Large Human Communication Networks: Patterns and a Utility-Driven Generator

Nan Du[†]* Christos Faloutsos* Bai Wang[†] Leman Akoglu*
[†]Beijing University of Posts and Telecommunications,*Carnegie Mellon University
{dunan, christos, lakoglu}@cs.cmu.edu, wangbai@bupt.edu.cn

ABSTRACT

Given a real, and weighted person-to-person network which changes over time, what can we say about the cliques that it contains? Do the incidents of communication, or weights on the edges of a clique follow any pattern? Real, and in-person social networks have many more triangles than chance would dictate. As it turns out, there are many more cliques than one would expect, in surprising patterns.

In this paper, we study massive real-world social networks formed by direct contacts among people through various personal communication services, such as Phone-Call, SMS, IM etc. The contributions are the following: (a) we discover surprising patterns with the cliques, (b) we report power-laws of the weights on the edges of cliques, (c) our real networks follow these patterns such that we can trust them to spot outliers and finally, (d) we propose the first utility-driven graph generator for weighted time-evolving networks, which match the observed patterns. Our study focused on three large datasets, each of which is a different type of communication service, with over one million records, and spans several months of activity.

Categories and Subject Descriptors: H.2.8 [Database Management]:Database Applications—*Data Mining*

General Terms: Experimentation, Measurement

Keywords: Social networks, Graph generators, Cliques

1. INTRODUCTION

Questions have emerged from research on social networks. What patterns should we expect in a network of human-to-human interactions? How can we spot anomalies (e.g., telemarketers, spammers)? What will be the net effect if we lower the price of each phone-call?

Social networks, and graphs in general, have had an increase of interest recently. The related applications are numerous and almost everywhere in people's modern life. Online social networks, like Facebook (`www.facebook.com`) and LinkedIn (`www.linkedin.com`), mimic publicly the telecommunication networks where and what people communicate privately. Product recommendation systems, such as Amazon(`www.amazon.com`) and Netflix(`www.netflix.com`), rely on a network of trust and collaboration [18]. Computer networks have predictable relations regarding intrusion detection [23], security, and virus propagation. It is important in all the above applications to spot anomalies and outliers [1] [4][13][20]. Anomaly detection [9] is tightly connected to patterns: if most of the nodes in our network closely follow a power-law, then the few deviations that do exist are probably outliers. Several patterns have been reported for un-weighted graphs, like small diameter ('six degrees') [35], shrinking diameter [21], scale-free (power-law) [34] or lognormal [6] or Double Pareto LogNormal (DPLN) distributions [24] [30] for the in- and out-degrees etc.

In this paper, we are investigating the following questions:

- When we isolate the cliques in a network, what patterns do they follow? How large are our social circles on average? If someone has many contacts, does that indicate popularity?
- What patterns do the edge weights follow, both in triangles and in general cliques? Specifically, in a triangle, all three nodes are equivalent in topology, but is it normal if all three weights are equal as well?
- How can we design an intuitive generator that will naturally reproduce all the above behaviors? Most existing generators try to mimic the skewed degree distribution, but fail to incorporate the weight information. Here, we want a utility-driven generator, which should try to model the way in which humans decide when and whom to contact. Our guiding principle is that humans balance a trade-off between the cost of the communication (in time and money), and its benefit (in valuable information and emotional support).

Let's elaborate on the last item, the utility-driven generator. Many preferential-attachment [5] guided models assume that a newly-added node is more likely to be linked to the most popular node of the current graph. However, in real world scenarios, incoming nodes are typically unaware of such global structural knowledge of the network. Moreover, most earlier generators dictate that nodes/humans will choose contacts at random; in contrast, we argue that they choose contacts to maximize some utility. Our goal is to design an intuitive graph generator, where each node (a) uses only the local information, and (b) uses no randomness, but instead tries to maximize a well-defined utility function. Such a generator should be carefully designed so that the resulting graphs follow all the observed patterns (old and

KDD'09, June 28–July 1, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-495-9/09/06 ...$5.00.

new). The major advantage over older generators is that it can answer *what if* scenarios. For example, if the connection price of each phone-call goes up, will this decrease the average number of friends/edges? What about a change in the price-per-minute? What if there is a flat rate?

We examine multiple large anonymized human communication networks, where we have the hash-codes of the source, and the destination, as well as the time-stamp of the contact (Phone-Call, IM, or SMS - the specific service is also anonymized). For ease of presentation, we will refer to these contacts as *Phone-Calls* generally. The analysis of human communication networks is important because various personal communication services and applications are ubiquitous. Furthermore, unlike many artificial social networks, such as the scientific collaboration network which emerges as a one-mode projection of the bipartite graph between authors and papers, the massive anonymized human communication networks are formed from the real-time direct contact events of people. They can fully capture the underlying realistic social structures, and lay a solid foundation for our upcoming work.

The paper is organized as follows. Section 2 reviews related work. Section 3 proposes background materials. Section 4 presents our observed patterns. Section 5 describes the utility-driven model. Section 6 gives the conclusion.

2. RELATED WORK

The network formation problem has been studied by many researchers from the fields of statistical physics, economics, game theory, combinatorial optimization and computer science. A major class of network models extend from the classic Erdös-Rényi(ER) random graph model [14] where edges are randomly placed among nodes. Many famous graph generators belong to this class, including the small-world model [36], the preferential-attachment model [5], the forest fire model [21], as well as the recent 'butterfly' model [22]. [See [3] and [8] for a detailed review]

There is another whole class of network models, often referred to as *games of network formation*, mainly from the fields of economics and game theory. Here, linking between two nodes is regarded as a strategic activity and the network structure can arise from the collective interactions between the nodes. Laoutaris [19] proposes a network formation game, where links have costs and lengths, and players have preference weights on the other players, to study the properties of pure Nash equilibria [26] in different settings. Albers [2], Demaine [11], and Fabrikant [16] study a similar game where players do not have fixed budgets and the cost function is defined in terms of the sum of the number of edges. Even-Dar [15] proposes a network creation game where nodes act as buyers and sellers such that the resulting graphs are bipartite.

Moreover, Onnela [28][29] and Nanavati [25] have also used mobile phone-call data to examine and characterize the social interactions of cell-phone users. Seshadri [31] further shows that the degree distribution of large scale mobile phone-call networks can be better fitted using the lesser known but more suitable DPLN distribution[24][30], which is close to yet more precise than the power-law distribution.

In summary, our work differs from earlier work as follows: most research work on network formation games is only interested in the effect of specific linking strategies on the properties of the system equilibria. By contrast, our work studies how the microscopic behavior of each node can collectively influence the emerged macroscopic network structure itself. We are the first to discover the patterns where people can form cliques, and how the edge weights can be distributed in cliques. Moreover, we give the first utility-driven graph generator that is able to reproduce the weighted time-evolving networks, which can have both the old and the new patterns.

