graph modeling approach has been proposed to infer ties between social network users. For a given seed user, her interaction data is analyzed to identify directly interacting (level-1) users, and the same process is repeated to identify the users that have interactions with the level-1 users forming a second-level user-set (level-2). Identified users are plotted as nodes in a 3D space considering the seed user as a reference point, and latent space model is used to calculate link probability between a pair of users in the multi-dimensional space. Finally, a link probability is calculated to predict social tie between two users. Based on the degree of a node (representing user), a sphere of influence is calculated demarcating the maximum influence area of the user. Unlike other modeling techniques that represent a social network in its entirety, the modeling technique followed in this paper is designed around a given seed user. Thus, the behavior of the seed user formulates and shapes the entire network. The user ties and interaction links formed are driven by the seed user and hence help in focusing on a particular trend in the network. If a seed user is criminal minded, the link probabilities will show the chances of two users to interact in a negative environment, and if the seed user behaves with a constructive mind-set,

[4] https://gephi.org/

the user-user links identified would be constructive in nature. Most of the existing social network modeling techniques generalize the users under consideration, whereas the proposed modeling technique is a specialized view of the user network, which results into the formation of a socio-centric model of the extracted user-set from the interactions around a seed user.

In our experiment, the users' profiles have been crawled using a customized crawler, which has access to a limited public data. In future, an enhanced crawler having proper support from Facebook, can be developed to retrieve a large and complete dataset to calculate link probabilities with more accuracy. Using private feature-set (currently unavailable) as dimensions in the latent space model can result into better insights about the users' interactions. Moreover, the modeling technique presented in this paper can be further enhanced and remains a topic for future research.

7. ACKNOWLEDGMENT

The authors acknowledge the support provided by the King Abdulaziz City for Science and Technology (KACST) and King Saud University (KSU), Kingdom of Saudi Arabia. This work has been funded by the KACST under the NPST project number `11-INF1594-02`.

8. REFERENCES

[1] M. W. Boyer, J.M. On the cutting edge: Simplifed o(n) planarity by edge addition. *Journal of Graph Algorithms and Applications*, 8(3):241–273, 2004.

[2] S. Catanese, P. De Meo, E. Ferrara, and G. Fiumara. Analyzing the facebook friendship graph. In *Proceedings of the 1st International Workshop on Mining the Future Internet*, pages 14–19, September 2010.

[3] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Crawling facebook for social network analysis purposes. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, pages 52:1–52:8, 25-27 May 2011.

[4] M. Gjoka, M. Kurant, T. Carter Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *Proceedings of the IEEE INFOCOM*, pages 1–9, 14-19 March, 2010.

[5] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370, April 10 âĂŞ 15, 2010.

[6] K. Lewisa, J. Kaufmana, M. Gonzaleza, A. Wimmerb, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks 30*, pages 330–342, 2008.

[7] S. G. Mazman and Y. K. Usluel. Modeling educational usage of facebook. *Computers and Education 55*, pages 444–453, 2010.

[8] A. Nosko, E. Wood, and S. Molema. All about me: Disclosure in online social networking profiles: The case of facebook. *Computers in Human Behavior 26*, pages 406–418, 2010.

[9] D. Samuel Gosling, S. Gaddis, and S. Vazire. Personality impressions based on facebook profiles. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*, 2007.

[10] P. Sarkar and A. W. Moore. Dynamic social network analysis using latent space models. *SIGKDD Explorations*, 7(2):31–40, 2005.

[11] A. Tiffany Pempek, A. Yevdokiya Yermolayeva, and L. Sandra Calvert. College students' social networking experiences on facebook? *Journal of Applied Developmental Psychology 30*, pages 227–238, 2009.

[12] S. Valenzuela, N. Park, and F. Kerk Kee. Is there social capital in a social network site?: Facebook use and college studentsâĂŹ life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication 14, International Communication Association*, pages 875–901, 2009.