Intergroup networks as random threshold graphs

Sudipta Saha, Niloy Ganguly, and Animesh Mukherjee
Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India

Tyll Krueger
Department of Computer Science and Engineering, Technical University of Wroclaw, Poland
(Received 9 February 2014; published 24 April 2014)

Similar-minded people tend to form social groups. Due to pluralistic homophily as well as a sort of heterophily, people also participate in a wide variety of groups. Thus, these groups generally overlap with each other; an overlap between two groups can be characterized by the number of common members. These common members can play a crucial role in the transmission of information between the groups. As a step towards understanding the information dissemination, we perceive the system as a pruned intergroup network and show that it maps to a very basic graph theoretic concept known as a *threshold graph*. We analyze several structural properties of this network such as degree distribution, largest component size, edge density, and local clustering coefficient. We compare the theoretical predictions with the results obtained from several online social networks (LiveJournal, Flickr, YouTube) and find a good match.

DOI: 10.1103/PhysRevE.89.042812 PACS number(s): 89.75.Fb, 89.90.+n

I. INTRODUCTION

Group formation [1–3] is a very common and popular feature among humans where a user (human) can participate in multiple groups [4]. Consequently, many social networking sites like LiveJournal,[1] Flickr,[2] YouTube,[3] etc., provide explicit facilities to form, maintain, and communicate within social groups [5]. These groups can be deemed as a medium of mass communication among its participating users [6–9]; there is, however, an interesting side effect to this communication. Common users belonging to multiple groups pass information of one group to another. Hence, analyzing the extent of connectivity among the groups can shed light into the amount of information propagated from one group to another. This connectivity structure can be best estimated by analyzing the properties of an evolving intergroup network, which is the primary focus of this paper.

Intergroup networks can be modeled as the one-mode projection of evolving *user-group* bipartite networks. In the bipartite process, the user partition grows with time, while, if we consider only the popular groups, the group partition remains fixed. Such bipartite networks where one partition remains fixed is termed a alphabetic-bipartite network (α-BiN) [10]. A projection on the groups allows us to obtain the group-group network (i.e., the intergroup network) where two groups are neighbors if they have at least one common user. However, it can be safely assumed that two groups will have high mutual interaction, thus allowing more information to propagate if there are at least a critical number of common users (determined by a threshold value) [5,11]. Therefore, an interesting structure to study is the "*pruned intergroup network*" where two nodes (groups) are connected if they have more than a threshold number of common users.

The first attempt to understand the "*pruned intergroup network*" was made in Ref. [5], where by mapping the underlying evolution dynamics to a Polya urn model, the degree distribution of the pruned intergroup network is derived. However, in order to gain a better insight into the connectivity structure, the more relevant structural properties such as largest component size, edge density, and local clustering coefficient (assuming a large number of groups) under various possible threshold values need to be analyzed as a first step. In this paper, we identify the mathematical relationship between the weight of an edge and the weights of the associated nodes (later these node weights are referred to as "*attractiveness parameters*"), derived in Ref. [5], as special importance. We heavily leverage on this relationship to derive the formula for the above mentioned structural properties of the "*pruned intergroup network*." We also show that this class of networks can be appropriately modeled by a special variant of "*random threshold graph*" [12], termed a "*multiplicative random threshold graph*." We compare the theoretical predictions with the same obtained from the available real datasets, and in most of the cases, we obtain a significantly accurate match.

Hence, the main contributions of this paper are twofold: (a) we show how the intergroup relationships in social systems and in a more general sense the pruned one-mode projection of a preferentially grown α-BiN can be studied as a special kind of random threshold graph model, and (b) we show how the mathematical analysis of various structural aspects (degree distribution, largest component, edge density, as well as local clustering coefficient) of this specific kind of multiplicative random threshold graphs can be done in a more transparent way.

The rest of the paper is organized as follows. In the next section, we precisely describe the basic model of α-BiN and the special property of the intergroup networks that allows us to map it to multiplicative random threshold graphs. In Sec. III we present a detailed description of the mathematical analysis of degree distribution, edge density, largest component, and the local clustering coefficient. Next, in Sec. IV we compare the mathematical findings with the observations made from the real dataset. Finally, in Sec. V we present a brief review of the state-of-the-art before drawing the conclusion.

[1]LiveJournal: www.livejournal.com
[2]Flickr: www.flickr.com
[3]YouTube: www.youtube.com

II. INTERGROUP NETWORK AND RANDOM THRESHOLD GRAPH

In the following we first very briefly describe the α-BiN model and the necessary constructions for the user-group system. Next we discuss the mapping of the class of a pruned intergroup network to a more basic and a simple graph theoretic concept called a "random threshold graph." This mapping allows us to use certain fundamental properties of a "random threshold graph" to characterize intergroup networks.

A. User group and α-BiN and its projections

User-group system as α-BiN: An α-BiN [10,13,14] is a special kind of bipartite network $[G = (V,U,E)]$, where one partition (V, containing the active members, or the top set, $|V| = t$) grows proportionally to time, whereas the other partition (U, containing the passive members, or the bottom set, $|U| = n$) remains fixed in size. The set of edges (E) represents the interactions between the elements of the active partition and the passive partition (see Fig. 1). For example, in Ref. [5] the authors model the user-group online social systems as an α-BiN where the users are considered to be the active members and the social groups the passive members.

Evolution: In order to model the evolution of α-BiN, it is assumed that at every time step one node joins the set V and creates some connections (edges of α-BiN) with the nodes in set U. The number of edges created by the nodes in V is considered to be a random variable having a certain probability distribution [denoted by $F(u)$]. This distribution (which is effectively the degree distribution of the top set V) is assumed to have known *finite* first moment $E(X) = \mu$ and *finite* second moment $E(X^2) = \mu'$. Reference [5] shows that many interesting aspects of the dynamics of the evolution depends on only μ and μ', and hence, no specific form of this distribution is assumed in the analysis. The probability that an edge created by a node $v \in V$ at time t attaches to node $u \in U$ is proportional to the current degree of the node u has the following equation:

$$\Pr\{d_t(u) = d_{t-1}(u) + 1\} = \frac{d_{t-1}(u)}{\sum_{x \in U} d_{t-1}(x)}. \qquad (1)$$

This attachment strategy has been formally studied in detail in Ref. [15].

Intergroup network: With the evolution of the user-group systems, the relationship among the groups also changes. This effect is generally captured by taking the one-mode projection of the two-mode user-group α-BiN. We term this one-mode projection as the intergroup network, which is a weighted graph where the nodes are the social groups and there is an edge between two such groups with weight w to indicate that there are w common users for the two groups. The weight of an edge fundamentally reflects the strength of the relationship between two groups; in other words, it is an indicator of how much information can be propagated between a pair of groups.

Pruned intergroup network: The node set in this network is the same as that in the intergroup network. A pruned intergroup network is associated with a threshold value Δ. An edge is put between two groups if and only if the weight of the edge in the intergroup network is above or equal to Δ. Thus, fundamentally this construction is the pruned one-mode projection of the user group α-BiN.

