Analysis of Large Multi-modal Social Networks: Patterns and a Generator

Nan Du*, Hao Wang*, and Christos Faloutsos†

*Nokia Research Center, Beijing
{daniel.du,hao.ui.wang}@nokia.com
† Carnegie Mellon University, Pittsburgh
christos@cs.cmu.edu

Abstract. On-line social networking sites often involve multiple relations simultaneously. While people can build an explicit social network by adding each other as friends, they can also form several implicit social networks through their daily interactions like commenting on people's posts, or tagging people's photos. So given a real social networking system which changes over time, what can we say about people's social behaviors ? Do their daily interactions follow any pattern ? The majority of earlier work mainly mimics the patterns and properties of a single type of network. Here, we model the formation and co-evolution of multi-modal networks emerging from different social relations such as "*who-adds-whom-as-friend*" and "*who-comments-on-whose-post*" simultaneously. The contributions are the following : (a) we propose a new approach called *EigenNetwork Analysis* for analyzing time-evolving networks, and use it to discover temporal patterns with people's social interactions; (b) we report inherent correlation between *friendship* and *co-occurrence* in on-line settings; (c) we design the first multi-modal graph generator *xSocial*[1] that is capable of producing multiple weighted time-evolving networks, which match most of the observed patterns so far. Our study was performed on two real datasets (Nokia FriendView and Flickr) with 100,000 and 50,000,000 records respectively, each of which corresponds to a different social service, and spans up to two years of activity.

Keywords: Social Network Analysis, Graph Generator, Multi-modal Networks

1 Introduction

Research of real world complex networks, like social networks [26], biological networks[11], topology [15] of WWW and Internet raises many significant and important problems. What patterns do the human-to-human interactions follow in large-scale social networks ? How can we use such patterns to facilitate existing applications, such as anomaly detection [1] [18] and collective classification [8], and make further innovations?

[1] http://research.nokia.com/people/hao_ui_wang/index.html

As a result of the widespread adoption of Web 2.0 technology, social networking sites or services(SNS) are becoming ubiquitous and penetrate into every corner of people's daily lives. In such systems, people often belong to multiple social networks because of different person-to-person interactions. For example, in Nokia FriendView, Flickr(`www.flickr.com`), Facebook(`www.facebook.com`), LinkedIn(`www.linkedin.com`), Twitter(`www.twitter.com`), and eBay(`www.ebay.com`), they all provide the basic function that enables people to add each other as friends through their content and conversations, which contributes to the emergence of our first type of social network, namely, the "*friend network*" or the "*buddy network*".

In addition, they also allow people to participate in specific activities. In FriendView, we can comment on the posts written by our colleagues. In Flickr, we can tag the photos uploaded by our friends. In eBay, we can rate the products sold by our partners. As a consequence, interactions with people centered around content form another type of social network called the "*comment network*" or the "*participation network*" from such activities as "commenting-on-posts", "tagging-photos" and "rating-products". Therefore, these two types of social networks describe different facets of the same social networking system. For each of them, recent research has reported fascinating patterns, like [28] or lognormal [7] or Double Pareto LogNormal (DPLN) distribution [24] [27] for the degree, as well as small and shrinking diameter [20].

In this paper, we are interested in answering the following questions:

- Do human social interactions and behaviors follow any temporal pattern ? Is there any regularity inherent in the daily activities of individuals and groups? Can we use such patterns to make predictions of their future behaviors ?
- Given a real social networking site, is there any correlation between the *buddy network* and the *participation network* ? For instance, can we infer the friendship between two people in *buddy network* according to the discrete observations of their co-occurrence in the *participation network* ?
- How can we produce an intuitive generator that will mimic the behaviors, and correlations of these networks within a real social networking site simultaneously ? Most existing generators try to mimic the skewed distribution of degree or weight of only a single network, and thus fail to incorporate the possible correlations with other networks. Here, we want a multi-modal graph generator, which should describe the way in which the different social networks discussed above could co-evolve over time through the local interactions and activities between individuals.

Answering these questions can have many practical applications. First, identifying meaningful patterns hidden in human activities contributes to classifying people into different groups according to the similarity of their social behaviors, based on which we can have a deep insight about the composition and evolution of the network they belong to. Discovering new patterns also helps to discard unrealistic graph models. Second, knowing the correlation between different social relations is good for us to design better systems that further expand the range of human interactions by offering particular friend or product recommendations according to specific user context. Finally, intuitive graph models are also vital for simulation studies of routing algorithms when it is hard or even impossible to collect large real data, for understanding how the macro and global patterns of networks can emerge through the micro and local interactions among people over time, and for compressing and summarizing the real networks by model parameters.

The paper is then organized as follows. Section 2 reviews related work. Section 3 presents our observed patterns. Section 4 describes the *xSocial* model in detail. Section 5 gives the conclusion.

2 Related Work

In this section, we mainly survey the various discovered properties of real world networks, and several well-known graph generators.

2.1 Network Patterns

Many interesting patterns that real graphs follow have been discovered in recent work like the power-law distribution of the number of messages(photos), power-law comment distribution, power-law interval distribution[16], power-law degree distribution[28], power-law edge-weight distribution[26], power-law node-weight distribution[26], snapshot power-law[23], clique-participation law[13], clique-degree power-law[13], triangle-weight law[13], eigenvalue power-law[2], shrinking diameter[20], and oscillating connected component[23]. These patterns are important for us to understand the static and temporal properties of real world networks, to identify authorities and subgroups, as well as to refine routing algorithms and recommendations. Moreover, they are also vital for eliminating unrealistic graph generators and guiding us to design better ones, because ideally a graph model should be able to mimic all these patterns as many as possible.

