Statistical Behavior of Embeddedness and Communities of Overlapping Cliques in Online Social Networks

Ajay Sridharan
University of Victoria,
Email: ajays@uvic.ca

Yong Gao
University of British Columbia,
Email: yong.gao@ubc.ca

Kui Wu
University of Victoria,
Email: wkui@ieee.org

James Nastos
University of British Columbia,
Email: jnastos@interchange.ubc.ca

***Abstract*—Degree distribution of nodes, especially a power law degree distribution, has been regarded as one of the most significant structural characteristics of social and information networks. Node degree, however, only discloses the first-order structure of a network. Higher-order structures such as the edge embeddedness and the size of communities may play more important roles in many online social networks. In this paper, we provide empirical evidence on the existence of rich higher-order structural characteristics in online social networks, develop mathematical models to interpret and model these characteristics, and discuss their various applications in practice. In particular,**

1) **We show that the embeddedness distribution of social links in many social networks has interesting and rich behavior that cannot be captured by well-known network models. We also provide empirical results showing a clear correlation between the embeddedness distribution and the average number of messages communicated between pairs of social network nodes.**
2) **We formally prove that random k-tree, a recent model for complex networks, has a power law embeddedness distribution, and show empirically that the random k-tree model can be used to capture the rich behavior of higher-order structures we observed in real-world social networks.**
3) **Going beyond the embeddedness, we show that a variant of the random k-tree model can be used to capture the power law distribution of the size of communities of overlapping cliques discovered recently.**

I. Introduction

Social networks essentially represent the relationship between two social entities such as individuals or organizations. It indicates the way in which they are connected, ranging from casual acquaintance to personal connections. The advent of the Internet has revolutionized the way of communication amongst the common masses. One of the popular services that spawned out of this revolution is the On-line Social Network (OSN), evidenced by the huge success and popularity of On-line Social Network (OSN) websites such as Facebook, MySpace and Twitter, all having hundreds of millions of users. These OSNs not only provide their users with a convenient environment to interact easily with their friends, colleagues, relatives, and even "strangers" who share common interests, but also serve as a mirror of real social networks, making the study of social structure and interaction much easier than before. As a result, the study on the structural behavior of on-line social networks has triggered unprecedented interests in many research areas, including telecommunication networks, social science, and business to mention a few [1]. It is also one of the core research problems in the newly emerging scientific discipline, network science [2].

The power law distribution of node degree has been regarded as one of the most significant structural characteristics of social and information networks. In 1999, Barabási and Albert discovered that the degree distribution of the World Wide Web (WWW) follows a power law [3]. Since then, this structural behavior has been broadly investigated in many other types of real-world networks [1] and a large variety of generative random models have been proposed for it [4]. Nevertheless, node degree distribution alone only discloses the first-order structure of such networks. Higher-order structures such as the edge embeddedness, a notion used to capture the "degree" of a social tie with regard to the number of common neighbors, and the size of communities may play a more important role in information propagation and on-line social networking.

Practically, the degree of a social tie has already been utilized in various application contexts. As an incomplete list, Zhu et al. use the "cellular-social relationship graph" constructed based on traffic amount between cellular users in the design of effective patching strategy to prevent the propagation of computer worms over smart phones [5]; Ioannidis and Massoulie use the implicit tie between friends, i.e., the shared common interests, to develop personalized strategies for searching the web [6]; Wolf et al. use social network analysis on the communication structure of development teams, another type of social tie between network nodes, to predict software build failures [7].

Despite the diverse practical applications, existing work is mostly based on empirical studies over real-world dataset [7], [8], [9], [10]. While empirical studies are valuable, their pitfalls are obvious: the data collection is time consuming; the size of dataset is usually huge and very hard to handle; the cost of human resources on analyzing and processing the data is nontrivial. As mathematical models can greatly alleviate

the above problems, the call for new mathematical models that are powerful and flexible enough to capture the statistical behavior of OSNs, especially the higher-order statistics like the embeddedness, becomes unprecedentedly urgent.

While there are numerous mathematical models designed to model the structural behavior of complex networks [1], [3], [11], [12], to the best of our knowledge, there is currently no unified mathematical framework to design *generative models that are able to model the statistical behavior of higher-order structures such as the embeddedness or communities.* In this paper, we address the above challenge with the following contributions:

1) We show that in some real-world OSNs like Facebook, the distribution of edge embeddedness, a notion used to capture the "degree" of a social tie with regard to the number of common neighbors, has interesting and rich behavior that cannot be captured by well-known network models designed to model the observed power law node degree distribution in information networks. We also provide empirical evidence showing a clear correlation between a power law embeddedness distribution and the average number of messages communicated between pairs of social network nodes.
2) We prove formally that random k-tree, a recent model for complex networks [13], has a power law distribution of embeddedness. We show empirically that random k-trees can be used to model and interpret the statistical behavior of embeddedness we have observed in real-world social networks. To the best of our knowledge, this is the first existing model for which a power law distribution has been established mathematically for higher-order structural measures of a network other than the node degree.
3) Going beyond the embeddedness, we show that a variant of the random k-tree model generates random networks that capture well the power law distribution of the size of communities of overlapping cliques as has been discovered by Palla et al. [8] in 2005.