Special property of α-BiN: In Ref. [16] the authors show that a single realization of the entire evolution of an α-BiN through preferential attachment can be obtained as follows: (a) first, a parameter (called an attractiveness parameter), sampled from a certain distribution (Dirichlet), is preassigned with each of the nodes in the set U of the α-BiN (denoted by ψ_u for node u); (b) next, the probability that a node $v \in V$ joins with node $u \in U$ is considered to be following a Bernoulli distribution with success rate ψ_u. The joint distributions of these random variables $(\psi_1, \psi_2, \ldots, \psi_n)$ were shown to be following a Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_n)$, α_i denoting the initial degree of the node i of set U in the α-BiN.[4] It should be noted that this initial configuration of the α-BiN, specifically the initial degrees of the nodes of set

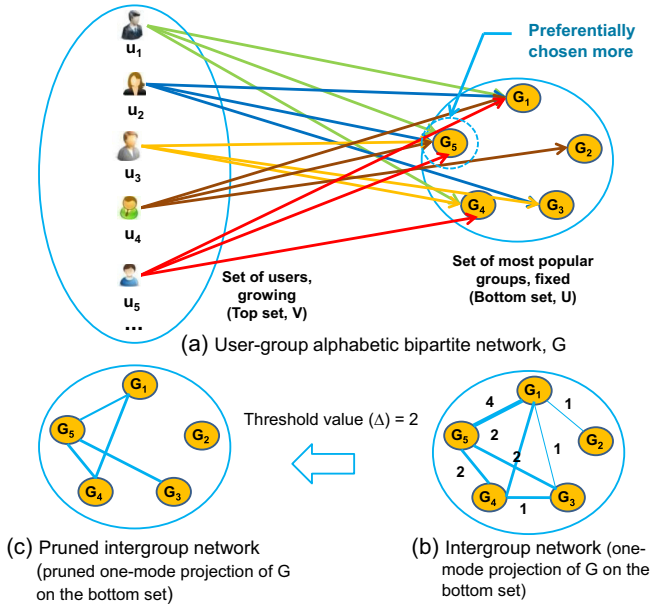

FIG. 1. (Color online) (a) An α-Bin, G, representing a user-group bipartite system with $n = |U| = 5$ social groups $(G_1, G_2, \ldots, G_5)$ and $t = |V| = 5$ users $(u_1, \ldots, u_2, \ldots, u_5)$, (b) intergroup network derived from G as the weighted one-mode projection of G on its bottom set, and (c) the pruned intergroup network as the pruned bottom projection of G on its bottom set G^T at time $t = 5$, i.e., when five users have entered set V (we assume, at each time step one user joins the system). The users u_1 to u_5 each has created three connections with the groups in U. The weight of an edge in the intergroup network tells the number of common users that have connected with both the end groups or in other words the number of paths between the two groups in set U of G.

[4]The attractiveness parameter ψ_i associated with a node i $(\in U)$ precisely defines the probability that a node j $(\in V)$ will make its next connection with node i. Hence, in a preferential attachment-based evolution process the attractiveness parameters associated with the nodes change with time. However, Ref. [20] shows that the mathematical analysis of such an evolution process can be simplified if the attractiveness parameters are sampled from a specific probability distribution and are preassigned to the nodes in U.

U, are an important component of the model. Reference [16] analyzes the special case where $\forall i, \alpha_i = 1$, i.e., the degrees of all nodes in set U of the α-BiN are 1. In this case the marginal distribution of the attractiveness value of a certain node i follows a beta distribution with parameters $(1, n-1)$. Using this modeling approach, Ref. [5] derives an important property regarding the relationship between the attractiveness values of two nodes and the weight of the edge between them in the one-mode projection of the α-BiN under the assumption of asymptotic growth in the number of elements in the top set (i.e., $t \to \infty$). Equation (2) describes this relationship:

$$\lim_{t\to\infty} \frac{W(i,j)}{t} = (\mu' - \mu)\psi_i \psi_j, \tag{2}$$

where ψ_i and ψ_j are the values of the attractiveness associated with nodes i and j, respectively; $W(i,j)$ is the weight of the edge between nodes i and j, and μ and μ' are the first and second moments of the distribution of the number of edges created by the nodes entering the top set V.

Threshold graph and random threshold graph: The notion of "*threshold graph*" [17,18] is defined as follows:

A graph G is called a threshold graph if there exists a set of weights X_i for all nodes i in the graph and a real value Δ (called the threshold value) such that there exists an edge between any pair of nodes (i,j) in the graph if and only if $X_i + X_j \geqslant \Delta$.

However, there are many other equivalent definitions of a threshold graph based on their structural properties [17,18].

A "*random threshold graph*" [12,19] is also a threshold graph where the node weights are random variables sampled from a certain distribution.

Pruned intergroup network as random threshold graph: From the special property as presented in the Eq. (2), it can be understood that, there exists an edge in a pruned intergroup network, if and only if the product of the attractiveness values of the two nodes is above a certain threshold [as $(\mu' - \mu)$ is constant for a given system]. Sinc a, product can be transformed to a sum by taking a logarithm on both sides, the class of pruned intergroup networks can be thought of as random threshold graphs.[5] We call this special variant of the random threshold graph the "*multiplicative random threshold graph*." In this work we assume an implicit transformation of the threshold values and hence, work with the multiplicative representation itself.

In the following we present two important properties of threshold graphs which directly follow from the definition. We use these properties extensively in the derivation of the theoretical results in the next section:

(1) *Component structure*: In any threshold graph, for any given threshold value there exists only one connected component of size larger than or equal to one.

(2) *Embedded star graph*: For any threshold value the largest connected component has at least one star node, that is, a node which is connected with all the other nodes within the connected component [12].

Random threshold graph model: We assume that each node i of the intergroup network is associated with an attractiveness parameter θ_i. Using the relationship between the weight of an edge and the attractiveness parameters of the nodes at its two ends [as depicted in Eq. (2)], we define the existence of an edge in a multiplicative random threshold graph with node weights θ_i as follows: *There exists an edge between two nodes if and only if the product of the attractiveness parameters of the nodes is larger than or equal to a certain threshold value Δ (i.e., there is an edge between node i and node j, if and only if $\theta_i \cdot \theta_j \geqslant \Delta$).*

It is to be noted that for the sake of simplicity we do not consider the factor t as well as $(\mu' - \mu)$ in the analysis. Hence, when we compare the mathematical results with the results obtained from real networks in Sec. IV, we require an adjustment of the edge weights.

Work has been done in the past to analyze the evolution of bipartite networks where both the partitions grow in size simultaneously [20]. However, in the current work we assume that only the top set undergoes an unbounded growth (i.e., $t \to \infty$), while the bottom set remains fixed. In addition, we also assume that the bottom set has a large number of nodes (i.e., $n \to \infty$) from the beginning of the evolution. The theoretical results presented in this paper have to be understood in the following way: the limit $t \to \infty$ is applied first, and then the limit $n \to \infty$ is applied. However, it is important to note that the limits cannot be exchanged. For a specific setup it implies that we must have the value of t much larger than n to meet the basic requirement of α-BiN. We simulated the evolution of α-BiN for different values of n. For each different n we considered the value of t to be $100\times n$ and find already a very accurate match between the results obtained from theory (derived in the next section) and the simulation.