2.2 Graph Generators

Generally, the graph generators of recent literature can be mainly classified as *emergent* graph models, and *generative* graph models. The basic principle of *emergent* graph models is that the macro network properties should *emerge* from the micro interactions of nodes over time. This type of models include Erdös-R*é*nyi(ER) model [14], small-world model [29], BA model [6], Copy model [9], Random Multiplication Model [9], Forest Fire model [20], 'butterfly' model [23], and 'RTG' model [2]. [See [5] and [9] for a detailed review and discussion]. Recently, Goetz [16] also provides models to mimic the evolving and spreading mechanism of blog systems [21]. Moreover, research from the fields of economics and game theory also brought utility-based models [17][4][12][13] where each node tries to optimize a specific utility function, and the network structure can arise from the collective strategic activities of all the nodes. *Generative* graph models often assume a global mathematic rule and perform iterations of such rule recursively until the generated networks meet several properties of real networks. Such models include kronecker multiplication model [19] and tensor model [3].

In summary, the majority of earlier graph generators often focused on modeling some main properties of only one single network. For example, [29][6][20][23] are limited in trying to model unweighted networks, and cannot be generalized to weighted networks. Goetz[16] describes the evolving process of blogs, but fail to incorporate the weights. Although RTG[2] can generate weighted graphs, it still only focused on one single network. As to the generative models, they usually cannot mimic the micro mechanism of node and edge addition, which makes it hard for us to understand the inherent natural process of real networks. In contrast, our work not only considers to mimic most of the known patterns, such as generating weighted networks from local nodes' interactions, but also focuses on co-evolution of different networks simultaneously.

3 Tools and Observations

In this section, we seek to find patterns inherent in large-scale on-line social networking sites. We first give a preliminary description of Nokia FriendView and Flickr datasets, and then we present the proposed *EigenNetwork* analysis method, and the discovered *CoParticipation Friendship Correlation* pattern.

3.1 Data Description

The datasets that we have analyzed include the interaction records from Nokia FriendView, and Flickr. Nokia FriendView is a location-enhanced experimental microblogging application and service operated by Nokia Beta Labs from the beginning of November 2008 to the end of September 2009 when the service was finished. It allows users to post messages about their status and activities from GPS-enabled Nokia S60 phones or from the web. Any two users can add each other to their buddy list through email request and confirmation. The users can also comment on the status messages posted by the buddies in their social network. As a result, we use three different types of record, $< usrID, msgID, postTime, length >$, $< usrID, buddyID, addTime >$, $< userID, msgID, commentTime, length >$, to describe these actions respectively.

Here, the edge weight of *buddy network* is the total number of comment times between them. For the dataset, there are 34,980 users, 20,873 buddy links, 62,736 status messages, and 22,251 comments [10]. The unique feature of this dataset is that it has recorded a complete evolving process of a social networking site from the very beginning to the end, over the course of 11 months. The detailed records enable us to have a deep insight about the way that people interact with each other. In the Flickr dataset (where people can upload photos, add contacts, and comment on or tag photos), we use similar tuples as Friend View to describe the data which includes about 542,105 users, 46,668,661 contact links, 101,520,484 photos, and 8,999,983 comments from 2005 to 2007. Because these datasets belong to different services, have different scales, and were collected from different time, the diversity of our data can thus be guaranteed. Notice we only use the encrypted user id in this study, and restrict our interest only in the statistical findings within the data.

3.2 EigenNetwork Analysis

While the activities and interactions where each of us is involved every day appear nearly random, intuition tells that there also seems to be some regular recurrence of patterns, especially when we take the temporal, spatial, and social context into consideration. For instance, we may check several emails, and see some news after arriving at the office in the morning. Then we might chat with our friends through instant messaging during the working hours, and in the evening, we might write blogs, make comments, upload photos, or even play on-line games. Since a social network is inherently the collection of people and their interactions, analyzing the temporal behaviors of individuals and subgroups can help us to have a deep insight about the overall composition of the entire network.

We formulate our approach as follows. Given graph $\mathcal{G}$, for $\forall e_{ij} \in \mathcal{E}(\mathcal{G})$, we characterize the temporal activity of all the edges by a two-dimensional $E \times D$ binary matrix $\mathcal{M}$, where $E = |\mathcal{E}(\mathcal{G})|$, and D is the total number of days that graph $\mathcal{G}$ has been in study.

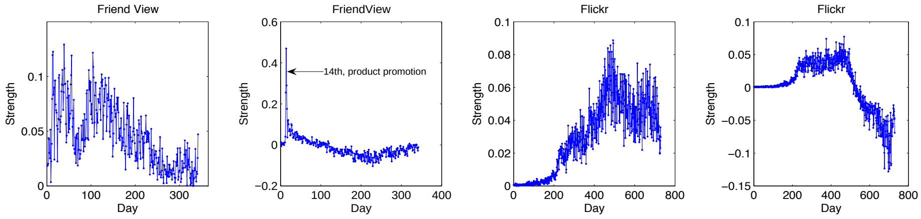
$$\mathcal{M}(p,q) = \begin{bmatrix} 0 & 0 & 1 & 0 & ... \\ 0 & 1 & 0 & 0 & ... \\ 1 & 0 & 1 & 1 & ... \\ ... & ... & ... & ... & ... \end{bmatrix} \qquad (1)$$