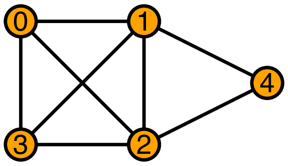


Figure 1: Maximal clique example. Here we have two maximal cliques {0,1,2,3} and {1,2,4}.

3. BACKGROUND

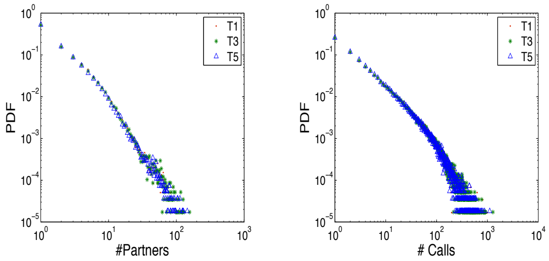
A simple graph $\mathcal{G}$ is represented as a set of nodes $V(\mathcal{G})$ and a set of edges $E(\mathcal{G})$. The weight of the edge $e_{ij} \in E(\mathcal{G})$ is quantified by the number of contact times between node i and j over the studied period, and is denoted by w_{ij}. The total weight w_i of node i is defined as $w_i = \sum_{k=1}^{di} w_{ik}$ where d_i is the degree of node i. In social network analysis, social cohesion [33] is often used to explain and develop sociological theories. Examples of cohesive subgroups include sports teams, work groups, student unions etc. Mathematical analysis of social cohesion has been a hot research topic for many years. The clique model is one of the classic and well-known graph models used for studying cohesive subgroups[33].

Given subgraph G_i, if $\forall u, v \in V(G_i), \exists (u, v) \in E(\mathcal{G})$, then G_i is called a complete subgraph or a *clique* of $\mathcal{G}$. In this paper, we assume that mathematically, a triangle is the smallest clique possible. If there is no other subgraph G_j that is also a clique of $\mathcal{G}$ with $V(G_j) \supset V(G_i)$, G_i is further called a *maximal clique* of $\mathcal{G}$. In Figure 1, {0,1,2,3} and {1,2,4} are two maximal cliques, because cliques {0,1,2}, {0,1,3}, {0,2,3}, and {1,2,3} are included in {0,1,2,3}. $\forall v_i \in V(\mathcal{G})$, let $C(v_i)$ denote the set of the maximal cliques which contain v_i, so $C(0) = \{\{0, 1, 2, 3\}\}$, and $C(1) = \{\{0, 1, 2, 3\}, \{1, 2, 4\}\}$.

The complete clique enumeration is a classic NP-complete problem [7]. However, real world social networks have several unique properties such as the sparsity and scale-freeness. People who share a common friend are highly likely to become friends themselves [17]. This kind of locality generates triangles which further form larger cliques. Consequently, we are able to design an efficient algorithm for practical situations. Following earlier literature, we use the algorithm *Peamc*[12] to find the complete set of all the maximal cliques in our human communication networks.

4. PATTERNS AND OBSERVATIONS

Here, we seek to find the patterns that our social networks have. Starting with a description of the datasets, and the known recurring patterns that hold for the real world networks, we report three newly discovered patterns that our datasets seem to follow. The first is *Clique-Degree Power-Law*(CDPL), correlating the ith largest degree with the av-



(a) PDF of #Partners (b) PDF of #Calls

Figure 2: Known properties of communication networks. (a) is the pdf of the number of partners in graph $\mathcal{G}^{S1}_{T1}$, $\mathcal{G}^{S1}_{T3}$, and $\mathcal{G}^{S1}_{T5}$. (b) is the pdf of the total calls in the same graphs. Both are in logarithmic scales, and follow a DPLN distribution. The rest networks behave similarly.

erage number of maximal cliques, which seems to remain rather stable over time so that we trust them to further detect outliers and spot anomalies. The second is *Clique Participation Law*(CPL), which gives the distribution of the number of maximal cliques that each node participates in. Finally, the third comes *Triangle Weight Law*(TWL), describing how the weights are distributed on the edges of triangles, based on which we could further make predictions about the missing values of the edge weights in time-evolving weighted networks.

4.1 Data Description

The datasets analyzed are made of a large collection of records from several human communication services including voice, data, IM, SMS etc. Each record is represented as a triple $< ID_i, ID_j, Time >$, where $< ID_i >$ and $< ID_j >$ are generally referred to as the *caller* and *callee*. During a particular time period, there can be multiple times for a pair of people to communicate with each other, and the accumulated number of communication times between ID_i and ID_j is defined as the edge weight between node i and j. We have the weighted graphs extracted from the records of three types of services ($S1$, $S2$ and $S3$), referred to as $\mathcal{G}^{S1}$, $\mathcal{G}^{S2}$, and $\mathcal{G}^{S3}$ respectively. Each type of service has on average about 1 million records, which were collected by different geographic locations. Apart from this spatial diversity and the service type variety, we also incorporate temporal diversity by collecting data for each type of service during five consecutive time periods represented from $T1$ to $T5$, so $\mathcal{G}^{S1}_{T1}$ is the graph of service type $S1$ in time period $T1$, and $\mathcal{G}^{S2}_{T5}$ is the graph of service type $S2$ in time period $T5$ etc.

Notice that we only focus on the link between the caller and callee. It is important to know that our work is only an aggregate statistical analysis, and therefore, we do not study any individual's behavior from any specific type of communication service. More importantly, any information that could identify users is stripped to access. We only use the encrypted user id in this study, and restrict our interest only in the statistical findings that are held within the networks.

4.2 Old Patterns

We first consider the total number of unique callers and callees which are often referred to as the partners associated with every user. This essentially corresponds to the degree of each node. Then we calculate the total number of calls made or received by each user, which is represented by the node weight. We show the full results in Figure 2 for $\mathcal{G}^{S1}_{T1}$ (the beginning), $\mathcal{G}^{S1}_{T3}$ (the middle), and $\mathcal{G}^{S1}_{T5}$ (the last), because $\mathcal{G}^{S2}_{T1} \sim \mathcal{G}^{S2}_{T5}$ and $\mathcal{G}^{S3}_{T1} \sim \mathcal{G}^{S3}_{T5}$ have similar observations.