III. STRUCTURAL PROPERTIES OF INTERGROUP NETWORK

In this section we derive the main theoretical results regarding the structure of the multiplicative random threshold graphs as defined in the previous section.

In Ref. [12] the authors present a detailed analysis of additive random threshold graphs, where an edge exists between two nodes if the sum of the weights (like the attractiveness parameters) associated with the nodes is larger than or equal to a given threshold value. In our work, we primarily focus on understanding the structural properties of the pruned one-mode projection of a preferentially grown α-BiN, which has a very specific setup, e.g., the attractiveness parameters associated with the nodes jointly follow a Dirichlet distribution. Furthermore the edge weights in our model are proportional to the product of the attractiveness parameters of the nodes. Although the Dirichlet distributed random variables are not independent, for large n they can be well approximated by independent exponential random variables. Next, we take an ordered list of these exponential random variables and thus, in order to calculate the fraction of nodes satisfying a specific property, we convert the exponential order statistics into uniform order statistics (a similar technique has been

[5] In general it can be easily shown that the properties of any random threshold graph hold true for all those graphs where the existence of an edge is decided based on the value of any monotonically increasing and symmetric function on the weight parameters assigned to the nodes of the graph.

followed in Ref. [12]). In the following we describe both of these techniques in detail:

(1) *Relationship between Dirichlet and exponential*: Let the random variables $\{\psi_1,\psi_2,\ldots,\psi_n\}$ jointly follow a Dirichlet distribution with parameters $\{\alpha_1,\ldots,\alpha_n\}$. It is well known that $\{\psi_1,\psi_2,\ldots,\psi_n\}$ can be represented as $\{X_1/S_n, X_2/S_n,\ldots,X_n/S_n\}$, where $X_i \sim \text{Gamma}(\alpha_i,1)$ and $S_n = \sum X_i \sim \text{Gamma}(\sum \alpha_i,1)\,(1 \leqslant i \leqslant n)$. For large n, S_n/n converges to 1, and since we assume that $\alpha_i = 1$ the X_i are exponentially distributed with rate parameter 1. Therefore, for large n we can well approximate the variables $\theta_i (= n\ \psi_i)$ by $\exp(1)$.[6]

(2) *Relationship between exponential and uniform order statistics*:
Let us consider n independent and identical exponential distributed random variables with rate parameter 1. Let the ordered sequence $\{X_i\}$, $1 \leqslant i \leqslant n$ be the order statistics of these random variables in a descending order, i.e., $X_{i+1} \geqslant X_i$, i is the rank of the variables in the sequence, and $\{U_i\}$ be the order statistics of n uniform (in [0,1]) distributed random variables. We use the following result to relate the two order statistics: an exponential random variable X having rate parameter γ can be represented as $\frac{-\ln U}{\gamma}$, and the ith order statistics X_i can be represented by $\frac{-\ln U_{(n-i)}}{\gamma}$. Let $\alpha_i = \frac{i}{n}$ be the normalized rank (hence, i can be represented as $n\alpha_i$). For the following it is crucial that for large n and fixed α the random variable $U_{\lfloor \alpha\cdot n \rfloor}$ becomes essentially localized in the sense that $\lim_{n\to\infty} U_{\alpha\cdot n}\cdot\frac{n}{i} = 1$ with probability one. Hence, for large n we can represent the exponential order statistics by its rank as

$$X_i = -\ln(1-\alpha_i). \tag{3}$$

In the following we calculate the degree distribution, the largest connected component size, the edge density, and the local clustering coefficient.

A. Degree distribution

There are essentially two equivalent ways to derive and characterize the degree distribution for threshold graphs. First, one can give the degree as a function of the value X_i of the node, and second, due to the localization property of the order statistics, one can present the degree as a function of the rank of the node. Further, due to the monotonicity of the degree in both variables (node value and rank) the ranking of the nodes according to their degrees is asymptotically the same as the ranking according to the node values. For real world networks one usually has no observable attractiveness value. Therefore the most natural way to compare with real networks would be to express the degree as a function of the node rank.

We fix the threshold Δ and consider a node x having rank $k = n\alpha$ (as defined above). To estimate the fraction of nodes which are connected with x we first consider the node that has the minimum rank among the nodes that are directly connected with node x. Let this specific node have rank $j = n\beta$. Clearly by definition of threshold graphs, all nodes with rank higher than or equal to j are directly connected with the node x. Figure 2(a) pictorially describes this scenario. Thus, the fraction of nodes connected with node x is precisely $1-\beta$. In other words, for the degree of node x with rank k we have

$$\frac{d(k)}{n} = 1-\beta.$$

For the evaluation of β we use the fact that j is the minimum integer in $[1,n]$ that satisfies the relationship $\theta_k \times \theta_j \geqslant \Delta$. Replacing the exponential order statistics by the corresponding uniform distribution order statistics [using Eq. (3)] we get

$$\frac{d(k)}{n} = e^{\frac{\Delta}{\ln(1-\alpha)}}. \tag{4}$$

The above formula relates the rank of a node with its degree in the network. The cumulative degree distribution can be now easily computed:

$$\begin{aligned} F_d(z) &= \text{Prob.}\left[\frac{d(i)}{n} \leqslant z\right] \\ &= \text{Prob.}\left[e^{\frac{\Delta}{\ln(1-\alpha)}} \leqslant z\right] \\ &= \text{Prob.}\left[\alpha \leqslant 1 - e^{\frac{\Delta}{\ln z}}\right] \\ &= 1 - e^{\frac{\Delta}{\ln z}}, \text{ since, } \alpha \text{ is a normalized rank} \end{aligned} \tag{5}$$

B. Largest component size

Owing to the star property and the fact that there is at most one component of size $\geqslant 2$ in any threshold graph, the highest ranked node, that is, the node with rank n, will always be a part of the largest connected component, and any other node

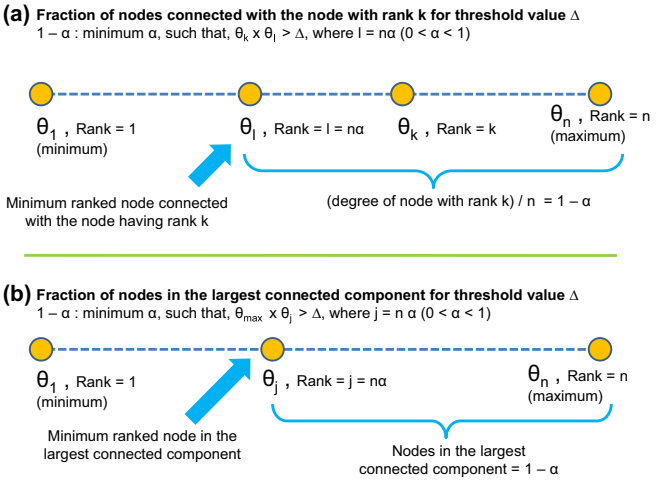

FIG. 2. (Color online) Pictorial representation of (a) the key idea in the computation of the degree distribution, i.e., the fraction of nodes in the system connected with a node having rank k, and (b) largest connected component size. The blue dotted lines indicate the list of nodes in the system sorted in a decreasing order (from left to right) according to their attractiveness parameter values (θ).