Therefore, the pth row represents the behavior of a particular edge e_{ij} spanning the D days. On a specific day q, if node v_i and v_j has at least one interaction with each other, then $\mathcal{M}(p,q) = 1$; otherwise $\mathcal{M}(p,q) = 0$. We then do Singular Value Decomposition(SVD) on matrix $\mathcal{M}$ and it is factorized as

$$\mathcal{M} = U \times \Sigma \times V^T \qquad (2)$$

where the columns of D-by-K matrix V form a set of orthonormal *input* basis vectors for $\mathcal{M}$, the columns of E-by-K matrix U form a set of corresponding orthonormal *output* basis vectors, and the diagonal values in K-by-K matrix Σ are the singular values arranged in the descending order by which each corresponding *input* is multiplied to give a corresponding *output*.

By intuition, the SVD on matrix $\mathcal{M}$ implicitly decomposes the E edges into K groups. Each column (or singular vector) i of the E-by-K matrix U describes the extent to which each edge of $\mathcal{G}$ participates in the ith group. Every column j of the D-by-K matrix V shows the extent to which the jth group is active on each day. The nonnegative real numbers on the diagonal of the K-by-K matrix Σ indicates the strength of each group. For each singular value s_i, the *energy* of s_i is defined as s_i^2, so we keep the first few strongest singular values whose sum covers 80-90 percentile of the total energy. Here, we build matrix $\mathcal{M}$ for the *participation network* which emerges from the comment interactions among users in FriendView and Flickr respectively. $\mathcal{M}(p,q) = 1$ means that for the pth edge e_{ij}, at least one of the two nodes (v_i and v_j) commented on the messages or photos posted by the other one on the qth day.

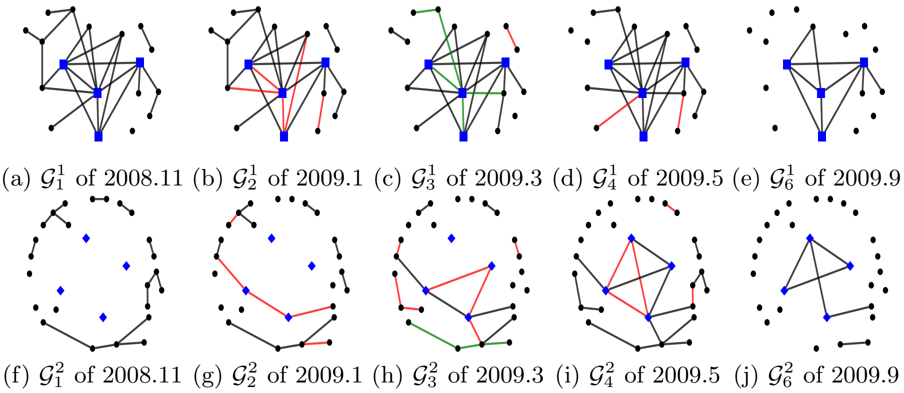


(a) 1st vector (b) 2nd vector (c) 1st vector (d) 2nd vector

Fig. 1. The 1st and 2nd singular vector of matrix V that describe the corresponding daily activities of the 1st and 2nd subgraph consisting of the selected edges in the *participation network* (formed by the *comment* relation) of FriendView (a-b), and Flickr (c-d) respectively.

Figure 1 shows the top two singular vectors of the matrix V from FriendView and Flickr. In Figure 1(a-b), we have two groups of edges that show different patterns of behavior. The first group of Figure 1(a) has basically a periodic pattern, while the second group of Figure 1(b) appears more bursty, where the spike occurs on the 14th day. Based on the complete records of FriendView, it was discovered that the 14th day was just during the week that Nokia did lots of advertising work to promote FriendView by calling for more open beta testers. For Flickr, both of the two groups shown in Figure 1(c-d) behave periodically. There is a clear trend of overall growth in the amplitude with some oscillation. We guess this may be caused by the quickly increased popularity and fast development of Flickr as more and more users joined in the system after the year 2006.

Figure 2 further presents the evolving process of the subgraph G_x^1 and G_x^2 consisting of the selected edges that actively participate in the 1st singular vector of matrix U. Being active means that we only keep the set of edges whose sum of the energy (which is the square of the corresponding value) covers 80-90 percentile of the total energy. In Figure 2, the evolving pattern of G_x^1 and G_x^2 are clearly different. Subgraph G_x^1 contains a size-4 clique (complete graph) where each blue-square node has connections with each other. This clique remains stable in topology and in total number of activities over the whole period, except for G_2^1 where five edges shown in red had significantly increased number of activities, and for G_2^1 where the the number of their activities dropped back. For $\forall e_{ij} \in \mathcal{E}(\mathcal{G}_x^1), x > 1$, *red* color of e_{ij} indicates that its weight (which is the total number of times that node v_i and v_j interact with each other in the xth month) is significantly higher than its previous value in graph $\mathcal{G}_{x-1}^1$, and *green* color


(a) $\mathcal{G}_1^1$ of 2008.11 (b) $\mathcal{G}_2^1$ of 2009.1 (c) $\mathcal{G}_3^1$ of 2009.3 (d) $\mathcal{G}_4^1$ of 2009.5 (e) $\mathcal{G}_6^1$ of 2009.9