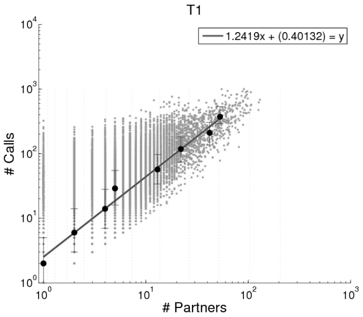


Figure 3: Partners-Calls distribution of $\mathcal{G}^{S1}_{T1}$ in logarithmic scales. Black dots are the medians by logarithmic binning. Least square fit slope is 1.24.

We also study the correlation between the number of partners and the total number of contacts per user (shown in Figure 3). It is observed that there is a "fortification effect" leading to a *Snapshot Power Law*(SPL) [22]. The more partners an individual has, the superlinearly more calls he makes and receives. Here, the result of service type $S1$ in the first time period $\mathcal{G}^{S1}_{T1}$ is reported, for that the fortification effect is very stable and leads to similar results in the rest.

4.3 New Patterns

In this section, we will give the newly discovered findings of our human communication networks, and discuss the potential ways in which they can be utilized.

4.3.1 Clique-Degree Power-Law

As defined previously, $\forall v_i \in V(\mathcal{G})$, d_i is the number of all the partners that v_i has, and $C(v_i)$ represents the set of all the maximal cliques that v_i participates in. Is there any relationship between d_i and $|C(v_i)|$? We can imagine that if a particular user has doubled his partners, it tends to be easier for him to participate in doubled social circles as well. This kind of relationship seems to be linear, and sounds reasonable. However, this is often not the case. For our real world social networks, the number of social circles actually over-doubles by following a *Clique-Degree Power-Law.*

Observation 1. (Clique-Degree Power-Law (CDPL)). *The number of maximal cliques that a node participates in, is super-linearly related to its degree. Given d_i and $C^{d_i}_{avg}$, they follow a power-law :*

$$C^{d_i}_{avg} \propto d_i^{\alpha} \qquad (1)$$

where α is the exponent of CDPL, and remains about constant over time.

Figure 4 plots the number of partners vs. the number of maximal cliques averaged over all the nodes with that many of partners, from $T1$ to $T5$. The result is surprising because for any given node, the clique-participation is superlinearly related to its degree. In addition, we notice that the

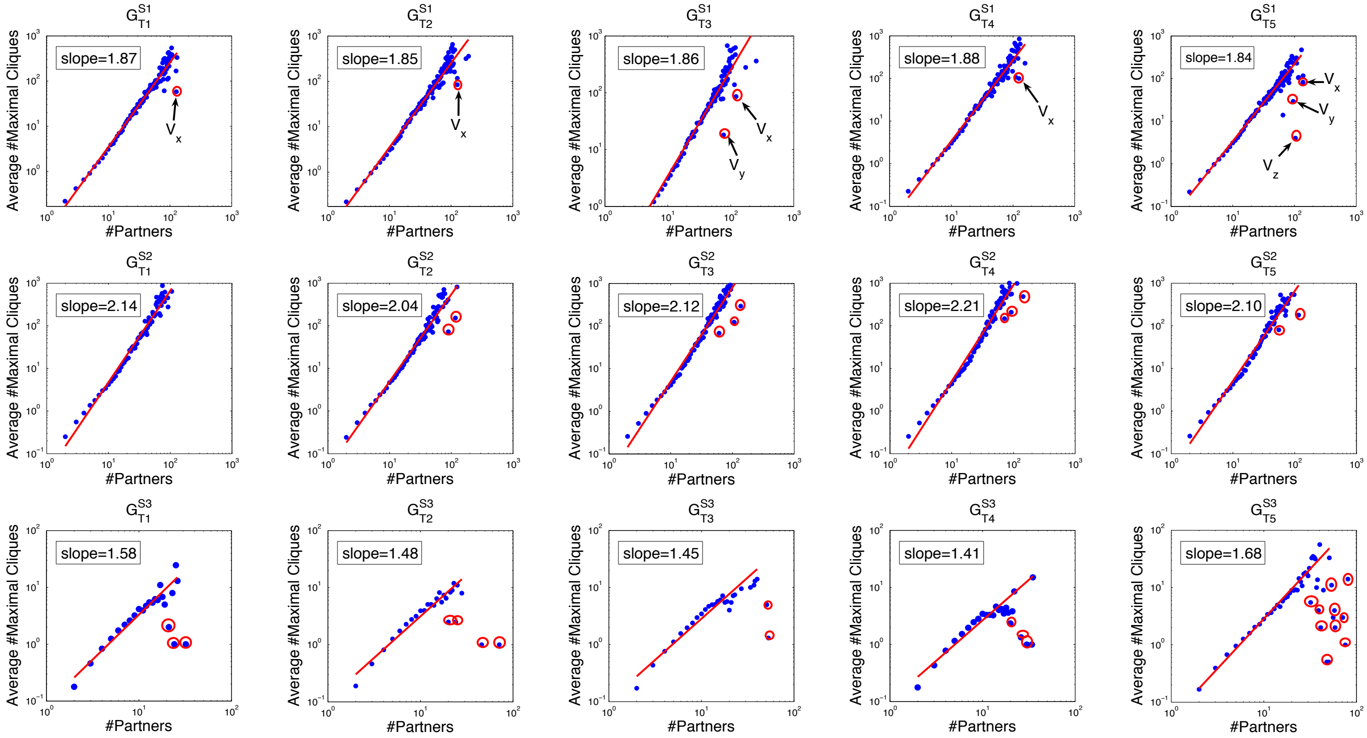
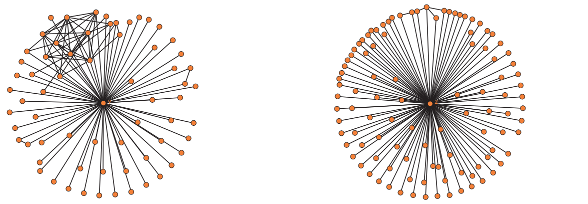


Figure 4: Clique-Degree Power-Law. Number of partners vs. the average number of maximal cliques in $G^{S1} \sim G^{S3}$ from $T1$ to $T5$. All of the exponents are fitted with $R^2 > 0.95$. Notice that *CDPL* is very stable over time. The detected outliers are marked by red circles.



(a) Centered with vertex v_y (b) Centered with vertex v_z

Figure 5: Detected typical outliers. Both v_y and v_z (in $\mathcal{G}^{S1}_{T3}$ and $\mathcal{G}^{S1}_{T5}$ from Figure 4) have too many unrelated partners, resulting in a star-like subgraph.

exponent takes values in the range $[1.84, 1.88]$, $[2.04, 2.21]$ and $[1.41, 1.58]$ for G^{S1}, G^{S2}, and G^{S3}, which seems to be stable over time.