II. Background

A. Structural Properties of OSNs

An OSN is usually modeled as an undirected graph $G = G(V, E)$ where V denotes the set of nodes and E denotes the set of edges. A node represents an individual entity (e.g., a people or an organization) and an edge between two nodes signifies a social connection between the individual entities established according to some given criteria such as friendship or colleagues. In graph theory, a node is also called a vertex, and in the sequel we will use the two terms interchangeably. Let $e = \{u, v\}$ denote an edge between the two nodes u and v. The degree of a vertex v in a network G, denoted by $\deg_G(v)$, is the count of its neighbors. Among the many structural properties of OSNs [1], the distribution of vertex degrees of a network is probably the most well-known one that has been broadly studied before.

Definition 1. ***Power law distributions****. A power law distribution, as the name suggests, is a distribution function of the form $F(x) = 1 - x^{-\alpha+1}$ for some constant $\alpha > 1$. The corresponding density function is $f(x) = -cx^{-\alpha}$ for $c = -\alpha+1$. Note that α is also called the power law exponent.*

Power law distributions have long been used as a tool to model and explain empirical observations in a large variety of research fields. A distinct feature of a power law distribution is that it has a "heavy tail" as compared to other well-known distributions such as the normal distribution, the Poisson distribution and the exponential distribution. People are interested in such a distribution because it is scale free, i.e., scaling the variable x does not change the shape of the function.

The degree sequence of a network G is a sequence of integers $\{d_1, \cdots, d_n\}$ where d_i is the degree of the i-th vertex. The degree distribution of a network G is a sequence $\{X_1, \cdots, X_n\}$ where X_d is the proportion of vertices with degree d. It has been discovered that the degree distribution of the World Wide Web (WWW) and many other real-world networks follows a power law distribution [3], [1]. To interpret their empirical observations, Barabási and Albert proposed an evolving random model that is known as the *preferential attachment model* (also called **the BA model**). According to this model, a graph grows by adding one vertex at a time. In each step, the newly-added vertex is connected to m existing vertices selected according to the preferential attachment mechanism, i.e., an existing vertex is selected with probability in proportion to its degree.

Bollobás et al. [14] later provided a formal proof showing that the vertex degree distribution of the BA model obeys a power law distribution [14]. Since then, quite a few similar models have been proposed and studied [15] in order to design models where the scaling exponent of the power law vertex degree distribution can be controlled by some parameters. However, we note that *none* of these models are designed to capture structural characteristics other than the vertex degree distribution and clustering coefficients.

As the vertex degree distribution only provides statistics of the degree of individual vertices in a network, we regard it as the first-order structural property. In graph theory, it is well known that the structure of a network can be fully characterized by its vertex degree distribution only if the network belongs to some very special class of graphs. To get a more general picture about the structural features of an OSN, we may in many cases care more about the statistics of other higher-order structural properties such as the embeddedness and communities, which are the main focus of this paper.

Definition 2. ***(Edge) Embeddedness****. The* ***embeddedness*** *of an edge $e = \{u, v\}$ in a network $G(V, E)$, denoted by $\deg_G(e)$, is defined to be the number of common neighbors of u and v. For an edge $e = \{u, v\}$ with $\deg_G(e) = d$, the subnetwork consisting of the vertices u, v, and their d common neighbors is called a d-triangle.*

Embeddedness, also called edge embeddedness interchange-

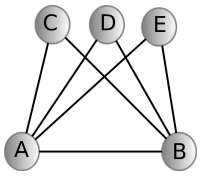


Fig. 1. Edge embeddedness of degree 3

ably, is an important measure of social networks [12]. Fig. 1 shows an example of a 3-triangle where the edge AB has embeddedness of degree 3. Edge embeddedness can be regarded as the "degree" of an edge. This is the reason why we use $\deg_G(e)$ to denote the embeddedness of an edge.

In addition to edge embeddedness, many OSNs typically include subgroups of nodes that are strongly connected to each other than to the rest of the network. Such subgroups are generally called communities, even though no unique definition is widely acceptable so far. As demonstrated in [8], statistics of communities and their overlapping disclose interesting social/natural relationships in many real networks. We follow the definition in [8].

Definition 3. ***Community of Overlapping Cliques****. A k-clique community in a network is defined as a union of all k-cliques (i.e., complete subgraph of size k) that can be reached from each other through a series of adjacent k-cliques, where adjacency means sharing $k-1$ nodes.*

B. Random k-Tree Model

In spite of the diverse applications of higher-order structural properties like embeddedness and community in OSNs, there are currently no generative mathematical models amenable for deriving the distribution of these higher-order structural properties. In [13], one of the authors of this paper introduced a random k-tree model and proved that the random k-tree model generates graphs with vertex degree distribution following a power law. To ease our further discussion, we introduce this model first.

Throughout this paper, a fully connected subgraph of k-vertices is defined to be a k-clique. All the graphs considered are undirected. The construction of a random k-tree is based on the following simple randomization of the recursive definition of k-trees. Starting with an initial k-clique $G^k(n)$, a sequence of graphs $\{G^k(n), n \geq k\}$ is constructed by adding vertices to the graph, one at a time. To construct $G^k(n+1)$ from $G^k(n)$, we add a new vertex v_{n+1} and connect it to the k vertices of a k-clique selected uniformly at random from all the k-cliques in $G^k(n)$. We call the graph process $\{G^k(n), n \geq k\}$ a *k-tree process*.

III. The Edge Embeddedness Distribution and Contact Strength of Social Links in OSNs

While the vertex degree distribution of real-world networks has been intensively studied since the seminal work of Barabási and Albert [3], we know of no previous work on the statistical behavior of the edge embeddedness. Our work has been motivated by the belief that an understanding on the statistical behavior of higher-order structures, such as the embeddedness, may help shed further light on the structure and the dynamics of OSNs. In this section, we report our empirical studies focusing on

1) the distribution of edge embeddedness in OSNS;
2) the impact of embeddedness distribution on social communication;

We will discuss the possible models that can capture the observed behavior of the edge embeddedness distribution in the next section.