[6]In order to get rid of the dependent random variables, we replace the Dirichlet distributed random variables for the attractiveness parameters (ψ_i) by exponential distributed random variables. However, the range of the ψ_i is [0,1], whereas the range of the exponential distributed random variables is $[0,n]$. Hence, when we compare the mathematical results with the results obtained from real networks in Sec. IV, we require an adjustment of the edge weights.

with rank j, must satisfy the following condition to be a part of the largest connected component:

$$\theta_j \times \theta_n \geqslant \Delta. \tag{6}$$

Figure 2(b) pictorially describes this scenario. For large n the maximal order statistics of n i.i.d exponential random variables with rate parameter 1 is $\ln n + z$, where z is a random variable following a Gumbel distribution with parameters (1,1). On the other hand, the expected value of the minimal order statistics (θ_1) is of the order $O(\frac{1}{\ln n})$. Thus, asymptotically, for a fixed value of the threshold Δ, the fraction of nodes in the largest connected component becomes 1. Therefore, in order to get a largest connected component of size $x \cdot n$ with $x < 1$, we need to take threshold values of the order constant times $\ln n$. In the following we express the largest component size as a function of c for the threshold $\Delta = c \ln n$. A straightforward computation shows that the minimum value of α such that $\theta_{\alpha n} \times \theta_n \geqslant c \ln n$ is $1 - e^{-c}$ because we have $\frac{\theta_n}{\ln n} \to 1$ and $\theta_{\alpha n} \to -\ln(1-\alpha)$. Hence $\mathcal{L}(\Delta)$, the fraction of nodes in the largest connected component is given by

$$\mathcal{L}(\Delta) = e^{-c}. \tag{7}$$

C. Edge density

In the following we derive the edge density $\mathcal{E}(\Delta)$ for a given threshold value Δ where *edge density* is defined as the ratio of the actual number of edges existing in the network at Δ to the total possible number of edges in the graph [i.e., $\binom{n}{2}$]. Due to the multiplicative relationship between the edge weight and the node parameters and the approximation of Dirichlet distributed variables as explained above, the edge weight distribution is asymptotically the distribution of the product of two independent random variables, each following exponential distribution. Thus the fraction of edges that exist above a certain edge weight, which is the same as the edge density, is described as follows:

$$\begin{aligned}\mathcal{E}(\Delta) &= \Pr\{XY > \Delta\} \\ &= \int_0^\infty e^{-x} e^{-\frac{\Delta}{x}}\, dx \\ &= 2\sqrt{\Delta} K_1(2\sqrt{\Delta}).\end{aligned} \tag{8}$$

X and Y are independent, and $\sim \exp(1)$ and $K_1(z)$ is the modified Bessel function of a second kind. In the derivation we used the fact that the cumulative distribution function $F(z)$ of the product of two independent nonnegative random variables with density $\varphi(x)$ is given by $F(z) = 1 - \int_0^\infty \varphi(x)\varphi(\frac{z}{x})dx$.

D. Local clustering coefficient

We want to derive the local clustering coefficient $Cl_\Delta(\alpha)$ for a given node with rank $n\alpha$ for the threshold value Δ. For a given threshold value Δ, there is a minimal rank $h = n\gamma$ such that all nodes with rank $\geqslant h$ are connected with each other. These nodes form the maximal clique in the network. To evaluate this minimal rank we use the condition $(\theta_h)^2 = \Delta$, which is equivalent to $\ln(1-\gamma) = \sqrt{\Delta}$. Therefore $\gamma = 1 - e^{-\sqrt{\Delta}}$, and hence $e^{-\sqrt{\Delta}}$ is the fraction of nodes in the maximal clique.

A crucial point to be noted here is that if rank h is the minimum rank which is in the maximal clique, then the nodes having rank less than h have only edges with nodes which have rank larger or equal to h. This is a simple consequence of the fact that the edge weights $\theta_k \times \theta_j$ are monotonically increasing functions of k and j.

Since the nodes in the maximal clique are all connected, the nodes having rank less than h have local clustering coefficient 1. However, the local clustering coefficient of the nodes above rank h have a clustering coefficient less than 1.

Let us take a node x having rank $n\alpha > h$ (i.e., it is a node inside the maximal clique). First, we categorize the nodes that are connected with x into two parts: (a) nodes having ranks lesser than h and (b) nodes having rank higher or equal to h. Each of these two categories of nodes contributes to the local clustering coefficient of x separately. In the following we derive these two parts.

The total number of nodes connected to x, i.e., the degree of x is $ne^{\frac{\Delta}{\ln(1-\alpha)}}$. Therefore, the number of unordered pairs of nodes that are neighbors of x are

$$\frac{\left(ne^{\frac{\Delta}{\ln(1-\alpha)}}\right)\left(ne^{\frac{\Delta}{\ln(1-\alpha)}} - 1\right)}{2} \simeq \frac{1}{2} n^2 e^{\frac{2\Delta}{\ln(1-\alpha)}}. \tag{9}$$

On the other hand, the number of pairs of connected nodes inside the maximal clique with which x is connected is

$$n^2 \tfrac{1}{2} e^{-2\sqrt{\Delta}}. \tag{10}$$

Note further that all nodes with node weights between $-\frac{\Delta}{\ln(1-\alpha)}$, i.e., the minimum weight of a node connected with x and $\sqrt{\Delta}$, i.e., the minimum node weight in the maximal clique, are connected to x. Let us denote this interval by $I_{\alpha,\Delta}$. The degree fraction of a node i, for $\theta_i \in I_{\alpha,\Delta}$, is given by Eq. (4) as follows:

$$\frac{d(i)}{n} = e^{-\frac{\Delta}{\theta_i}},$$

where θ_i is the attractiveness value of node i.

To calculate the number of node pairs (i,j) with $\theta_i \in I_{\alpha,\Delta}$ and $\theta_j \geqslant \sqrt{\Delta}$ such that there is an edge between i and j, we need to sum over the degrees of the nodes i. There will be no double counting, since all the nodes in $I_{\alpha,\Delta}$ are connected only with the nodes above the upper bound of the region, i.e., above node weight $\sqrt{\Delta}$. This can be calculated as the following sum:

$$n \sum_{\theta_i \geqslant -\frac{\Delta}{\ln(1-\alpha)}}^{\theta_i \leqslant \sqrt{\Delta}} e^{-\frac{\Delta}{\theta_i}}.$$

For large value of n, this sum can be approximated by the following integral:

$$n^2 \int_{-\frac{\Delta}{\ln(1-\alpha)}}^{\sqrt{\Delta}} e^{-\theta - \frac{\Delta}{\theta}}\, d\theta. \tag{11}$$

Note that to obtain the formula above, we have used the fact that the θ_i are distributed as exp(1).