The direct application of *CDPL* is to spot outliers. In Figure 4, all of the detected anomalies are marked by red circles. We can see that these points present a clear pattern which does not conform to the established normal behavior. In other words, for these users the actual number of the maximal cliques that they belong to is significantly distant from the one that they should have according the number of their friends.

It is also interesting to notice that some outliers are stable and persistent, such as node v_x and v_y from $\mathcal{G}^{S1}_{T1}$ to $\mathcal{G}^{S1}_{T3}$, while others are more casual and bursty, such as node v_z in $\mathcal{G}^{S1}_{T5}$, and the circled outliers in $\mathcal{G}^{S3}_{T5}$. Figure 5 shows the egocentric subgraphs centered with node v_y and v_z, which are composed of the connections among their neighbors in $T5$. Clearly, for node v_y, although it has a large number of partners, it only belongs to few maximal cliques on the left upper part of Figure 5(a). As to node v_z, almost no connections exist among its partners. Because any automatic customer service id is excluded from our communication networks, the anomalous behavior of v_y and v_z makes them more like the tele-marketers. In fact, there are more outliers in the last time period $T5$ than the others, especially in the network $\mathcal{G}^{S3}_{T5}$ of the third communication service. We guess it is probably because there is actually a big holiday in $T5$, and the third communication service is the cheapest for broadcasting messages.

4.3.2 *Clique Participation Law*

Based on the discovered maximal cliques, we are able to study how people could get involved into them. Figure 6 shows the distribution of the number of maximal cliques that people actually participate in. That is, in graph $\mathcal{G}$, it plots the correlation between the number of maximal cliques (x-axis) and the pdf of nodes (y-axis) that get involved in that many of maximal cliques. We observe that there exists a power-law followed by this kind of relationship, which is called *Clique Participation Law.*

Observation 2. (Clique Participation Law (CPL)). *For a given number of maximal cliques, say n_{clique}, and the set $V_{clique} = \{v_i | v_i \in V(\mathcal{G}), |C(v_i)| = n_{clique}\}$, we have*

$$n_{clique} \propto |V_{clique}|^{cp} \qquad (2)$$

where cp is the clique participation exponent of CPL, and keeps about constant over time.

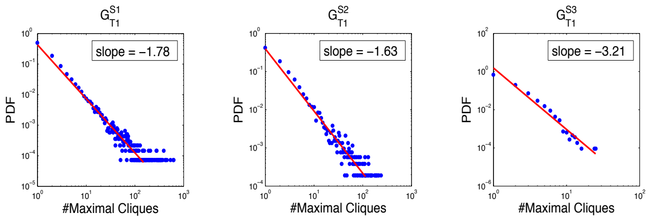


Figure 6: Clique-Participation Law. PDF of #Maximal Cliques in $\mathcal{G}^{S1}_{T1} \sim \mathcal{G}^{S3}_{T1}$. The rest graphs behave similarly.

	T1	T2	T3	T4	T5
G^{S1}	-1.78	-1.74	-1.76	-1.70	-1.68
G^{S2}	-1.63	-1.56	-1.52	-1.56	-1.54
G^{S3}	-3.21	-3.50	-3.46	-3.50	-3.01

Table 1: Power-Law exponents of CPL in $G^{S1} \sim G^{S3}$ from $T1$ to $T5$. Notice the stability.

According to the above discussion, for most people in real world social networks, they are often involved in a small number of maximal cliques (or social circles). Only a few of them are really 'social butterflies' that can actively span many social circles simultaneously. In Figure 6, we report the results from $\mathcal{G}^{S1} \sim \mathcal{G}^{S3}$ only in $T1$ for brevity, because in Table 1 we observe that *CPL* is rather stable over time, leading to similar plots in the rest.

Actually, the *CPL* pattern could be potentially applied to help the operators to make better designed family plans. Because we have a model of the distribution of user behavior to form close-knit groups, we can propose better pricing strategies that charge users differently according to the size of their social circles. For example, in most cases people only belong to one or two cliques, which may be formed by their families or best friends. We can design specific billing plans which are favorable to the communications among members of the same clique who are also the customers of the same operator. Even if our friends are the customers of other operators, we may still like to invite them to join us, because we know that it will be good for all of us. As a result, this could implicitly improve the loyalty of the current users, and may further help to increase the rate at which new customers sign up the plans. Moreover, we can also reward a few loyal users who span multiple social groups, because they might help to achieve a quick market promotion by introducing new products and services to their friends.

4.3.3 Triangle Weight Law

According to the clique definition, each node in a clique has connections with all the other nodes. Although it is very intuitive that all these nodes are equivalent in topology, will this also mean that they could have equally close relationships? In our communication networks, the edge weight w_{ij} gives the total number of contact times between i and j, which is an important indicator to show how intimately they could relate to each other. Since that triangle is the base case of a clique, given any triangle $\{i, j, k\}$, will w_{ij}, w_{ik}, and w_{jk} hold approximately equal values because of the structure equivalence between i, j, and k? Although this intuitive conjecture seems to make sense, we have made very unexpected and striking discoveries in the real social networks, which are described as follows.

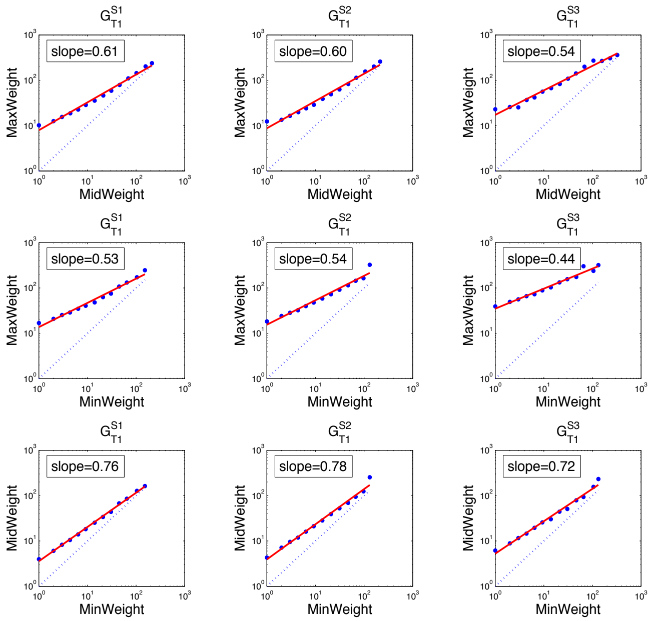


Figure 7: Triangle Weight Law. Minimum, medium, and maximum weights in all 3 pairs are plotted in logarithmic scales. Least square fits all have $R^2 > 0.95$ in $\mathcal{G}^{S1}_{T1} \sim \mathcal{G}^{S3}_{T1}$.