A. Datasets

We studied the datasets collected from two real-world OSNs: Facebook and Orkut, which are hugely popular OSN services that helps their users to get connected to each other through constant interactions. Both OSNs have seen tremendous growth in the past few years. The size of these two OSNs is still growing, making the collection of complete data on these networks extremely hard, if not impossible. As a result, researchers resort to various methods for collecting a representative sample of these networks. The datasets for Facebook and Orkut considered in our empirical studies consist of representative samples and were made available to us by Mislove et al [9], [10].

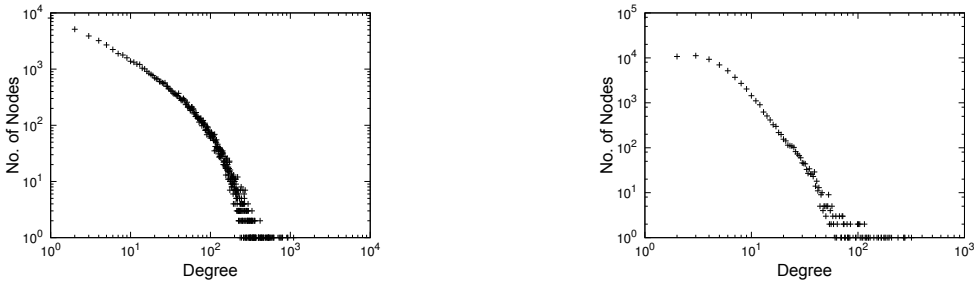
The Facebook dataset [10] consists of user links and their wall[1] posts from the New Orleans regional network. The dataset had a total of around 63k users with more than 800k user-to-user links and 870k wall posts amongst these users for a period of around 3 years. More details on the method of extracting the dataset can be obtained from [10].

The Orkut dataset [9] consists of more than 3 million users. The sheer size of the dataset as a whole makes it extremely hard and very time consuming to obtain its statistical results. Hence, we generated a sample of the network for our empirical study. Briefly, the sampling method, based on the Metropolis-Hasting Random Walk (MHRW) algorithm proposed in [16], selects an initial node, v, at random and then proceeds to select the next node, w, from the list of its neighbors with a probability of $min(1, \frac{deg(v)}{deg(w)})$, where min means the minimum value. More subtle details on the sampling method of MHRW could be found at [16]. We use this sampling technique since it results in an unbiased sample of the network when compared to the traditional Breadth First Search (BFS) and Random Walk (RW) techniques [16]. After sampling, we obtained a smaller Orkut dataset consisting of around 60k users with more than 175k user-to-user links.

B. The Vertex Degree Distribution

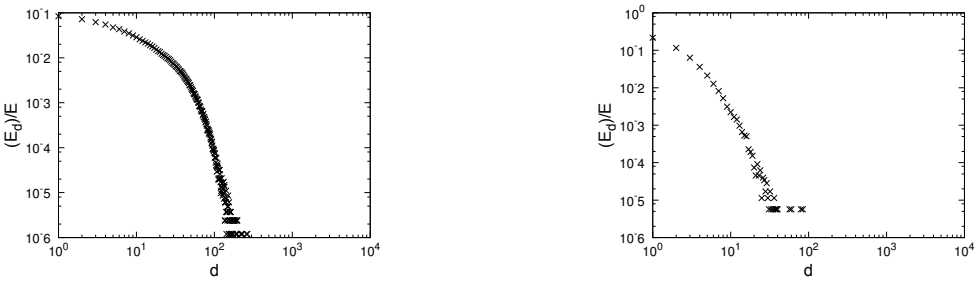
As many real-world networks have been empirically shown to have a power law vertex degree distribution, it is not a surprise for us to observe roughly a power law vertex degree distribution in the sampled Orkut dataset as shown in Fig. 2b.

[1]Wall is a space on every user's profile page that allows friends to post messages to the user.



(a) Facebook (b) Orkut

Fig. 2. Node degree distribution in Facebook and Orkut on a log-log scale. The x-axis represents the vertex degree and the y-axis shows the proportion of nodes having the corresponding node degree.



(a) Facebook (b) Orkut

Fig. 3. Embeddedness distribution in Facebook and Orkut on a log-log scale. The x-axis represents the degree of embeddedness and the y-axis shows the proportion of edges having the corresponding degree of embeddedness.

For the Facbook dataset, we observed that the vertex degree distribution can hardly be called power law as has been previously argued [16]. Instead, we can identify two regimes, roughly $[1, 40)$ and $[40, 1098]$, with each approximated by power law exponents 1.70 and 2.59, respectively (Fig. 2a). The similar multistage behavior has been observed before in [16], although the range of the two regimes and their corresponding exponents are different due to the different datasets used in the studies in [16].

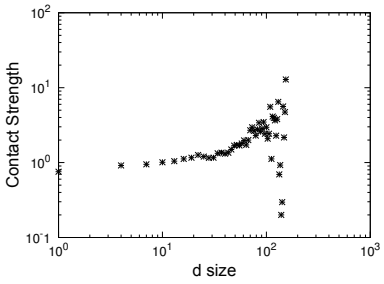


Fig. 4. Facebook communication pattern and edge embeddedness. The x-axis represents the degree of edge embeddedness and the y-axis represents the average contact strength of the edges with the corresponding embeddedness.