(f) $\mathcal{G}_1^2$ of 2008.11 (g) $\mathcal{G}_2^2$ of 2009.1 (h) $\mathcal{G}_3^2$ of 2009.3 (i) $\mathcal{G}_4^2$ of 2009.5 (j) $\mathcal{G}_6^2$ of 2009.9

Fig. 2. The evolving process of the subgraph G_x^1 and G_x^2 consisting of the selected edges belonging to the 1st(top row) and 2nd(bottom row) singular vector of matrix U in the *participation network* of FriendView. $\forall G_x^1$ (G_x^2) where $x > 1$, *red* indicates that the weight (representing the number of times that two users comment on each other's messages) is at least an order of magnitude higher than its previous value in G_{x-1}^1 (G_{x-1}^2), *green* means the reverse, and *black* shows the same level.

means the reverse. We made further investigations into the egocentric subgraph of around such 4 blue-square nodes in the entire network. Their average *degree*, and *node betweenness* [26] are 39 and 0.42 respectively. Because *degree*, and *node betweenness* are two popular measures to quantify a node's authority or centrality in a social network, the subgraph formed from these active edges in the 1st singular vector of matrix U actually represents the central part or the core of FriendView's *participation network*.

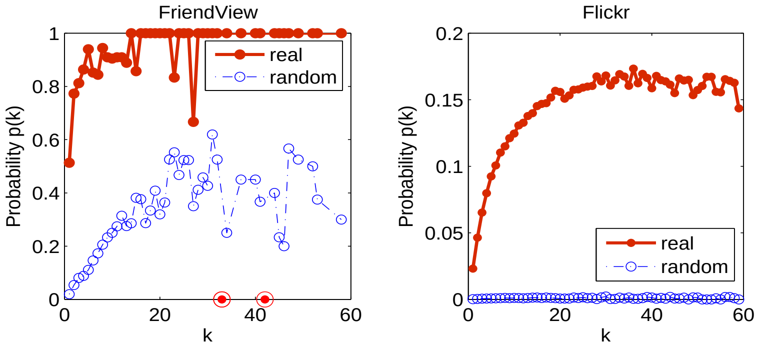
We see that in November, 2008 and January, 2009, there are two significant increases in the number of interactions as most edges in the subgraph are red compared with the previous graph, which also coincides with the two spikes in Figure 1 (a). Moreover, because the open beta testing for FriendView actually finished in September, 2009, in Figure 2, the subgraph becomes sparse, when the interactions between users dropped gradually, and also conforms with the decreasing trend in Figure 1(a). In contrast, the subgraph G_x^2 is loosely connected. In the beginning, it only consisted of several separated edges. Notice in Figure 1(b), there is a bursty in the first month when Nokia did a lot of publicity work. As a result, there were many separated short-term interactions at that time.

Therefore, because subgraphs formed by the selected edges from the singular vectors of matrix U (which are also the eigenvectors of $\mathcal{M} \times \mathcal{M}^T$) hold different local temporal patterns, and represent different compositions of the overall network, they are defined as the *EigenNetwork*s, and our methodology is thus called *EigenNetwork* analysis.

Observation 1 *EigenNetwork. The EigenNetworks can reveal local compositions of real world social networks, and hold different temporal patterns over time.*

3.3 CoParticipation-Friendship Correlation

In real social networking sites like FriendView or Flickr, on the one hand, people spend their daytime in following the updated status of their friends in the explicit *buddy network*. On the other hand, people are also the major players in the implicit *participation network* that emerges from the activities we adopt. As a consequence, is there any correlation between these two types of interaction ? Will the reoccurrence of one particular implicit activity contribute to a formation of the corresponding explicit interaction ? More specifically, can we quantify the extent to which two people will become friends in the *buddy network* according to the discrete observations of their co-occurrences in the corresponding *participation network* ? An underlying premise is that the probability for two people to become friends increases with the number of activities in which they have engaged together.



(a) Friend View (b) Flickr

Fig. 3. The probability $P(k)$ of being friends as a function of the number of co-commented messages k in Friend View (a), and photos in Flickr (b) respectively. For each k, *red* curve indicates the actual probability of being friends, and *blue* curve shows the expected value in random graphs. The outliers are marked by *red* circles.

Figure 3 shows this basic relationship in red color for FriendView and Flickr respectively, that is, the probability $P(k)$ of two people to become friends as a function of the total number of times that they have participated in k common activities. $P(k)$ is calculated as follows. We first find all tuples $< i, j, k >$ such that node v_i and v_j have k participated activities in common. Then $P(k)$ is the fraction of such tuples for a given k that node v_i and v_j are also friends in the *buddy network*. We see that when k is roughly small ($k < 30$), $P(k)$ has a strictly monotonic increase as k increases. However, as k becomes large, the marginal effect diminishes as k increases.