Observation 3. (Triangle Weight Law (TWL)). *For any triangle, let $MaxWeight$, $MidWeight$, and $MinWeight$ denote the maximum, medium, and the minimum edge weight respectively. In all our graphs, they follow three power-laws:*

$$MaxWeight \propto MidWeight^{\alpha} \tag{3}$$

$$MaxWeight \propto MinWeight^{\beta} \tag{4}$$

$$MidWeight \propto MinWeight^{\gamma} \tag{5}$$

where α, β, and γ are the power-law exponents which remain constant in weighted time-evolving social networks.

As a result, for the given triangle $\{i, j, k\}$, rather than being approximately equal, w_{ij}, w_{ik} and w_{jk} are significantly different from each other. Figure 7 gives the results from the networks $\mathcal{G}^{S1} \sim \mathcal{G}^{S3}$ in the same time period $T1$. To achieve a good fit, we bucketize the x-axis with logarithmic binning [27], and for each bin, we compute the average value of y. The dotted line means the value of x equals to the value of y. Moreover, Figure 8 shows the three exponents of *TWL* in $G^{S1} \sim G^{S3}$ from $T1$ to $T5$. Notice that α, β, and γ of these graphs take values in the range [0.5,0.7], [0.4,0.6], and [0.7,0.8], which seem persistent and stable.

In practical situations, due to missing data we can only have partial network information to analyze. For example, in Figure 9, given the weighted egocentric subgraph that link e_{23} belongs to, what can we say about the missing w_{23}? Where the link prediction [10] tries to predict between which unconnected nodes a link will form next, our problem here concerns how to estimate the value of an edge weight, because we already know there is a link between node 2 and 3. We formulate this problem as the ***weight prediction*** prob-

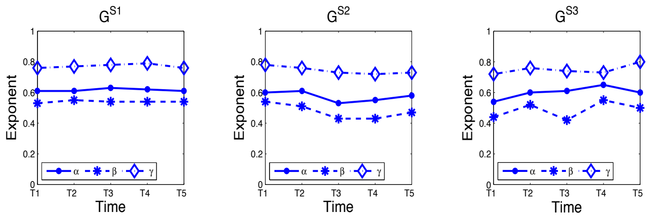


Figure 8: Persistence of Triangle Weight Law. Exponent α, β, and γ (red, blue, green) in $G^{S1} \sim G^{S3}$ remain about constant from $T1$ to $T5$.

lem, which not only is important to fill and complete the missing values, but also is useful for discovering anomalous links, because if the actual value of w_{23} is significantly different from the predicted value, it would be highly unusual.

Based on the above discussion, TWL can help us to solve the weight prediction problem. Formally, given $e_{ij} \in E(\mathcal{G})$, let $\triangle$ denote the set of all the edges (excluding e_{ij} itself) of the triangles that e_{ij} belongs to. $\forall e_k \in \triangle$, $w(e_k)$ denotes the weight of e_k. The minimum and maximum values of $w(e_k)$ are represented as $\triangle_{min}$ and $\triangle_{max}$ accordingly. On one hand, if $w_{ij} < \triangle_{min}$ or $w_{ij} > \triangle_{max}$, the numerical relationship between w_{ij} and the weights on the other two edges is determined, so we can use either equation (4) and (5) or (3) and (4) to estimate w_{ij} directly. On the other, if $w_{ij} \in [\triangle_{min}, \triangle_{max}]$, w_{ij} might be the minimum in one triangle, while might be the maximum in another triangle. Thus, for $\forall e_k \in \triangle$, we define $\phi_{(e_{ij},e_k)}(x)$ to represent one of the three equations (3) $\sim$ (5) based on the particular numerical relationship that e_{ij} and e_k could hold. The return value of $\phi_{(e_{ij},e_k)}(x)$ is the estimated weight for edge e_k given the possible value x of w_{ij}. Here, we assume that all edge weights are positive integers. Let w_{min} be the minimum estimated value of w_{ij} when $w_{ij} < \triangle_{min}$, and w_{max} be the maximum estimated value of w_{ij} when $w_{ij} > \triangle_{max}$. Then the optimal value of w_{ij} is given as:

$$\hat{w}_{ij} = argmin \sum_{e_k \in \triangle} (w(e_k) - \phi_{(e_{ij},e_k)}(x)) \qquad (6)$$

where $x \in [w_{min}, w_{max}]$. We evaluate this approach in $\mathcal{G}^{S1} \sim \mathcal{G}^{S3}$ by comparing $\hat{w}_{ij}$ with w_{ij} for each edge in $T1$, $T3$ and $T5$. Due to the persistence of TWL, we set $\alpha = 1.5$, $\beta = 2.2$, and $\gamma = 1.2$ for $\mathcal{G}^{S1}$; $\alpha = 1.3$, $\beta = 1.7$, and $\gamma = 1.4$ for $\mathcal{G}^{S2}$; $\alpha = 1.4$, $\beta = 2.1$, and $\gamma = 1.3$ for $\mathcal{G}^{S3}$. Let $\epsilon = |\hat{w}_{ij} - w_{ij}|$ denote the prediction error. The the average prediction accuracy of $\epsilon = 0$ (the exact prediction) and $\epsilon = 1$ is around 0.21 and 0.32 accordingly. One problem of this simple method is that it can not predict w_{ij}, if the edge e_{ij} does not belong to any triangle. To solve this problem, and further improve the prediction accuracy is an area of future work.

5. UTILITY-DRIVEN GENERATIVE MODEL

The next goal is to design a generative model that mimics people's natural communication behaviors. The guiding principle is that such a model should be *utility driven*, as opposed to earlier models (preferential attachment [5], forest-fire [21], butterfly [22], etc.) which are mainly randomness-guided generators.

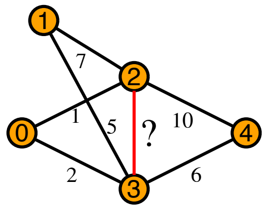


Figure 9: Weight Prediction Problem. What can we say about w_{23} ?

On one hand, every communication, such as phone-call, SMS, and e-mail, has a cost in terms of money, time, and equipment. On the other hand, it has a benefit, otherwise humans would not do it. The benefits can be psychological and emotional (talking to friends makes us happy), or monetary (stock tip), or desirable in other ways. For ease of presentation, we refer to the benefit as if it is measured by *emotional dollars*. The point of this thought experiment is to set up a utility-driven model for the social contacts of humans, which should be more realistic and more informative than the ones using randomness.