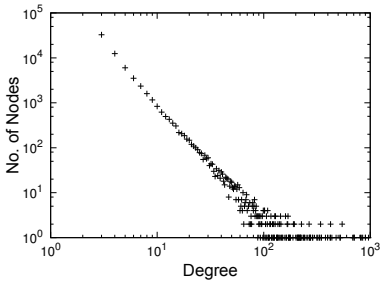
C. The Edge Embeddedness Distribution and Contact Strength of Social Ties

What intrigued us is that the edge embeddedness distribution in both datasets are found to have a behavior similar to that of their node degree distribution. In Figs. 3a and 3b, we plot the embeddedness distribution in the Facebook and Orkut datasets, respectively.

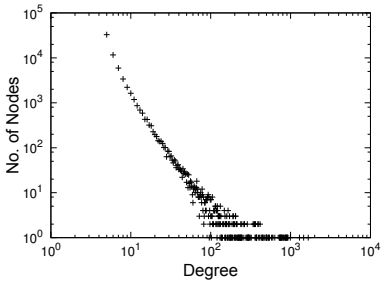
From these figures it can be observed that the sampled Orkut network tends to have a power law embeddedness distribution with the power law exponent of 2.91. The embeddedness distribution in the Facebook dataset does not follow a power law distribution. Similar to its node degree distribution, we can also roughly identify two regimes $[1, 50)$ and $[50, 265]$ with power law exponents around 1.69 and 3.50, respectively. We know of no previous work reporting the distribution of the edge embeddedness of a real-world OSN.

Another interesting observation obtained in our study is the correlation between the embeddedness of the edges and the contact strength of the social ties represented by the edges in the Facebook dataset. The **contact strength** of an edge is defined as the total number of wall posts posted by the two end nodes of the edge to each other's wall and can be regarded as a metric measuring the level of communications between the two end nodes.

In Fig. 4, we plot the contact strength of edges as a function of their embeddedness, on a *log-log scale*. The plot clearly shows that the contact strength increases with the increase of the degree of embeddedness. Even more interesting to note in Fig. 4 is that a two-stage pattern can also be observed of the contact strength with the boundary between the two stages coinciding well with the boundary between the two stages in the edge embeddedness distribution shown in Figure 3a.

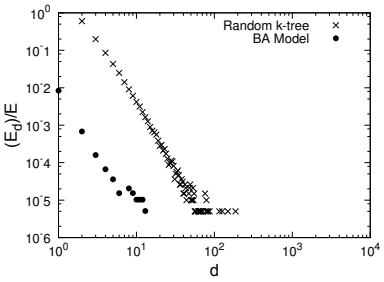


(a) Random k-tree, k=3

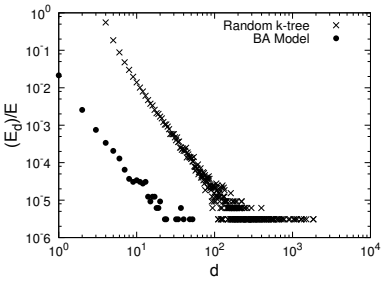


(b) Random k-tree, k=5

Fig. 5. Node degree distribution of random k-tree model. The x-axis represents the vertex degree and the y-axis shows the proportion of nodes having the corresponding node degree.



(a) Random k-tree ($k = 3$), BA model ($m = 3$)



(b) Random k-tree ($k = 8$), BA model ($m = 8$)

Fig. 6. Embeddedness distribution of random k-tree model. The x-axis represents the degree of embeddedness and the y-axis shows the proportion of edges having the corresponding degree of embeddedness.

It is well-known that a heavy-tailed node degree distribution has significant implications on the robustness of an information network. We believe that the behavior of the embeddedness distributions and its correlation with the contact strength of social ties as we have observed in the Facebook dataset are of great significance in many practical applications such as those in [5], [6].

IV. Modeling the Embeddedness Distribution with Random k-trees

While there have been numerous mathematical models in the literature, most notably the well-known BA model, designed to model the power law node degree distribution observed in real-world networks, none of them has been proved theoretically or shown empirically to have a power law embeddedness distribution.

In fact, our simulations show that in networks generated with the preferential-attachment-based models, the number of triangles is too low to draw any meaningful empirical observations on the edge embeddedness distribution. Refer to Fig. 6 for a comparison of the edge embeddedness distributions of networks generated from the BA model and the random k-tree model with the same edge density. We also note that none of these previous models can be used to model the multi-stage degree distributions and embeddedness distributions.

In an effort to search for a good generative model for the power law edge embeddedness distribution, we found out that the random k-tree not only has a power law degree distribution as established in [13], but also has a power law edge embeddedness distribution. A formal proof of the power law embeddedness distribution will be provided in the next section.

Fig. 5 shows a *log-log plot* of the node degree distribution in networks generated from the random k-tree model with different values of k. We see that the node degree distribution of random k-tree model follows a power law distribution, a fact that has been mathematically established in [13].

In Fig. 6, we compare the embeddedness distribution of networks generated from the random k-tree model and the BA model with the same edge density. From the figures, we can infer that the distribution of edge embeddedness in the random k-tree model follows a power law, which will further be proved in Section V of this paper. On the other hand, it is evident that the BA model fails to capture the richness of the edge embeddedness distribution simply because the number of triangles in the networks it generates is too low.