Moreover, we would also like to evaluate how this empirical correlation compares to the corresponding result if comments were produced randomly with the same background distribution in real datasets. Specifically, for each node v_i, while we keep the number of posts(photos) on which she would comment the same as that in real dataset, we let her randomly choose among the posts(photos) this time. We then use $P_0(k)$ to denote the expected probability for a pair of nodes to become friends. If $P(k) > P_0(k)$, we say that the correlation is over-represented in the data compared to chance;on the other hand, if $P(k) < P_0(k)$, then this correlation is underrepresented. To quantify the significance of $P(k)$ being over-or-underrepresented, we use the *surprise*[22] $S(k)$ which is defined as

$$S(k) = \Delta(k) \times (P(k) - P_0(k))/\sqrt{\Delta(k) \times P(k) \times (1 - P(k))} \tag{3}$$

where $\Delta(k)$ is the total number of tuples that have k common posts(photos). $S(k)$ indicates the number of standard deviations by which the actual number of pairs being friends deviates from the expected number in random graphs. In Figure 3, the plots in *blue* dash-line describe $P_0(k)$ vs. k in random graphs for FriendView and Flickr respectively. According to the Cental Limit Theorem, the distribution of each $S(k)$ conforms approximately to a standard normal distribution, and it is expected on the order of tens to already be significant ($S(k) = 6$ gives a *p*-value of 10^{-8} approximately)[22]. However, in our datasets, we have found that the average $S(k)$ for $k < 60$ and $P(k) \neq 0$ is 54.0 and 371.6 for FriendView and Flickr respectively, which is much larger and means that this correlation is statistically significant.

There are also some outliers(marked by red circles). The existence of such outliers means that although two people have engaged in many common activities together, they are still not friends yet. We guess this might be caused by users' ignorance or unawareness of each other. These types of users may only care about the messages or photos themselves by ignoring other people's comments at all. As a result, one possible application of the correlation shown in Figure 3 may be to help us with better recommendation systems, especially for the situation where k is large, because as k continuously increases , the number of pairs who have k activities together decreases significantly, which makes it easier to give specific recommendations.

Observation 2 CoParticipation-Friendship Correlation(CPF). *Given a real social networking site, the probability $P(k)$ of being friends for any pairwise persons increases with their k activities in common. Although the marginal effect diminishes as k increases, the effect remains significant.*

4 xSocial Model

Next, we present our *xSocial* model where the "*x*" means that it is a multi-modal graph generator that mimics real social networking sites to produce the *buddy* *network* and the *participation network* simultaneously. The guiding principle is that based on our understanding of existing patterns, we will devise a set of simple rules that each user would follow, and the entire social network will arise and evolve through the local interactions between individuals over time. Notice that this is actually a very challenging task, because the majority of prior work mostly focused on modeling only a single network. Our work is different as all the synthetic networks generated by *xSocial* should follow both the old and the new patterns mentioned in the previous section.

4.1 Model Description

xSocial model consists of the following four essential components:

- It is designed by using agent-based modeling approach. We have a set $\mathcal{A}$ of n distinct agents, each of which has a *preference* value f_i.
- At each time, every agent performs three independent actions(*write a message*, *add a friend* and *comment on a message*) guided by the one-dimensional random walk mechanism.
- An agent chooses his friends either by their popularity or by the number of messages on which they have commented together, which is determined by his *preference* f_i.
- An agent can also follow the updated status of his friends by putting comments on the corresponding newly written messages.

Preference Value. For any agent $a_i \in \mathcal{A}$, $f_i \in (0, 1)$ represents two different behaviors of people while they are using on-line social networking services. f_i approaching to 1 means the agent likes to follow those active agents who have already written many messages, and continuously put new messages, while f_i close to 0 indicates that he is interested in and pays more attention to the comments put by other agents.

Random Walk. In every step, each agent does a random walk on a line, and then chooses to write a message, or add a friend, or comment on a message whenever the walk returns to the origin (at state 0). We use three integers: S_w, S_a, and S_c to represent the state of the corresponding action respectively. The initial state of S_w, S_a, and S_c is 0. For each variable, there are two types of transition. An agent a_i adds or subtracts 1 from the variable's current state with probability p_i and $1 - p_i$ respectively. The agent a_i performs the corresponding action whenever S_w, S_a, or S_c returns to 0 again. Newman [25] shows that the inter-posting times follow a power-law distribution with exponent -1.5. Intuitively, the random walk reflects how frequently an agent uses the social networking service. On the one hand, when the probability p_i approaches to 0 or 1, the agents may just use the system in the very beginning, and never come back, just like that most users only register an account for curiosity in the beginning, but almost seldom use the service later. On the other hand, when p_i is near 0.5, the corresponding agents are relatively active users who can successively use the service, although they may be distracted by some random events to the nearby state around the origin.

Add a friend. For $\forall a_i \in \mathcal{A}$, once $a_i.S_a$ hits zero, he decides to expand his *buddy network* by exploring more friends. With probability f_i, a_i trusts word-of-mouth and chooses an agent a_j proportionally with the number of messages she has written, because the user who has published many messages will naturally attract attention of others so that she can expand the number of her followers. Once she has more followers, she would probably like to publish even more messages. In the opposite case, with probability $1 - f_i$, a_i picks an agent a_k proportionally with the number of messages on which they have put comments together.

Make a comment. For $\forall a_i \in \mathcal{A}$, when a_i decides to comment on some other messages at the moment $a_i.S_c = 0$, she prefers the candidates newly written by her friends, because for most social networking services, we can often receive a notification once any one of our friends has updated her status. Therefore, a_i chooses the message proportionally with $\frac{\#comments+1}{age+1}$, where $\#comments$ is the number of existing comments on such message, and age is the number of times since its publication. As a result, the newly written messages which already have many comments will be chosen with very high probability.