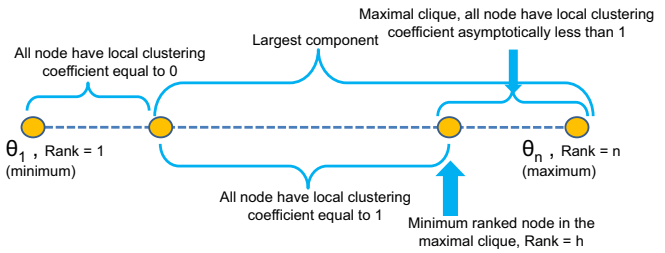

FIG. 3. (Color online) Pictorial representation of different classes of nodes arranged in an ascending order of their attractiveness parameter values and having different values of the local clustering coefficient.

Combining Eqs. (9), (10), and (11), we get the expression for the local clustering coefficient $L_\alpha(\Delta)$ as follows:

$$Cl_\Delta(\alpha) = 2e^{\frac{-2\Delta}{\ln(1-\alpha)}}\left[\int_{-\frac{\Delta}{\ln(1-\alpha)}}^{\sqrt{\Delta}} e^{-\theta-\frac{\Delta}{\theta}}\,d\theta + \frac{e^{-2\sqrt{\Delta}}}{2}\right]. \quad (12)$$

See Fig. 3 for a pictorial description of range of the ranks of the nodes having different classes of local clustering coefficient. The analysis also reveals the fact that, for any threshold value, the largest component of the network will have two distinct classes of nodes: (a) the nodes that are in the maximal clique and hence, completely connected with each other, and (b) the nodes that are in the largest component but outside the maximal clique; they have no connection among each other, but rather are connected to some node in the maximal clique. Hence, the nodes in part (b) have clustering coefficient 1, while the nodes in part (a) have a clustering coefficient less than 1. Thus, the structure corresponds to a perfect core-periphery organization [21,22].

Due to Eqs. (3) and (4) there is an asymptotic one-to-one correspondence between degree, rank, and the weight (θ) associated with a node in the network. Equation (12) expresses the local clustering coefficient of a node as a function of its rank (α). However, the problem with this formula is that the rank of a node is based on the node weights (θ), which are only intrinsic variables and are not an observable in real networks. Asymptotically the rank according to the θ values can be replaced by the rank according to the degrees which are an observable although not local. In contrast, rank computation needs a complete survey of the graph. On the other hand, degree of a node can be easily computed through local observations of the neighbors of a node. Hence, for testing whether a given graph is like a threshold graph, it is very useful to have an expression of the local clustering coefficient of a node as a function of its degree [$Cl_\Delta(d)$], which we provide in the following.

Let us consider a node x having degree d. From Eq. (4) and the fact that the node weight lesser than $\sqrt{\Delta}$ has clustering coefficient 1, we get immediately that $Cl_\Delta(d) = 1$ for $d \leqslant ne^{\sqrt{\Delta}}$. The value of $Cl_\Delta(d)$ for $d \geqslant ne^{\sqrt{\Delta}}$ is derived as follows.

Using Eqs. (3) and (4), the right side of Eq. (9) can be rewritten for node x as

$$\frac{1}{2}n^2\left(\frac{d}{n}\right)^2. \quad (13)$$

Similarly, Eq. (11) can be rewritten as

$$n^2 \times \int_{\ln n-\ln d}^{\sqrt{\Delta}} e^{-\theta-\frac{\Delta}{\theta}}\,d\theta. \quad (14)$$

Thus, combining Eqs. (10), (13), and (14), the local clustering coefficient for a given degree d and threshold Δ ($Cl_\Delta(d)$) can be expressed as follows:

$$Cl_\Delta(d) = \frac{2}{\left(\frac{d}{n}\right)^2}\left[\int_{\ln n-\ln d}^{\sqrt{\Delta}} e^{-\theta-\frac{\Delta}{\theta}}\,d\theta + \frac{e^{-2\sqrt{\Delta}}}{2}\right]. \quad (15)$$

IV. COMPARISON WITH REAL DATASETS

We compare the mathematical results described in the previous section with the same obtained from the measurements done on publicly available user-group membership datasets [23] from three different online social systems: (1) YouTube, a video-sharing site, (2) Flickr, a photo-sharing site, and (3) LiveJournal, which allows users to share blogs, diary, journals, and so on. The details of the formation of the real intergroup networks are described below.

Real intergroup networks: YouTube, Flickr, and LiveJournal datasets originally contain 30 087 groups and 1 157 827 users, 103 648 groups and 1 846 198 users, 7 489 073 groups and 5 284 457 users, respectively. However, as pointed out earlier, many of these groups can be ignored from the perspective of the number of users joining the groups. Therefore, instead of considering all the groups present in the data set, we extract out a certain number of the most popular groups

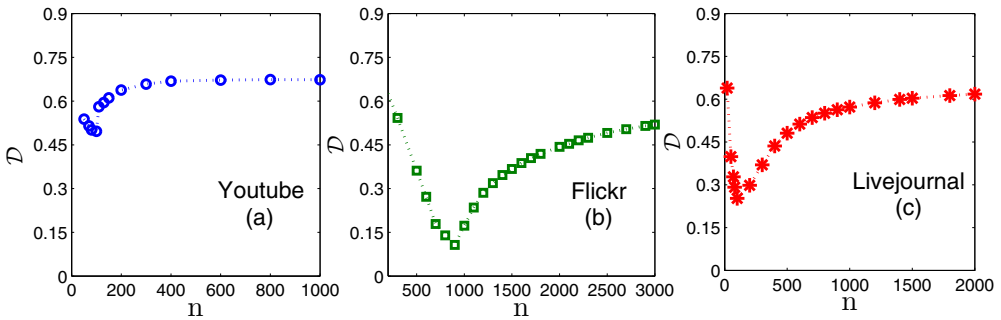

FIG. 4. (Color online) Plot of the difference between the theory and the real datasets calculated using Eq. (17). Part (a), (b), and (c) show the results obtained from the dataset of YouTube, Flickr, and LiveJournal, respectively.

TABLE I. Details of the real data set.

Description	$n = \|U\|$	$t = \|V\|$	μ	μ'
YouTube	100	37 163	1.88	7.99
Flickr	1000	220 827	12.94	760.86
LiveJournal	100	2 550 531	6.14	70.52

and corresponding users who joined those groups. Using these datasets we first form the intergroup network as described in the previous section.