Therefore, we assume that people are rational agents, which means tele-marketers are excluded from the model, and we design our generator to guide the behavior of each agent according to a well-defined utility function. Ideally, the fundamental macro-phenomena of a social network should then emerge from the simple local behavior of each agent/human.

5.1 Model Description

Following the above discussion, we now present our utility-driven model PaC as a ***Pay and Call*** game. Assume a setting where a set $\mathcal{A}$ of n distinct agents create links to one another through phone calls. In every round of the game, each agent's strategy is to choose among other peers to whom he will make calls and build links. Links are undirected. Once agent $a_i \in \mathcal{A}$ calls $a_j \in \mathcal{A}$, there will be a link between them. The total number of phone-calls that a_i and a_j give to each other is treated as the *weight* on the undirected link between them. The PaC model essentially includes the following four ingredients:

- It adopts the agent-based modeling approach. Each agent has a *friendliness* value, an *exponential lifetime*, a certain amount of *capital*, and the *expected* payoffs from talking to strangers.
- The goal of each agent is to invest his limited capital into phone-calls and maximize the potential payoffs from each conversation.
- The per-minute gain of a conversation will be gradually *saturated*, and finally both of the callers and callees will lose interest, and stop the conversation.
- Each agent a_i can ask his partners for recommendations. Every partner recommends the profitable agents from his own partners, so a_i benefits from talking to the most profitable agent within the recommendations.

Friendliness and ***Exponential Lifetime***. Each agent has a *friendliness* value $F_i \in (0, 1)$ to show his personality. F_i approaching to 1 means the agent is very open and friendly, and F_i close to 0 means he is very shy and introverted. a_i has a probability P_l, uniformly chosen from 0 to 1, to stay in the game, and has the probability $1 - P_l$ to leave the game, so that we can simulate the mixture of different ages in a real network. Once an old agent leaves, all his links will be removed, and a new agent replaces his position with the *friendliness* and P_l initialized to new values.

Utility-Driven Phone-Calls and ***Saturation***. An agent's payoffs are the difference between the benefits and costs. The benefits are defined based on the following considerations. Two open agents usually can benefit emotionally from a happy conversation. When an open agent meets a shy agent, they may benefit less from their conversation. Finally, two shy agents might gain little in the end. In addition, after two agents have been talking for a while, they may gradually lose interest, and gain less emotionally as time goes by. For agent a_i and a_j, they can achieve $\sqrt{F_i \times F_j} \times \alpha^{m-1}$ emotional dollars per minute from a conversation, where $\alpha \in (0, 1)$ is called the *saturation* factor to represent the loss of interest, and m is the number of minutes for which they have been talking.

For an m-minutes long conversation, the total benefits are defined as

$$\begin{aligned} benefits &= \sqrt{F_i \times F_j} \times \sum (1 + \alpha + \alpha^2 + ... + \alpha^{m-1}) \\ &= \sqrt{F_i \times F_j} \times \frac{1 - \alpha^m}{1 - \alpha} \qquad (7) \end{aligned}$$

The costs are the expenses of phone-calls, which include C_{ini} and C_{pm}. C_{ini} is the cost to initiate a phone-call, and C_{pm} is the per-minute fee. The total costs for an m-minutes call will be $C_{ini} + m \times C_{pm}$, so our utility function is defined as

$$payoffs = benefits - C_{ini} - m \times C_{pm} \qquad (8)$$

and each agent starts and maintains a conversation until the payoffs by equation 8 reach the maximum value or the agent has used all his money.

Expected Payoffs on Strangers. At first, each agent is given an initial *capital* which is enough to make one call only. Since none of the agents have ever talked before, for agent a_i, he first uniformly calls a stranger a_j, and keeps the conversation until either the payoffs by equation 8 begin to decrease or he spends all his money in the call ($a_i.capital <= 0$). When the call is finished, a_i and a_j will achieve the payoffs P_j from the conversation. A link is built between a_i and a_j with weight 1, and a_i will remember the payoffs P_j earned by talking to a_j. Because a_j was first a stranger to a_i before they met, a_i also updates his *expected* payoffs from talking to strangers as :

$$S_{exp} = \frac{\sum P_i}{1 + S} \qquad (9)$$

where S is the total number of times talking to strangers, and P_i is the payoffs achieved at each time. S_{exp} is initialized to 0 in the beginning. In each round of the game, agent a_i is only allowed to call a_j for one time. If a_i still has some money left (note that the payoffs earned in the current round can only be used in the next round), he will continue to interact with other strangers.

Recommendations: Once agent a_i has some partners, he will first prioritize his partners according to the remembered payoffs, and talk to them respectively. If the payoffs of the currently chosen partner is less than a_i's expected payoffs from strangers ($S_{exp} \geq 0$), a_i will stop talking to partners and choose to call strangers again. He first asks his partners for recommendations. Every partner will tell a_i how much money he actually earned by talking to his own partners last time. a_i can then pick the most profitable agent out of the partners of his partners. If all the recommended agents are already his partners, a_i will uniformly choose a stranger from the rest.

In summary, the *PaC* model is formulated in the pseudocode of algorithm 1.

Algorithm 1: PaC Model

```
Input: $C_{ini}$, $C_{pm}$, $\alpha$
foreach $a_i \in \mathcal{A}$ do
    if $a_i$ stays with probability $a_i.P_l$ and $a_i.capital \geq C_{ini} + C_{pm}$ then
        if $N(a_i) = \emptyset$ then
            Talk2Strangers($a_i$)
        else
            Talk2Partners($a_i$)
    if $a_i$ quits with probability $1 - a_i.P_l$ then
        Replace $a_i$ with a newly born agent
```