Algorithm 1: *xSocial* Model

```
Input: $\mathcal{A}$, $\mathcal{T}$, $time \leftarrow 0$
while $time < \mathcal{T}$ do
    foreach $a_i \in \mathcal{A}$ do
        with probability $p_i$, add $a_i.S_w$, $a_i.S_a$, and $a_i.S_c$ by 1;otherwise, subtract them all by 1;
        if $a_i.S_w = 0$ then $a_i$ writes a message;
        if $a_i.S_a = 0$ then
            if $SampleUniform(0, 1) < f_i$ then
                the probability of $a_i$ choosing $a_j$ is $P(a_i \rightarrow a_j) \propto \#messages(a_j)$;
                ($\#messages$ is the number of $a_j$'s messages);
            else
                the probability of $a_i$ choosing $a_j$ is $P(a_i \rightarrow a_j) \propto \#cocomments(a_i, a_j)$;
                ($\#cocomments$ is the number of messages commented together);
        if $a_i.S_c = 0$ then
            the probability of $a_i$ choosing a message $o_j$ is $P(a_i \rightarrow o_j) \propto \frac{\#comments+1}{age+1}$;
    $time \leftarrow time + 1$;
```

In summary, all these four major components in our *xSocial* model include very simple rules without assuming any prior sophisticated distributions or constraints. However, as we will show in the next section, both the *buddy network* and the *participation network* generated by this simple model can still match most patterns found on the real datasets. Pseudocode for *xSocial* is shown in algorithm 1.

4.2 Model Analysis

The *xSocial* model incorporates the interlinked evolving process of *buddy network* and *participation network* together, which is much more challenging to model jointly than separately. On the one hand, because the majority of existing graph generators mostly considers modeling a single type of network, there is no natural model to compare with our model. On the other hand, since the *xSocial* model also uses random walk to determine when to put a message or photo, for this point, we can at least make a comparison to the *ZC* model [16]. However, as we will show as follows, even *ZC* model still cannot give the correct distribution of the number of messages that users have posted. In Figure 4, *ZC* model actually gives a folded normal distribution for people's posting behavior with 0 mean and T deviation where T is the number of times that random walk repeats. In contrast, our *xSocial* model matches the power law exponents well : -1.95 vs. -2.07 in Figure 6(m) and 6(q).

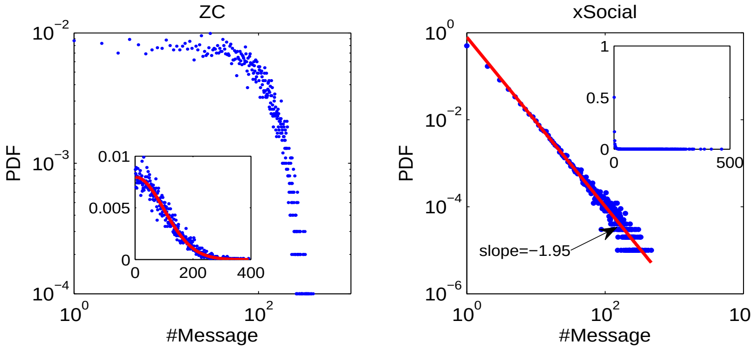


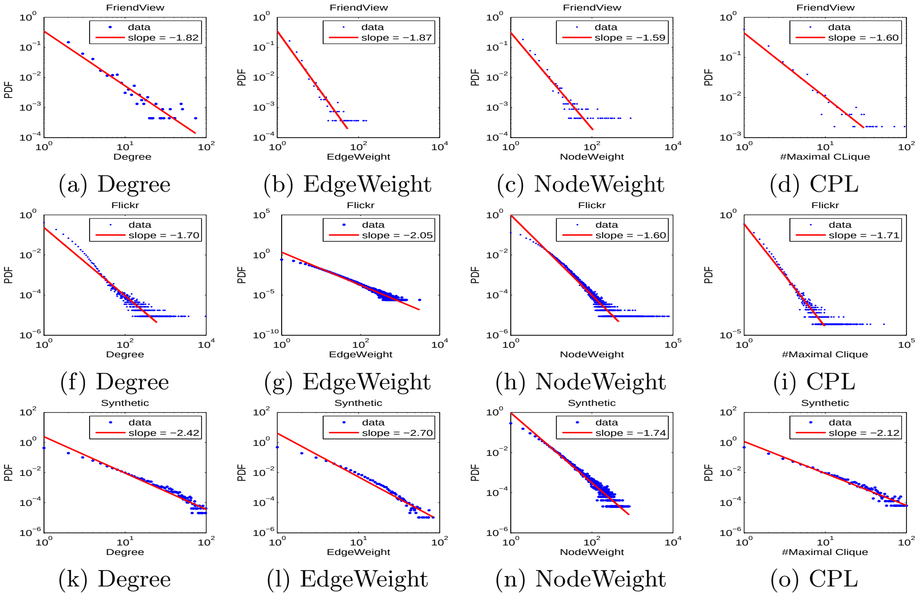
Fig. 4. PDF of #Message. Left : ZC model gives a folded normal distribution $p(x) \propto N_f(0, T)$ where T is the number of times that random walk repeats; Right: xSocial model produces a power-law distribution $p(x) \propto x^{-1.95}$.

Because normal distribution and power-law distribution corresponds to two extreme cases : a *homogeneous* network, vs. a *heterogeneous* network, we believe that this difference arises as a result of the different random walk behaviors. In *ZC* model, the probability for an agent to change his state is all set to 0.5, then each agent has equal opportunity to cross zero (make a post), although they may be distracted by some random events to the nearby state. However, this only models the behavior of active users who frequently use the system although they can be away from his computer for some random distractions. However, a real social networking site not only includes active users, but also involve lots of inactive users. These users just register an account for curiosity in the beginning, but seldom come back and use the system later. As a result, the probability of a random walk to change state in *xSocial* is designed to be different for each agent, which essentially enhances the system's heterogeneity. Because such heterogeneity significantly increases the model complexity, rigorous mathematical proofs are our current ongoing work.