As has been discussed in the previous section, the degree distribution and the embeddedness distribution of Facebook have a behavior different from that of the random k-tree model. To understand the two-stage degree and embeddedness distributions observed in the Facebook dataset, we made the following assumption:

> *There are several types of users, each with a different social connection behavior. As the network evolves over time, when a user joins the network, he may create social ties to other users of a different type*

as well as social ties to users of its own type. While the vertex degree and the edge embeddedness of a given type of users viewed in isolation may have a power law behavior, it is the aggregate effect of users of different types that results in the observed two-stage (or multi-stage) power law distributions shown in Figs. 2a and 3a .

We have to emphasize that the validity of this assumption, just as the assumption made when Barabási and Albert [3] proposed their preferential attachment model, needs to be further verified in a variety of OSNs and we leave it as an interesting future work.

Based on the above assumption, we propose the following mixed random k-tree model to model the phenomena:

Definition 4. ***Mixed random k-tree model*** *is a variant of the random k-tree model by mixing different k values in the k-tree process. Formally, given two integers $k_1, k_2 (k_1 < k_2)$ and starting with an initial k_2-clique $G^{k_2}(n)$, we construct a sequence of graphs $\{G^{k_i}(n), n \geq k_i\}$ by adding vertices to the graph one at a time, where k_i is a randomly chosen integer in the range of $[k_1, k_2]$ with a pre-defined probability $p_i (\sum_{i=k_1}^{k_2} p_i = 1)$. When a new vertex v_{n+1} is added, it is connected to the k_i vertices of a k_i-clique selected uniformly at random from all the k_i-cliques in the previous graph.*

The intuition is that in the mixed random k-tree model, all nodes assigned to the same value of k when joining the network are of the same user type. When viewed in isolation, vertices of the same type evolve in exactly the same way as they do in a pure random k-tree model. By allowing a vertex of a given type to be adjacent to vertices of another type (and thus contribute to the degree and embeddedness of vertices of that type), an overall multi-stage power law distribution emerges.

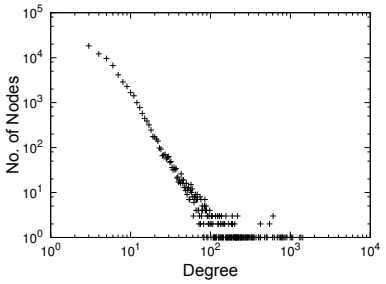


Fig. 7. Node degree distribution of mixed random k-tree, $k = 3$ to 12.

Figs. 7 and 8 show the node degree distribution and the embeddedness distribution, respectively, in the graphs generated with the mixed random k-tree model, using parameters $k = 3$ to 12 (i.e., $k_1 = 3, k_2 = 12$) and preset probabilities of $0.30, 0.20, 0.16, 0.11, 0.06, 0.05, 0.04, 0.03, 0.03, 0.02$, respectively. From the figures, the mixed random k-tree model can generate graphs having multi-stage statistics similar to that of Facebook. We note that by adjusting the parameters, we could obtain different multi-stage statistics. We leave the question of how to choose the parameters to fit the behavior of a particular real-world network as an interesting future work.

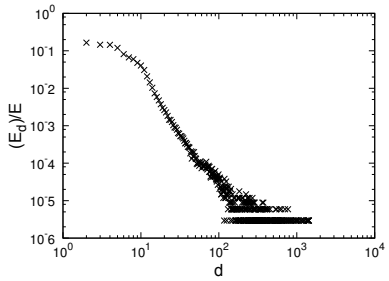


Fig. 8. Embeddedness distribution of mixed random k-tree, $k = 3$ to 12. The x-axis represents the degree of embeddedness and the y-axis shows the proportion of edges having the corresponding degree of embeddedness.

We remark that other random models, such as the BA model and its variants, can also be modified to define a mixed random model to capture the phenomenon of a multi-stage node degree distribution. But our simulations indicate that these mixed variants of the BA model fail to capture the rich statistical behavior of the edge embeddedness in OSNs simply because of the extremely low number of triangles in the networks they generate.

V. Power Law Distribution of Embeddedness of Random k-Trees

In this section, we prove the following theorem to establish the power law distribution of the edge embeddedness of a random k-tree.

Theorem 1. *Assume that $k > 2$. In the random k-tree process $\{G^k(n), n \geq k\}$, the proportion of the edges e with edge embeddedness $deg_{G^k(n)}(e) = d$ has the following power law distribution with high probability*[2]*:*

$$d^{-(1+\frac{k}{k-2})}. \tag{V.1}$$

To begin with, we first consider the number of k-cliques containing a particular edge. Let $e = \{u, v\}$ be an edge and assume w.l.o.g that u is added before v, i.e., the edge e is "born" when v is added to $G^k(n)$. We have the following observation:

Lemma 1. *Let $k > 2$ and c_e^* be the number of k-cliques that contain the edge e. Then,*

$$c_e^* = \binom{k-1}{k-2} + \binom{k-2}{k-3}(deg_{G^k(n)}(e) - (k-1)).$$

Proof: When e is created as a result of adding the vertex v, exactly $\binom{k-1}{k-2} = k-1$ k-cliques containing e are created. For each vertex w added after v is in the network, new k-cliques containing e are created if and only if w is made adjacent to both u and v (and consequently, $\deg_{G^k(n)}(e)$ increased by 1.

[2]In the theory of random graphs, by "with high probability" we mean that the probability of an event tends to 1 as the size of the graph tends to infinity. All the existing mathematical results on the power law degree distribution of models for complex networks are established in this form.

If this occurs, exactly $\binom{k-2}{k-3} = k-2$ new k-cliques are created that contain the edge e.