Transformation of the edge weights: As expressed in Eq. (2), the weight of an edge depends on t, $(\mu' - \mu)$ as well as the attractiveness parameters associated with the end nodes of the edge. However, as already pointed out, for the sake of simplicity in the analysis, we consider only the attractiveness parameters. In order to do the necessary adjustment we divide the weights of the individual edges by $t(\mu' - \mu)$. [This makes the edge weights independent of the number of users as well as the distribution of the number of connections with the social groups the users make, i.e., $F(u)$.] Furthermore, in the mathematical derivations, we also convert the Dirichlet distributed attractiveness parameters (i.e., ψ), each having range $[0,1]$, to exponentially distributed parameters, each having range $[0,n]$. Therefore, as an adjustment, we also multiply the weights of each of the edges in the real intergroup network by n^2. In summary, we multiply the weights of the edges in the real intergroup network with the following factor:

$$\frac{n^2}{t(\mu' - \mu)}. \tag{16}$$

The datasets: It is to be noted that in all our theoretical analysis we assumed a large value of n. Hence, as per the theory, the corresponding value of t should be even larger in comparison to the value of n (to satisfy the basic definition of α-BiN). However, the cardinalities of the top and the bottom sets in all the real datasets are finite. Therefore, from each of the three distinct base datasets, we first prepare many different subdatasets with a different number of most popular groups (i.e., the value of n), and for each such subset, we consider t as the number of users who are members of at least one of these most popular groups. Owing to simple empirical computations of the largest connected component size, we first compare this quantity with the theoretical predictions for different values of n.

We quantify the extent of match between theory and real data for all these different values of n, by computing their difference as follows. The absolute difference between the largest connected component size obtained from theory and real data for each threshold value is first normalized by the largest connected component size obtained from the real data. Next, these normalized absolute differences are averaged over all the threshold values. This is given in the following formula:

$$\mathcal{D} = \sum_{\Delta=a}^{b} \frac{|\mathcal{L}_{th}(\Delta) - \mathcal{L}_{re}(\Delta)|}{\mathcal{L}_{re}(\Delta)}, \tag{17}$$

where $\mathcal{L}_{th}(\Delta)$ and $\mathcal{L}_{re}(\Delta)$ are the values of the fraction of the nodes in the largest connected component obtained from theory and the real data for the threshold value Δ, respectively: a and b being the minimum and the maximum threshold values considered for the comparison, respectively. We select the results obtained from the real datasets to normalize the absolute difference because real results were always found to remain higher than the theory. Figure 4 shows this result for the three real datasets.

It can be seen from the figure that as n increases up to a certain value, a match between the theory and the real dataset also increases. This happens due to our assumption of a large value of n in the mathematical computation. However, after a certain value of n, the match starts degrading. The reason is that, with the increase in n, the associated value of t also increases. However, due to the finiteness of the real datasets, as n becomes large the value of t becomes insufficient in comparison to n to satisfy the fundamental property of α-BiN, $t \gg n$. Therefore, we find that for each dataset there is a certain value of n for which the match is the best (below which the

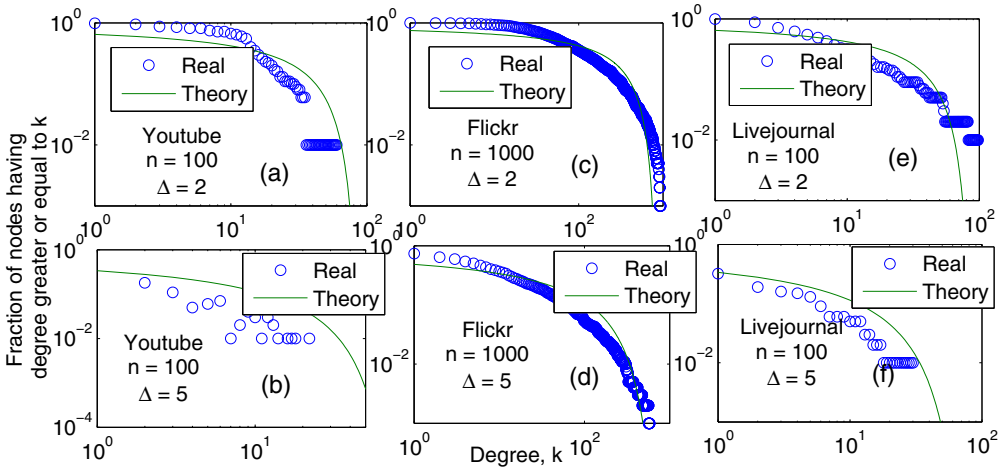

FIG. 5. (Color online) Comparison of the fraction of nodes having degree larger or equal to a certain value, i.e., a complementary cumulative degree distribution [calculated from Eq. (5)] with the same obtained from the pruned intergroup network derived from YouTube (a, b), Flickr (c, d), and LiveJournal (e, f) datasets.

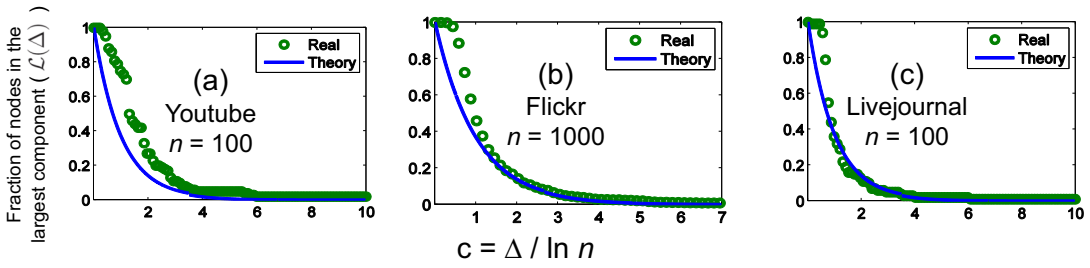

FIG. 6. (Color online) Comparison of the analytically derived fraction of the nodes in the largest component of the multiplicative random threshold graph with the same obtained from the pruned intergroup network derived from YouTube (a), Flickr (c), and LiveJournal (e) datasets with a few different values of n.

value of n is not sufficient and above which the value of t is not sufficient). This n value in case of YouTube, Flickr, and LiveJournal is around 100, 1000, and 100, respectively (see Fig. 4).

For the rest of the comparisons between theory and real data, we select these specific n values. The details of all parameters (n, t, μ, and μ') for these specific subdatasets of all three real systems are given in Table I. In the following we present the comparison of the theoretically derived formula of degree distribution, largest connected component size, edge density, and local clustering coefficient given in Eqs. (5), (7), (8), and (12), respectively, with the same obtained from the pruned intergroup networks derived from these three subdatasets. Figures 5, 6, 7, and 8, respectively, depict these comparisons for different values of threshold Δ.

YouTube: The dataset derived from YouTube was seen to be matching accurately in many cases [e.g., see Fig. 5(b) for degree distribution, Fig. 6(a) for largest component size, and Fig. 7(a) for edge density]. However, in some cases the match was not so accurate (e.g., see Fig. 5(a) for degree distribution as well as Figs. 8(a) and 8(b) for the local clustering coefficient).

Flickr: The match between the theory and the results derived from the dataset of Flickr was the most accurate among all three different datasets in all four observable properties [see Figs. 5(c) and 5(d) for degree distribution, Fig. 6(b) for largest component size, Fig. 7(b) for edge density, and Figs. 8(c) and 8(d) for local clustering coefficient].

LiveJournal: The dataset derived from LiveJournal was also seen to be matching well with theory in terms of all four observable properties [see Figs. 5(e) and 5(f) for degree distribution, Fig. 6(c) for largest component size, Fig. 7(c) for edge density, and Figs. 8(e) and 8(f) for local clustering coefficient]. However, the match was not as well as with Flickr, and not as bad as with YouTube.