4.3 Model Validation

How accurate is our model? A model is considered to be good if it is able to produce patterns and properties similar to those found in real world networks as many as possible. So, our next goal is to compare the synthetic networks generated by *xSocial* with the networks of FriendView and Flickr. We simulated the model 15,000 times with 100,000 agents. For each agent a_i, as f_i and p_i are independently and uniformly chosen at random from 0 to 1, there is intrinsically no user-predefined parameters for *xSocial* to set.

Our target is to match the following 12 patterns in both *weighted buddy network* and *participation network*. Specifically, for weighted network, we are going to check with Edge-Weight Distribution[26], Node-Weight Distribution[26], and Triangle-Weight Law[13]; for *unweighted* network, we will check with Degree Distribution[28], Snap-shot Power Law(SPL)[23], Clique-Participation Law(CPL)[13], Clique-Degree Power-Law(CDPL)[13], Oscillating connected component size (GCC&NLGCC)[23], Eigen-value Power-Law(EPL)[2]. For each agent, we check with the distribution of the number of written posts, and the CoParticipation Friendship Correlation(CPF). For each post, we check with the distribution of the number of received comments.



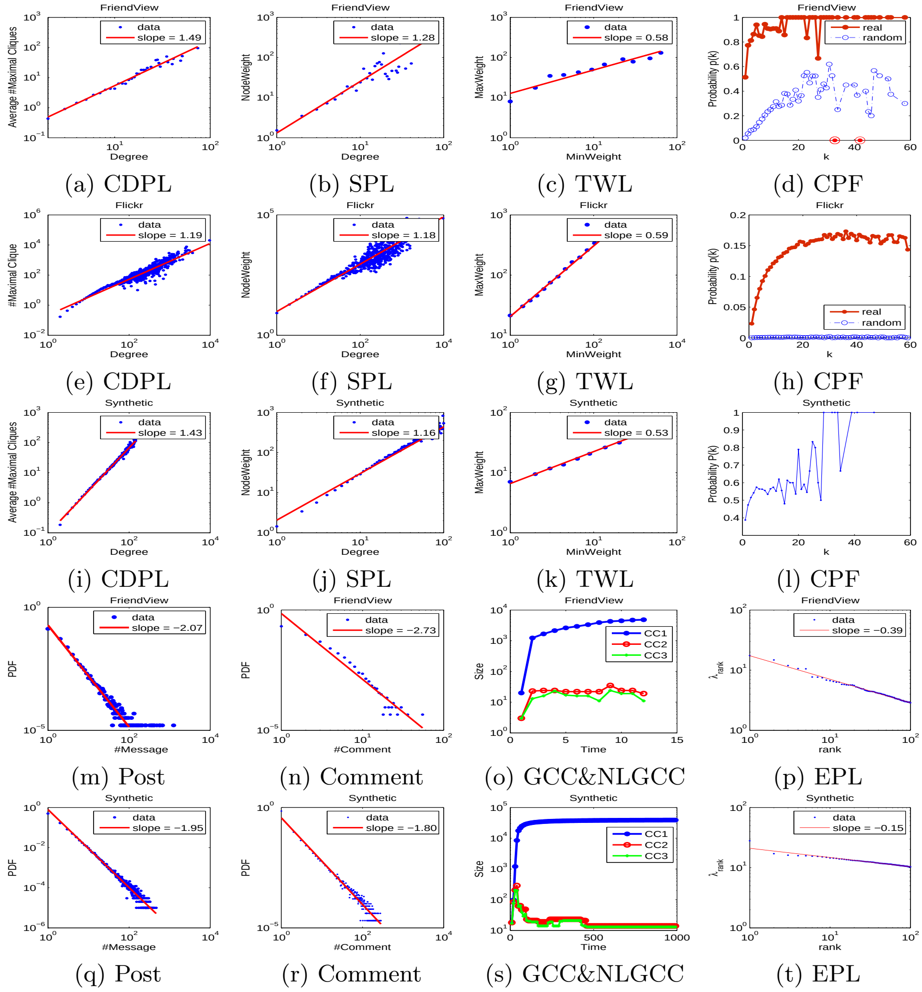
(a) Degree (b) EdgeWeight (c) NodeWeight (d) CPL

(f) Degree (g) EdgeWeight (h) NodeWeight (i) CPL

(k) Degree (l) EdgeWeight (n) NodeWeight (o) CPL

Fig. 5. Comparison of the weighted buddy network, between the two real graphs (top two rows) for FriendView and Flickr, and our synthetic graph (bottom row).

Figure 5 and 6 show the related old and new patterns of the *buddy network* for Flickr and FriendView as well as for *xSocial* results, respectively. Figure 7 further compares the *participation network* (formed by the comment relation) between the real graph and the synthetic graph. Here, we only show the results on Friend View for simplicity, because we have very similar observations on Flickr as well.