Note that the edge embeddedness of e is initially $k-1$ when e is created. This is because when v is added to the graph, it is made adjacent to a k-clique that contains u. Therefore, the number of newly-added vertices that form a triangle with e is equal to $\deg_{G^k(n)}(e) - (k-1)$. The lemma follows. ∎

Let $\mathcal{C}_n$ be the number of k-cliques in $G^k(n)$. Since every time a new vertex is added to the network, exactly k different new k-cliques are created, the number of k-cliques in $G^k(n)$ is $(n-k)k+1$, i.e., $\mathcal{C}_n = (n-k)k+1$. It follows from Lemma 1 that given $G^k(n)$, the probability for the new vertex v_{n+1} to form a triangle with the two endpoints of an edge $e = \{u, v\}$ is

$$\begin{aligned}
&\mathbb{P}\{u \text{ and } v \text{ are adjacent to } v_{n+1}\} = \frac{c_e^*}{\mathcal{C}_n} \\
&= \frac{(k-1) + (k-2)(\deg_{G^k(n)}(e) - (k-1))}{(n-k)k+1} \\
&= \frac{(k-2)\deg_{G^k(n)}(e) - b_k}{c_k n} \qquad \text{(V.2)}
\end{aligned}$$

where $b_k = (k-1)(k-3)$ and $c_k = k - \frac{k^2-1}{n}$. Note that in addition to the constants, the above conditional probability only depends on $\deg_{G^k(n)}(e)$. With these preparations, we are now ready to prove Theorem 1.

Proof of Theorem 1: Let $T_d(n)$ be the number of edges e with edge embeddedness $\deg_{G^k(n)}(e) = d$. We now derive a system of recursive equations for the expectation $\mathbb{E}[T_d(n)]$ of $T_d(n)$. We will focus on the case of $d > k-1$. The case of $d = k-1$ is similar.

Let $I_d(e, n)$ be the indicator function for the event that the embeddedness of an edge e is d, i.e.,

$$I_d(e,n) = \begin{cases} 1, & \deg_{G^k(n)}(e) = d \\ 0, & \text{otherwise.} \end{cases}$$

Then $T_d(n)$ is the sum of $I(e,n)$'s over all the edges, i.e., $T_d(n) = \sum_{e \in G^n(k)} I_d(e,n)$. By the definition and properties of conditional expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[T_d(n+1) \mid G^k(n), n \geq k\right] \\
&= \mathbb{E}\left[\sum_{e \in G^k(n)} I_d(e, n+1) \mid G^k(n), n \geq k\right] \\
&= \sum_{e \in G^k(n)} \mathbb{E}\left[I_d(e, n+1) \mid G^k(n), n \geq k\right]
\end{aligned}$$

For a particular edge e, we see that $\deg_{G^k(n+1)}(e) = d$ if and only if either one of the following two cases occurs:

1) $\deg_{G^k(n)}(e) = d$ and v_{n+1} does not form a triangle with e;
2) $\deg_{G^k(n)}(e) = d-1$ and v_{n+1} forms a triangle with e.

So, if we write $f_k^d(n) = \frac{(k-2)d - b_k}{c_k n}$, we have from Equation (V.2) that

$$\begin{aligned}
&\mathbb{E}\left[I_d(e, n+1) \mid G^k(n), n \geq k\right] \\
&= f_k^{d-1}(n) I_{d-1}(e,n) + (1 - f_k^d(n)) I_d(e,n).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[T_d(n+1) \mid G^k(n), n \geq k\right] \\
&= f_k^{d-1}(n) T_{d-1}(n) + (1 - f_k^d(n)) T_d(n).
\end{aligned}$$

Taking unconditional expectation on both sides in the above equation, we have from the properties of conditional expectation that

$$\mathbb{E}[T_d(n+1)] = f_k^{d-1}(n)\mathbb{E}[T_{d-1}(n)] + (1 - f_k^d(n))\mathbb{E}[T_d(n)]. \qquad \text{(V.3)}$$

Based on the above recursion, it can be proved by induction that for any small constant $\epsilon > 0$, there exists a constant n_ϵ such that for all $n > n_\epsilon$, we have

$$\mathbb{E}[T_d(n)] = \beta_d n + \epsilon,$$

namely, $\mathbb{E}[T_d(n)] = \beta_d n + O(1)$, where β_d satisfies the following simple equation

$$\beta_d = \frac{a_k(d-1) - b_k}{a_k d - b_k + c_k} \beta_{d-1}.$$

where $a_k = k-2$, $c_k = k$, and $b_k = (k-1)(k-3)$. The unique solution for the simple recursive equation for β_d is

$$\beta_d = \prod_{i=k-1}^{d} \frac{a_k(i-1) - b_k}{a_k i - b_k + k} = \frac{\Gamma(d - \frac{b_k}{a_k})}{\Gamma(d - \frac{b_k}{a_k} + \frac{k}{a_k} + 1)}.$$

By Stirling's approximation for the Gamma function, β_d is asymptotically equivalent to $d^{-(1+\frac{k}{k-2})}$.

Since there are kn edges in a random k-tree, we see that the average proportion $\frac{1}{kn}\mathbb{E}[T_d(n)]$ of the number of edges with embeddedness d is asymptotically $d^{-(1+\frac{k}{k-2})}$. By applying Azuma's Inequality [17], it can be shown that the proportion $\frac{1}{kn}T_d(n)$ of the edges with embeddedness d is equivalent to $d^{-(1+\frac{k}{k-2})}$ with high probability. We omit the details here due to space limit. However, a similar argument for the case of power law node degree distribution could be found in [13]. This completes the proof of Theorem 1. ∎

For random 2-trees, we have the following theorem showing that its edge embeddedness follows an exponential law. Its proof is omitted due to page limit.