A. Discussion

The comparison between real data and theory shows quite a good match from an overall perspective. The few cases of inaccurate matching were from the dataset obtained from both YouTube, and LiveJournal. However, LiveJournal always shows a better match than YouTube. The gradation in results directly correlates with the margin of error ($\mathcal{D}$), and the three data sets manifest when compared with theory (see Fig. 4). It can be seen that LiveJournal shows a behavior closer to theory in comparison to YouTube in the case of more complex properties such as degree distribution and local clustering coefficient. In order to have a visual understanding of the structure, we also present snapshots (Fig. 9) of the pruned intergroup networks from all three datasets. Although pruned intergroup networks are unweighted graphs by definition, to perfectly reflect the organization of the edges and their weights in the network, Fig. 9 shows them as weighted graphs. It can be seen that unlike Flickr and LiveJournal, in YouTube there are more small components which deviate from the basic property of a random threshold graph. Moreover, the hierarchical arrangement of the nodes as well as the edges are not so clear in YouTube, unlike Flickr and LiveJournal.

Thus, it can be concluded that the small mismatches resulting from the YouTube dataset are due to the deviation of its structure from the definition of the random threshold

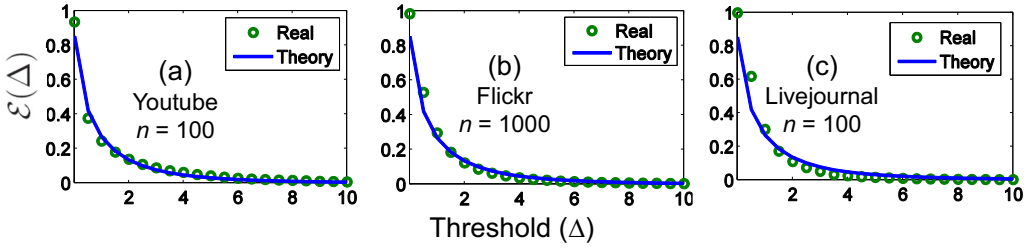

FIG. 7. (Color online) Comparison of the analytically derived fraction of the edges in the multiplicative random threshold graph with the same obtained from the pruned intergroup network derived from YouTube (a), Flickr (c), and LiveJournal (e) datasets with a few different values of n.

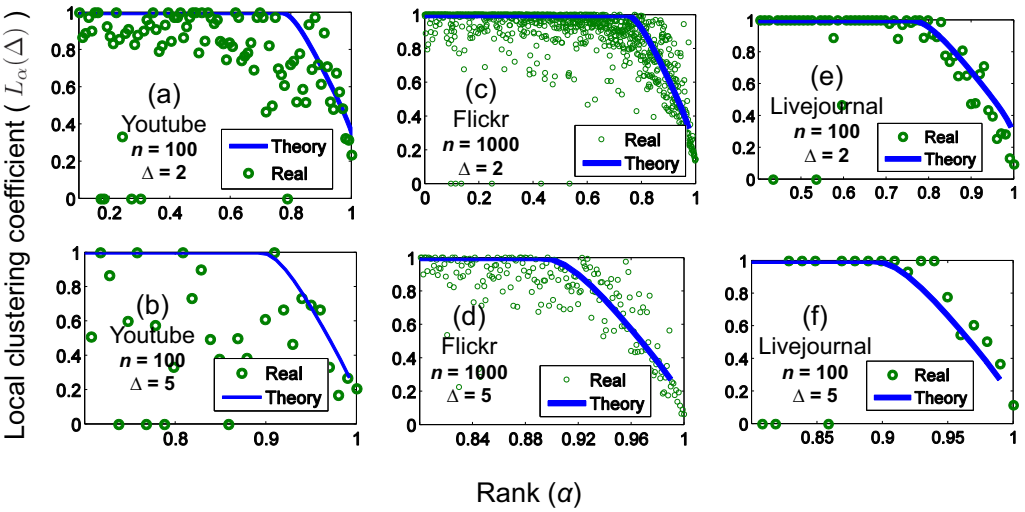

FIG. 8. (Color online) Comparison of the analytically derived expression for the local clustering coefficient of the nodes in a multiplicative random threshold graph with the same obtained from the pruned intergroup network derived (in log-log scale) from YouTube (a, b), Flickr (c, d), and LiveJournal (e, f) datasets with a few different values of n.

graph structure. Fundamentally, this might have come from several reasons such as partial nonpreferential selection of the groups by the users or simultaneous growth of the set of groups at a rate comparable to the rate of growth of the set of users. These observations together may indicate that users' choice of categories of videos are more diverse, possibly due to a higher number of available options, than the selection of categories of pictures or blogs/journals. However, the cases of a good match with the theory also support the fact that a significant portion of YouTube follows the fundamental definition of multiplicative random threshold graphs as well.

V. RELATED WORKS

Analysis of the evolution and the structure of social groups has been the central focus of much research in the past [24]. Common membership of users in the groups has been also studied as an overlapping community detection and analysis problem [1,2]. However, the groups that are specifically defined by the users themselves have special importance. Reference [25] characterizes them as ground-truth communities. In this work the authors consider many real world social networks along with their ground-truth communities. They analyzed the existing structural definitions of community and