Procedure `Talk2Strangers(`a_i`)`

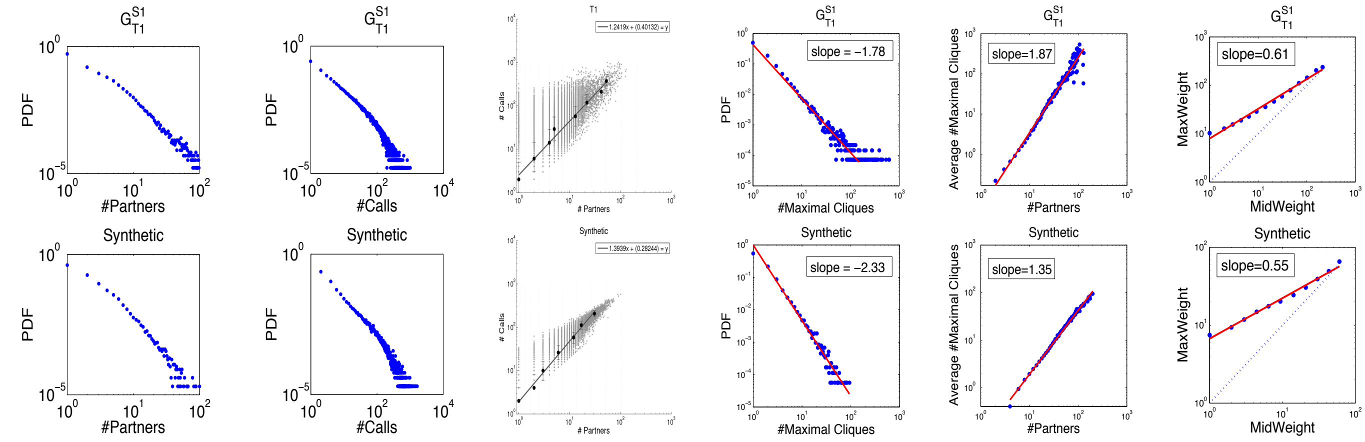
```
Input: current agent $a_i$
$total \leftarrow 0$
if $N(a_i) \neq \emptyset$ and $a_i$ finds the most profitable agent he never talks to from the recommendations then
    $a_j \leftarrow$ the most profitable agent
else
    $a_j \leftarrow$GetRandom($\mathcal{A}$)
while $a_i.capital \geq C_{ini} + C_{pm}$ and $a_i.S_{exp} \geq 0$ do
    maximize $payoffs$ with constraint $a_i.capital \geq 0$ by equation 8
    add $a_j$ to $N(a_i)$, add $a_i$ to $N(a_j)$
    $a_j.capital \leftarrow a_j.capital + payoffs$
    $total \leftarrow total + payoffs$
    update $a_i.S_{exp}$ and $a_j.S_{exp}$ by equation 9
$a_i.capital \leftarrow a_i.capital + total$
```

Procedure `Talk2Partners(`a_i`)`

```
Input: current agent $a_i$
$total \leftarrow 0$
prioritize $N(a_i)$ according to the descending order of the remembered payoffs
for $k \leftarrow 1$ to $|N(a_i)|$ do
    $P_k \leftarrow a_i$'s remembered payoffs from $a_k$
    if $P_k \geq a_i.S_{exp}$ then
        maximize $payoffs$ with constraint $a_i.capital \geq 0$ by equation 8
        increase the weight on the link between $a_i$ and $a_k$ by 1
        $a_j.capital \leftarrow a_j.capital + payoffs$
        $total \leftarrow total + payoffs$
    else
        Talk2Strangers($a_i$)
        break
    if $a_i.capital \leq 0$ then break
$a_i.capital \leftarrow a_i.capital + total$
```

5.2 Model Validation

How accurate is our model? Our goal here is to show that our model is able to generate degree, weight and clique distributions that mimic a real graph like our communication networks. Notice that we only want to show qualitative match of the properties. Exact fitting is outside the scope of this paper. We decided to test our model with respect to all the usual patterns, and specifically the degree distribution, weight distribution, as well as the snapshot power law. We also want to qualitatively check against our newly discovered clique-related patterns, the *CDPL*, *CPL*, and the *TWL*. We simulated the model 35 times for 100,000 nodes, with $C_{ini} = 0.1$, $C_{pm} = 0.4$ and $\alpha = 0.9$. For each agent



(a)PDF of #Partners (b)PDF of #Calls (c) SPL (d) CPL (e) CDPL (f) TWL

Figure 10: Qualitative comparison between the real graph (top row) and our synthetic graph (bottom row). *PaC* gives skewed distributions like the real ones.

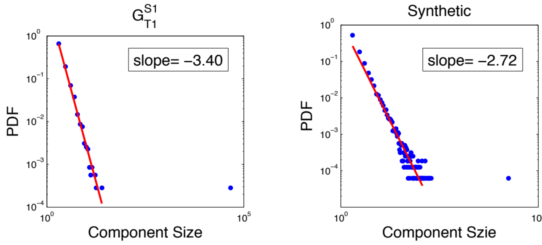


Figure 11: PDF of Connected-Component Size. The sizes of the connected-components in G_{T1}^{S1} (the left) and in the synthetic graph (the right) follow the power-law distribution.

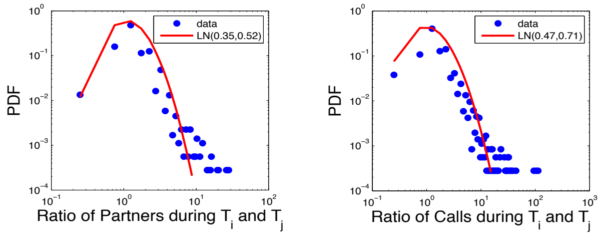


Figure 12: The ratio of partners (left), and calls (right) between two different snapshots of *PaC* follow the lognormal distribution. The parabolic line is fitted in red.

a_i, F_i and P_l are randomly chosen from 0 to 1. Figure 10 shows the results of these checkpoints. The top row is the actual graph $\mathcal{G}_{T1}^{S1}$, and the bottom row is a synthetic graph, generated by our *PaC* model. Figure 10(a)∼(c) show the old patterns, and Figure 10(d)∼(f) illustrate the new ones. Moreover, in Figure 11, we see that except for the giant connected component which is an isolated point distant from the rest, the size distribution of the connected-components conforms to a power-law. The exponents take values within the range observed in real world networks with a least-square fit of $R^2 > 0.95$. In all cases, notice that *PaC* gives skewed distributions that are remarkably close to the real ones.

5.3 Model Analysis

From earlier research [24][30], we understand how heavy-tailed distributions such as power-law, lognormal and DPLN could arise for the degree distribution and the node weight distribution. According to Mitzenmacher [24], lognormal distributions can be naturally generated by *multiplicative processes*. For a biological example, at each step j, an organism may grow or shrink by a certain percentage according to a random variable F_j. If X_j denotes the current size of the organism, $X_j = F_j X_{j-1}$ where F_j is independent of X_{j-1}. Consider $\ln X_j = \ln X_0 + \sum_{k=1}^{j} \ln F_k$ if $F_k, 1 \leq k \leq j$, are independent lognormal distributions, then X_j is always lognormal. If F_k are not lognormal, but are independent and identically distributed with finite mean and variance, by Central Limit Theorem, $\sum_{k=1}^{j} \ln F_k$ converges to a normal distribution, and X_j will asymptotically approach a lognormal distribution [24]. If X_j is lower bounded by a minimum value, then the distribution will become a power-law. If we sample the series from X_0 to X_j by a geometrically distributed random time k, we will have a geometric mixture of lognormal distributions. This will turn out to be a DPLN distribution with two power-laws at both tails [24].