Theorem 2. *The distribution of the edge embeddedness of a random 2-tree follows the exponential law 3^{-d}.*

Finally, we note that Theorem 1 can be generalized to the case of "embeddedness" of small-sized cliques. Let $\deg_G(C)$ be the number of common neighbors of the vertices in a clique C. We have the following theorem whose proof is omitted to save space.

Theorem 3. *For the random k-tree $G^k(n)$, the proportion of h-cliques C with $deg_{G^k(n)}(C) = d$ follows the power law distribution $d^{-(1+\frac{k}{k-h})}$. where $h < k$ is a constant.*

VI. Random Partial k-Trees and Size distribution of k-Clique Communities

Besides edge embeddedness, communities and their structures have drawn much interest in recent years. There is also a continuing effort to find better definitions for a network community [8], [18], [19]. Recently, Palla et al. [8] introduced the notion of a k-clique community. As defined in Section II, a k-clique community is a collection of k-cliques where every pair of k-cliques can be reached from each other through a sequence of k-cliques that share $k-1$ vertices. One of the intriguing findings in the study of Palla et al. [8] is that the size distribution of the k-clique communities follows a power law in several real-world networks such as the co-authorship networks, word-association networks, and the protein interaction networks.

Toivonen et al. [20] refer to k-clique communities as *clusters*. They compare a number of random network models for their parameter values of some higher-order structures under the umbrella of generating random networks that match the number of nodes and edges to real-world network examples.

In this section, we show that a simple variant of the random k-tree model is able to capture the characteristic of the community structure much better than other existing models such as the BA model.

Definition 5. *(Random Partial k-Trees) A partial k-tree is a subgraph of a k-tree. A random partial k-tree $G^k(n,r)$ is a graph obtained by removing uniformly at random r edges in a random k-tree $G^k(n)$.*

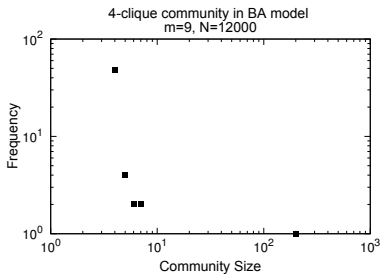


Fig. 9. k-clique communities in graphs created with the BA model

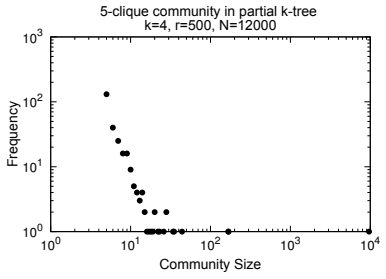


Fig. 10. Power law community size in partial k-trees

We use the *clique percolation method* of Palla et al. [8] to find the k-clique community sizes in a number of randomly generated networks. We analyze the $(k+1)$-clique community sizes of $G^k(n,r)$ with various values for $r>0$. We compare BA model networks and random partial k-trees with identical densities: that is, we set the parameter m in the BA model to the parameter k in partial k-trees and keep r relatively small.

Toivonen et al.'s study [20] uses two datasets: the *lastfm* network (www.last.fm) of edge density 4.2 and an email communication network of edge density 9.6. Their experiments indicate that some variants of the BA model are able to model the size distribution 4-clique communities in the *lastfm* while all the models they considered have difficulties in creating a sufficiently large number of 5-clique communities [20].

Our experimental results confirm the findings of Toivonen et al. As depicted in Fig. 9, the network generated by the BA model with similar edge density ($m=9$) revealed no power law size distribution of 4-clique communities. In order to find any nontrivial 5-clique community structure, we had to adjust m to 10 or higher. Nevertheless, even increasing the density to $m=9$ yields no significant 5-clique community structure.

On the other hand, as shown in Fig 10, networks generated by the random partial 4-tree model reveal a clear power law size distribution of 5-clique communities. Further experiments on our random partial k-tree model $G^k(n,r)$ with $k=9$ and $n=2000$ shows that the random partial k-tree model can generate networks with non-trivial s-clique communities with s up to 9 and has great potentials to model real-world datasets with higher edge density such as the email dataset used by Toivonen et al.'s study [20]. We omit these experiment results due to the page limit.

In conclusion, we observe that the partial k-tree model produces community distributions similar to those observed in real-world networks and such community shapes can be tuned with the parameter r. On the other hand, the BA model networks do not reveal any nontrivial community structure unless the density measure is increased beyond a reasonable threshold.

VII. Related Work

Research in OSNs has been steadily increasing in the past decade. Analysis of huge on-line social networks [21], [22] shows that some networks have a scale-free behavior [3] and also exhibit small-world properties [23]. There have been numerous mathematical models proposed to model the power law node degree distribution in real-world networks. The well-known one is probably the BA model [3], which stimulates many similar preferential-attachment-based methods [4]. Nevertheless, none of them has been proved theoretically or shown empirically to have a power law embeddedness distribution.