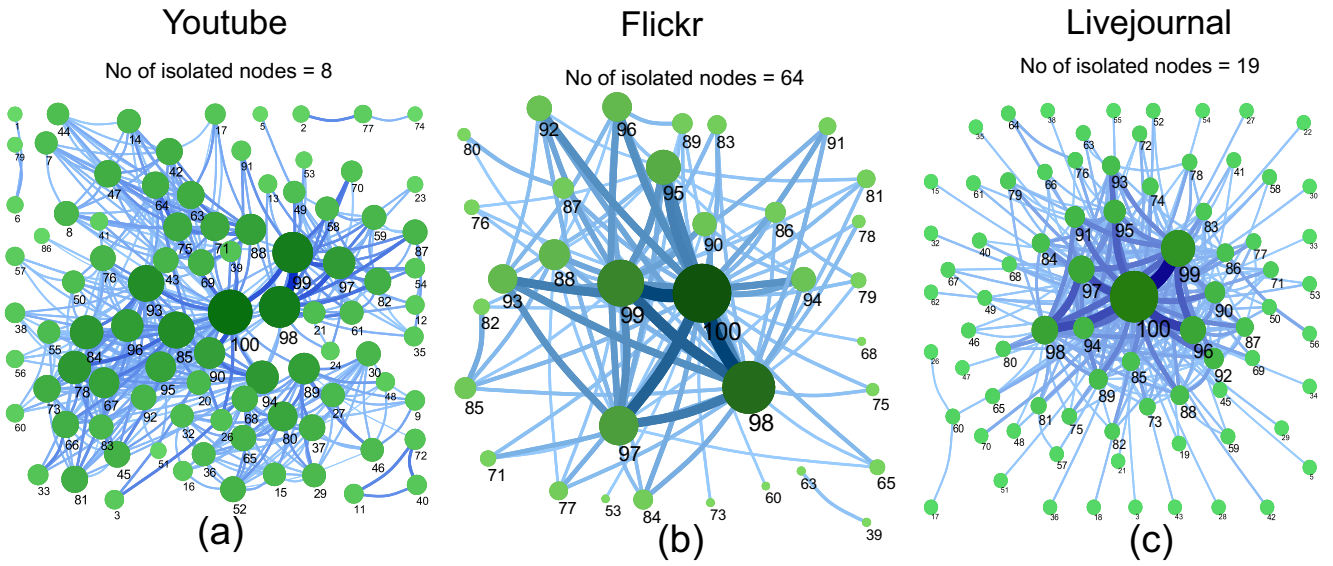

FIG. 9. (Color online) The graphical representation of the network structure of the pruned intergroup networks (with 100 most popular nodes) obtained from the dataset of (a) YouTube, (b) Flickr, and (d) LiveJournal for threshold value 3. Only the nodes in the largest connected component have been displayed (number of isolated nodes are mentioned in the boxes). The thickness of an edge is proportional to its weight in the corresponding intergroup networks. The size of a node is proportional to its page rank in the network. The numbers associated with the nodes represent their ranks (the top rank is 100).

present a comparative study of how well these definitions can appropriately capture the user-defined communities. A number of other studies have been performed on these user-defined communities also in online social networks. For example, Ref. [26] studies the growth of these groups and shows that the underlying network structure among users plays a significant role in this growth. A recent work [27] does a deeper analysis of the relationship between the group selection and the object selection (objects within a group, e.g., artists within a group of users who like a specific category of music) strategy followed by the users in online social systems by user-object and user-group bipartite network models. Moreover, the concept of pruned one-mode projection of bipartite networks was also used in Ref. [28] to understand the structural robustness or modularity of various synthetic as well as real networks under random failure. In our work we focus on the impact of the common membership of the users in different groups. Specifically, we lay the foundation for the study of social influence and spread of information or knowledge abstracted through the connectivity structure of the network.

On the other hand, threshold graphs have been studied independently by many researchers for a long time under different names, e.g., difference graph, intersection graph, interval graphs, biorders [17], etc. References [19] and [12] introduced the concept of random threshold graphs where the node weights follow a specific probability distribution, and these works analyze their graph theoretic properties. In fact, the same concept was introduced in the physics community as the fitness model (a nongrowing model) for generating scale free-networks [29,30]. Later the same idea was used in many other related fields, e.g., nongeographical and geographical threshold graphs, gravity models [31–33], etc. In the current work we show that threshold graph can also be used to study the intergroup relationships in social networks, and in general the thresholded one-mode projection of α-BiN. Specifically, we analyze a new class of random threshold graphs, called multiplicative random threshold graphs, with node weights jointly following a Dirichlet distribution.

VI. CONCLUSION

In this paper we analyze the structure of an intergroup network (in user-group systems), which is fundamentally a reflection of the social behavior of humans. We show that many of the structural properties of the pruned intergroup network can be well explained by the theory of multiplicative random threshold graphs. We could come up with closed form equations for the important properties: like-degree distribution, largest component size, and edge density (for large number of social groups and users). In fact, a general mathematical framework has been developed and used in this analysis which can be further extended to understand various other properties of the network or even similar structures based on multiplicative random threshold graphs. We also find that the theoretical predictions match very well with several real world intergroup networks.

ACKNOWLEDGMENT

S.S. thanks Tata Consultancy Services (TCS) and Samsung Pvt. Ltd for financial assistance.

[1] G. Palla, A.-L. Barabási, and T. Vicsek, Nature (London) **446**, 664 (2007).
[2] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761 (2010).
[3] A. Grölund and P. Holme, Phys. Rev. E **70**, 036108 (2004).
[4] J. Yang and J. Leskovec, in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (ACM, New York, 2013), pp. 587–596.
[5] S. Ghosh, S. Saha, A. Srivastava, T. Krueger, N. Ganguly, and A. Mukherjee, IEEE J. Selected Areas Commun. **31**, 584 (2013).
[6] D. Kimura and Y. Hayakawa, Phys. Rev. E **78**, 016103 (2008).
[7] Z. Zhao, J. P. Calderón, C. Xu, G. Zhao, D. Fenn, D. Sornette, R. Crane, P. M. Hui, and N. F. Johnson, Phys. Rev. E **81**, 056107 (2010).
[8] W. Zeng, A. Zeng, M.-S. Shang, and Y.-C. Zhang, CoRR arXiv:1308.3059.
[9] J. Koskinen and C. Edling, Social Netw. **34**, 309 (2012).
[10] M. Choudhury, N. Ganguly, A. Maiti, A. Mukherjee, L. Brusch, A. Deutsch, and F. Peruani, Phys. Rev. E **81**, 036103 (2010).
[11] D. M. Romero, B. Meeder, and J. Kleinberg, in *Proceedings of the 20th International Conference on World Wide Web* (ACM, New York, 2011), pp. 695–704.
[12] P. Diaconis, S. Holmes, and S. Janson, Internet Math. **5**, 267 (2008).
[13] F. Peruani, M. Choudhury, A. Mukherjee, and N. Ganguly, Europhys. Lett. **79**, 28001 (2007).
[14] A. Mukherjee, M. Choudhury, and N. Ganguly, Physica A **390**, 3602 (2011).
[15] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[16] N. Ganguly, S. Ghosh, T. Krueger, and A. Srivastava, Theor. Comput. Sci. **466**, 20 (2012).
[17] N. V. R. Mahadev and U. N. Peled, *Threshold Graphs and Related Topics* (North Holland, Amsterdam, 1995).
[18] V. Chvátal and P. L. Hammer, Ann. Discrete Math. **1**, 145 (1977).
[19] E. P. Reilly and E. R. Scheinerman, Elec. J. Combin. **16**, R130 (2009).
[20] F. Chung, S. Handjani, and D. Jungreis, Ann. Combin. **7**, 141 (2003).
[21] S. P. Borgatti and M. G. Everett, Social Netw. **21**, 375 (2000).
[22] P. Holme, Phys. Rev. E **72**, 046111 (2005).
[23] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, California, USA* (ACM, New York, 2007), pp. 29–42.

[24] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[25] J. Yang and J. Leskovec, in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics* (ACM, New York, 2012), p. 3.
[26] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2006), pp. 44–54.
[27] W. Zeng, A. Zeng, M.-S. Shang, and Y.-C. Zhang, Eur. Phys. J. B **86**, 375 (2013).
[28] J. P. Bagrow, S. Lehmann, and Y.-Y. Ahn, arXiv:1102.5085.
[29] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[30] V. D. P. Servedio, G. Caldarelli, and P. Buttà, Phys. Rev. E **70**, 056126 (2004).
[31] N. Masuda, H. Miwa, and N. Konno, Phys. Rev. E **71**, 036108 (2005).
[32] N. Masuda, H. Miwa, and N. Konno, Phys. Rev. E **70**, 036124 (2004).
[33] A. Hagberg, P. J. Swart, and D. A. Schult, Phys. Rev. E **74**, 056116 (2006).