(a) CDPL (b) SPL (c) TWL (d) CPF

(e) CDPL (f) SPL (g) TWL (h) CPF

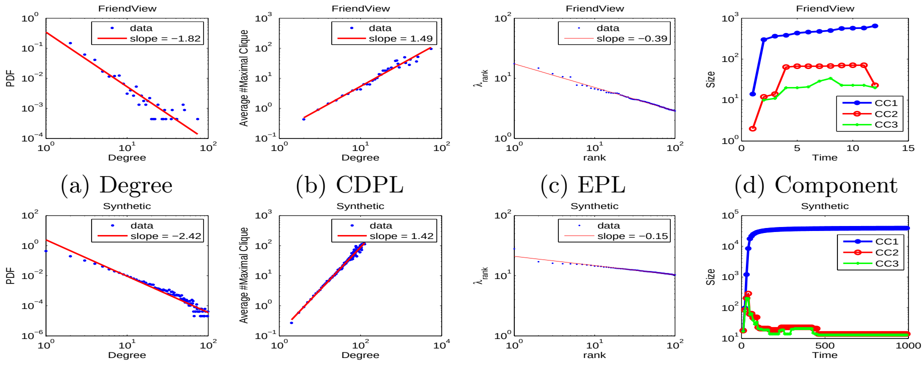
(i) CDPL (j) SPL (k) TWL (l) CPF

(m) Post (n) Comment (o) GCC&NLGCC (p) EPL

(q) Post (r) Comment (s) GCC&NLGCC (t) EPL

Fig. 6. Comparison of the weighted buddy network cont'd, between the two real graphs from (a) to (h), and (m) to (p) for FriendView and Flickr, and our synthetic graphs from (i) to (l), and (q) to (t).

The effective diameter of the weighted *buddy network* and *participation network* of FirendView is roughly 9 and 10, while *xSocial* gives 9.7 and 9.5 respectively. In all cases, *xSocial* can give skewed distributions for both the *buddy network* and *participation network* which are remarkably close to the real ones.



(a) Degree (b) CDPL (c) EPL (d) Component

(e) Degree (f) CDPL (g) EPL (h) Component

Fig. 7. Qualitative comparison of the comment network, between the real graph (top row), and our synthetic graph (bottom row).

5 Conclusion

We study multi-modal networks formed by *friend* and *comment* relations in two different datasets which have over 50 million records and span the course of 2 years. The main contributions are: (a) we proposed the *EigenNetwork* approach to analyzing time-evolving networks, and revealed that there exists temporal regularity with people's on-line social interactions; (b) we discovered inherent correlations between friendship and occurrence in on-line social networking settings; (c) we design the first multi-modal graph generator *xSocial* that stands out from the rest, because it does not include any user predefined parameters, it only uses local information, and it is capable of describing the co-evolving process of multiple weighted social networks that match the old and new patterns observed so far.

References

1. C. C. Aggarwal and P. S. Yu. Outlier detection with uncertain data. In *SDM*, pages 483–493, 2008.
2. L. Akoglu and C. Faloutsos. Rtg: A recursive realistic graph generator using random typing. In *PKDD*, pages 13–28, 2009.
3. L. Akoglu, M. McGlohon, and C. Faloutsos. Rtm: Laws and a recursive generator for weighted time-evolving graphs. In *ICDM*, pages 701–706, 2008.
4. S. Albers, S. Eilts, E. Even-Dar, Y. Mansour, and L. Roditty. On nash equilibria for a network creation game. In *SODA*, pages 89–98, 2006.
5. R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 2002.
6. A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
7. Z. Bi, C. Faloutsos, and F. Korn. The "DGX" distribution for mining massive, skewed data. *KDD*, Aug. 2001. Runner up for Best Paper Award.

8. M. Bilgic and L. Getoor. Effective label acquisition for collective classification. In *KDD*, pages 43–51, 2008.
9. D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1), 2006.
10. A. Chin. Finding Cohesive Subgroups and Relevant Members in the Nokia Friend View Mobile Social Network. *CSE(4).*, pages 278–283, 2009.
11. F. Chung, L. Lu, T. G. Dewey, and D. J. Galas. *J Comput Biol*, 10(5):677–687, 2003.
12. E. D. Demaine, M. Hajiaghayi, H. Mahini, and M. Zadimoghaddam. The price of anarchy in network creation games. In *PODC*, pages 292–298, 2007.
13. N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: patterns and a utility-driven generator. *KDD09*, pages 269–278,2009.
14. P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.*, 5:17–61, 1960.
15. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM*, pages 251–262, Aug-Sept. 1999.
16. M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos. Modeling blog dynamics. *ICWSM09*,2009.
17. N. Laoutaris, L. J. Poplawski, R. Rajaraman, R. Sundaram, and S.-H. Teng. Bounded budget connection (bbc) games or how to make friends and influence people, on a budget. *CoRR*, 2008.
18. J.-G. Lee, J. Han, and X. Li. Trajectory outlier detection: A partition-and-detect framework. In *ICDE 08*, pages 140–149, 2008.
19. J. Leskovec, D. Chakrabarti, J. M. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *PKDD*, pages 133–145, 2005.
20. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD05*, pages 177–187,2005.
21. J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.
22. J. Leskovec, D. Hunttenlocher, and J. Kleinberg. Signed Networks in Social Media. In *CHI*, 2010.
23. M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *KDD08*, pages 524–532, 2008.
24. M. Mitzenmacher. Dynamic models for file sizes and double pareto distributions. 2002.
25. M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.
26. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, A. M. de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New J. Phys.*, 9(6), June 2007.
27. W. Reed and M. Jorgensen. The double pareto-lognormal distribution - a new parametric model for size distribution. 2004.
28. D. Watts. *Small Worlds:The Dynamics of Networks between Order and Randomness.* Princeton University Press, Princeton, 1999.
29. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442,1998.