In the traditional social network literature, the so-called exponential random graph model is also well-known. An exponential random graph model is specified by a distribution from an exponential family of distributions over the space of all networks [11], [12]. An exponential random graph model does have model parameters for higher-order structures and it is possible to use some sophisticated statistical technique to estimate these parameters. Nevertheless, the exponential

random model has its own deficiency. First, the model is not generative; in fact, generating a random sample from the distribution is a highly non-trivial task. Second, mathematical results have been established showing that for many parameter settings, as the generated network gets larger, the exponential random graph model degenerates and becomes trivial in the sense that it only produces the complete network containing all possible links or networks similar to those generated from the pure Erdös-Rényi random graphs ([11], [12]).

In [13], Gao proposed the random k-tree model. The work in [13], however, only focuses on the power law distribution of node degree distribution of random k-trees. The higher-order statistics such as edge embeddedness and community size has not been touched.

There has been a lot of work in the literature concerning the presence and the significance of edge embeddedness. Embeddedness has been utilized in many applications [5], [6], [7]. It has also been employed to defend against attacks in distributed systems [24] and to detect and prevent email spam [25]. SPROUT [26] is a DHT routing algorithm that uses the embeddedness in an OSN to find reliable routes. Nevertheless, the above works are intended to better utilize the embeddedness in practice. Our paper lays a solid theoretical foundation for the above work, with which distribution of embeddedness can be modeled and mathematically analyzed.

VIII. Conclusion

In this paper, we showed that real-world OSNs have rich higher-order structural statistics of practical significance that cannot be modeled by well-known generative models such as the BA model. We studied a newly proposed random k-tree model and its different variants, and showed by both theoretical analysis and empirical simulations that these k-tree based models can be used to model not only the node degree distribution but also higher-order statistics such as the embeddedness and the size of communities of OSNs. The random k-tree model and its variants can be used to easily generate large graphs with statistical features similar to real-world OSNs such as Facebook and Orkut, and more importantly its unique structure lends itself to amenable mathematical analysis for higher-order statistics. To the best of our knowledge, this paper is the first that has proved the power law distribution of embeddedness with random k-tree model.

As the norm of mathematical modeling by George Box [27], "all models are wrong, but some are useful." While we should not expect that the random k-tree model and its variants capture all statistical features of OSNs, we believe, with empirical study and rigorous mathematical proof, that they are extremely useful in the study and utilization of OSNs.

References

[1] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[2] T. G. Lewis, *Network Science: Theory and Applications*. Wiley & Sons, 2009.

[3] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, p. 509, 1999.

[4] R. Durrett, *Random Graph Dynamics*. Cambridge University Press, 2007.

[5] Z. Zhu, G. Cao, S. Zhu, S. Ranjan, and A. Nucci, "A social network based patching scheme for worm containment in cellular networks," in *Proceedings of IEEE Infocom*, April 2009.

[6] S. Ioannidis and L. Massoulie, "Surfing the blogosphere: Optimal personalized strategies for searching the web," in *Proceedings of IEEE Infocom*, March 2010.

[7] T. Wolf, A. Schrter, T. Nguyen, and D. Damian, "Predicting build failures using social network analysis on developer communication," in *Proceedings of the 31st International Conference on Software Engineering*, May 2009.

[8] G. Palla, I. Dernyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435(7043), pp. 814–818, June 2005.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

[10] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

[11] T. Jonasson, "The random triangle model," *Journal of Applied Probability*, vol. 36, no. 3, pp. 852–867, 1999.

[12] T. Snijders, P. Pattison, G. Robins, and M. Handcock, "New specifications for exponential random graph models," *Sociaological Methodology*, vol. 36, no. 1, pp. 99–153, 2006.

[13] Y. Gao, "The degree distribution of random k-trees," *Theoretical Computer Science*, vol. 410(8-10), pp. 688–695, 2009.

[14] B. Bollobas, O. Riordan, J. Spencer, and G. Tusnady, "The degree sequence of a scale-free random graph process," *Random Structures and Algorithms*, vol. 18, pp. 279–290, 2001.

[15] C. Cooper and A. Frieze, "A general model of web graphs," *Random Structures and Algorithms*, vol. 22, pp. 311–335, 2003.

[16] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *Proceedings of IEEE INFOCOM 2010*. IEEE, March 2010, pp. 1–9.

[17] M. Michael and U. Eli, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[18] L. Falzon, "Determining groups from the clique structure in large social networks," *Social Networks*, vol. 22, pp. 159–172, 2000.

[19] N. Mishra, R. Schreiber, I. Stanton, and R. Tarjan, "Finding strongly knit clusters in social networks," *Internet Mathematics*, vol. 5, no. 1-2, pp. 155–174, 2008.

[20] R. Toivonen, L. Kovanena, M. Kiveläa, J. Onnelab, J. Saramkia, and K. Kaskia, "A comparative study of social network models: Network evolution models and node attribute models," *Social Networks*, vol. 31, no. 4, pp. 240–254, 2009.

[21] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *EuroSys '09: Proceedings of the 4th ACM European conference on Computer systems*. New York, NY, USA: ACM, 2009, pp. 205–218.

[22] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 835–844.

[23] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, june 1998.

[24] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks," in *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM, 2006, pp. 267–278.

[25] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazières, and H. Yu, "Re: reliable email," in *NSDI'06: Proceedings of the 3rd conference on Networked Systems Design & Implementation*. Berkeley, CA, USA: USENIX Association, 2006, pp. 22–22.

[26] S. Marti, P. Ganesan, and H. Garcia-Molina, "Sprout: P2p routing with social networks," in *First International Workshop on Peer-to-Peer Computing and Databases (P2P&DB 2004)*, March 2004.

[27] G. Box and N. R. Draper, *Empirical Model-Building and Response Surfaces*. Wiley & Sons, 1987.