Following the *PowerTrack* method in [30], we analyze empirically the generative process of our *PaC* model by taking two snapshots $\mathcal{G}_{T_i}$ and $\mathcal{G}_{T_j}$ at time step T_i and T_j with $j - i \geq 1$. Among the common agents between $\mathcal{G}_{T_i}$ and $\mathcal{G}_{T_j}$, we calculate the ratio X_{T_j}/X_{T_i}, where X_t represents either the degree or the weight for each node. In Figure 12, the distributions of the ratio for both of the degree and weight appear to be parabolic in logarithmic scales. This provides good evidence that a lognormal multiplicative process is involved in the temporal evolution of our model. Another important issue is that we also need to test the independence between partners and their ratios, and the same for the calls. Here, the correlation coefficients, which are necessary but not sufficient for independence[30], are very small: -0.02 and -0.04 for partners and calls respectively. Finally, in each round of the game, every agent has the probability P_l to stay or leave the game, which essentially leads to a geometric lifetime. Therefore, although we do not explicitly assume any prior distribution about the ratio (the File Model in [24] explicitly assumes a lognormal distribution for F_k), the *PaC* model can still mimic the DPLN degree distribution and the node weight distribution which are identical with the real social networks. By comparing with the existing graph generators, we see that preferential-attachment guided models usually ignore the weight information, and only generate the giant connected component [22]. The *butterfly* [22] model can reproduce all the connected components, however it does not include the weight either. In contrast, our model is able to reproduce the networks that have not only the patterns holding in un-weighted networks, but also the patterns followed by the weighted networks.

6. CONCLUSION

The main contributions are: (a) we found surprising patterns that cliques follow, like the *CDPL* and *CPL*; (b) we observed the weights on the edges of triangles followed power-laws *TWL*; (c) the discovered patterns are stable and persistent in several, diverse, real social networks, and finally (d) we propose the first utility-driven graph generator for weighted time-evolving networks.

The (anonymized) datasets had over one million records, spanning several months, and over various (anonymized) services. Thanks to our new patterns, we discovered several outliers. Closer inspection showed that they indeed had very suspicious behaviors. Further investigation was impossible, due to privacy issues. Moreover, our *PaC* model stands out from the rest, because (a) it does not use randomness (using a utility function instead) (b) it only uses local information (c) it still generates graphs that follow all the old and new patterns. Based on its utility function of *PaC*, we can explore what is the impact of, say, lower prices, on the shape of the network, as well as several other 'what if' questions.

Acknowledgments. This material is based upon work supported by the National Science Foundation of China under Grant No.60402011, the National Key Technology R&D Program of China under Grant No.2006BAH03B05, the IBM China Research Laboratory Project No.JTP200806014-3, the National Science Foundation under Grant No. IIS-0705359, the iCAST project sponsored by the National Science Council, Taiwan, under Grant No.NSC97-2745-P-001-001, and the Ministry of Economic Affairs, Taiwan, under Grant No.97-EC-17-A-02-R7-0823. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, or other funding parties. We would like to thank Charlotte Yano and Mary McGlohon for valuable comments, and Qi Ye for preprocessing the datasets.

7. REFERENCES

[1] C. C. Aggarwal and P. S. Yu. Outlier detection with uncertain data. In *SDM*, pages 483–493, 2008.

[2] 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.

[3] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.

[4] B. Wu and D. B. Davison Identifying link farm spam pages. In *WWW 2005*, pages 820–829, 2005.

[5] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[6] Z. Bi, C. Faloutsos, and F. Korn. The "DGX" distribution for mining massive, skewed data. *SIGKDD 2001*, pages 17–26

[7] F. Cazals and C. Karande. Reporting maximal cliques: new insights into an old problem. *Research Report, http://cgal.inria.fr/Publications/2005/CK05b*, (5615), 2005.

[8] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1), 2006.

[9] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *To Appear in ACM Computing Survey.*

[10] L.-N. David and K. Jon. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[11] E. D. Demaine, M. Hajiaghayi, H. Mahini, and M. Zadimoghaddam. The price of anarchy in network creation games. In *PODC*, pages 292–298, 2007.

[12] N. Du, B. Wu, and B. Wang. A parallel algorithm for enumerating all maximal cliques in complex networks. In *ICDM2006 Mining Complex Data Workshop*, pages 320–324.

[13] Z. Elena, K. Aleksander, and G. Lise. Trusting spam reporters: A reporter-based reputation system for email filtering. *ACM Trans. Inf. Syst.*, 27(1), 2008.

[14] P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.*, 5:17–61, 1960.

[15] E. Even-Dar, M. J. Kearns, and S. Suri. A network formation game for bipartite exchange economies. In *SODA*, pages 697–706, 2007.

[16] A. Fabrikant, A. Luthra, E. N. Maneva, C. H. Papadimitriou, and S. Shenker. On a network creation game. In *PODC*, pages 347–351, 2003.

[17] J. Leskovec, L. Backstorm, R. Kumar, and A. Tomkins Microscopic evolution of social networks. In *SIGKDD 2008*, pages 462–470.

[18] Y. Koren. Tutorial on recent progress in collaborative filtering. In *RecSys 2008.*

[19] N. Laoutaris, L. J. Poplawski, R. Rajaraman, R. Sundaram, and S.-H. Teng. Bounded budget connection games or how to make friends and influence people on a budget. *CoRR*, 2008.

[20] J.-G. Lee, J. Han, and X. Li. Trajectory outlier detection: A partition-and-detect framework. In *ICDE 2008*, pages 140–149.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *SIGKDD 2005*, pages 177–187

[22] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *SIGKDD 2008.*

[23] W. Michael and H. Mattord. *Principles of Information Secuirty.* Thomson, Canada.

[24] M. Mitzenmacher. Dynamic models for file sizes and double pareto distributions. 2002.

[25] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. In *CIKM 2006.*

[26] J. F. Nash. Non-cooperative games. *Annals of Mathematics*, 54, 286-295, 1951.

[27] M. E. J. Newman. Power laws, pareto distributions and zipf's law. May 2006.

[28] 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):179, 2007.

[29] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks. *PNAS*, 104(18):7332–7336, May 2007.

[30] W. Reed and M. Jorgensen. The double pareto-lognormal distribution a new parametric model for size distribution. 2004.

[31] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. In *SIGKDD 2008.*

[32] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.

[33] S. Wasserman and K. Faust. *Social Network Analysis.* Cambridge University Press, Cambridge, 1994.

[34] D. Watts. *Small Worlds:The Dynamics of Networks between Order and Randomness.* Princeton University Press, 1999.

[35] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 1998.

[36] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.