Community Structures in Classical Network Models *

Angsheng Li
State Key Laboratory of Computer Science
Institute of Software
Chinese Academy of Sciences
P. O. Box 8718
Beijing, 100190, P.R.China
angsheng@ios.ac.cn

Pan Peng
State Key Laboratory of Computer Science
Institute of Software
Chinese Academy of Sciences
P. O. Box 8718
Beijing, 100190, P.R.China
pengpan@ios.ac.cn

ABSTRACT

Communities (or clusters) are ubiquitous in real world networks. Researchers from different fields have proposed many definitions of communities, which are usually thought of a subset of nodes that vertices in the set are well connected with other vertices in it and have relatively less connections with vertices outside of the set. Unlike the traditional research which mainly focuses on detecting and/or testing the clusters, we propose a new definition of community and a novel way to study community structure, with which we are able to investigate mathematical network models to test whether they have the *small community phenomenon* or not, i.e., whether every vertex in the network belongs to some small community. We examine various models and establish both positive and negative results: we show that in some models small community phenomenon exists while in some other models it does not.

Categories and Subject Descriptors

G.2.2 [**Mathematics of Computing**]: Graph Theory—*Network problems*

General Terms

Theory

Keywords

network models, small community phenomenon, conductance

1. INTRODUCTION

There has been a number of interesting phenomena in the study of large scale networks. For example, the degree sequence in many networks obey the power law distributions [5, 34], which means that the number of nodes with k links is proportional to $k^{-\gamma}$ for some constant γ. Typical social, technological and biological networks exhibit the *small world phenomenon* [41, 26], namely, for almost every pair of nodes in the graph, they are connected by a short path through the network and in some cases we can find the short path efficiently by using only local information. Other typical characters include "triadic closure" property [38], the "densification and shrinking diameters" in the evolution of networks [30] and the property of community structures [19], which is the main focus in this paper.

Communities (also called "clusters" or "modules") are naturally thought of cohesive subgraphs in a network. Informally, vertices in a community are well interconnected with members therein and have relatively less connections with vertices outside of the community. Communities appear in a wide range of areas. For instance, in protein-protein interaction (PPI) networks, groups of proteins sharing the same or similar functions are clustered together [24]; in society, the communities may correspond to groups of friends or coworkers [21]; in scientific collaboration networks, scientists who have similar either of research topic or of methodology grouped together to form communities [20].

Former research focuses heavily on how to find and test these common clusters in networks. Many algorithms have been proposed to detect communities. To name a few, agglomerative or divisive ideas combined with some specific vertex (or link) similarity measures are used to find clusters [23, 20]. Due to the many similar characters between clustering and graph partitioning, in which spectral techniques work practically well, spectral algorithms are also used to find clusterings [40]. Modularity-based methods have been very influential in recent research [36, 10]. Other works may first treat communities in some specific perspective and then utilize it to achieve their specific goals, e.g. Palla Derenyi, Farkas and Vicsek [37] view communities as a chain of adjacent cliques and using this they can find overlapping and/or nesting communities. Testing the quality of a community has also been studied [29, 28]. For more applications and experimental results on community detection, see a recent survey [19].

Though there are extensive studies on finding and testing communities, there is no uniform or standard definition for communities. In fact, many papers for finding clusters do not give a precise definition (mathematically) but give algorithms which will output cohesive subsets of the nodes of graphs and then these sets are treated as the communities (e.g., [20, 1]). Traditional definitions vary much from fields to goals. Some definitions involve the global structure of the community (e.g., one can expect a partition of a graph to contain good communities if the partition is an (approximate) optimal solution to a global modularity function, which involves some quantities in the real world network and

*The research is partially supported by NSFC distinguished young investigator award number 60325206, and its matching fund from the Hundred-Talent Program of the Chinese Academy of Sciences. Both authors are partially supported by the Grand Project "Network Algorithms and Digital Information" of the Institute of Software, Chinese Academy of Sciences.

the corresponding quantities in a random model preserving the degree sequence of the original network [36]) and some are based on the local property of clusters (e.g., a clique or clique-like subgroup is supposed to be a good community [27, 37]).

In this paper, we introduce a new definition of community. In our definition, a community is allowed to overlap and/or nest other communities. Besides, our definition provides a quantitative way to compare the quality of two communities (compare the *community components*, see Section 2). This definition uses the concept of the conductance of a subset of the graph. Conductance measures somehow the relative ratio between the number of edges incident to the subset and the number of edges in the set, and it plays important roles in graph theory, algorithms and statistical physics [9, 22]. Some of the conductance results of random graphs have also been investigated [13].

There were several papers which connect conductance to clustering. Kannan, Vempala and Vetta [25] have proposed a bicriteria measure for assessing the quality of a clustering. They define a good clustering as a set of clusters in which each cluster has high conductance and the weight of inter-cluster edges only takes a small portion of the total edge weight. Their main goal is to analyze a spectral algorithm which gives a good approximation solution to the clustering problem under their definition. Leskovec et al. [31] directly use the conductance to measure the goodness of a community. A good community is supposed to have low conductance. The authors of [31] considered the quality of network communities at different size of scales. Specifically, they studied a quantity which is the minimum conductance over all sets of size k in the entire graph and they plot this quantity as a function of k over 100 large scale networks. In this way, they can analyze the relationship between the quality and the size of a community. One of the many interesting observations is that the size of the best community (with minimal conductance) in many large-scale networks is around 100. This observation matches the Dunbar's number [12] which predicts that a stable community has size upper bounded by 150 or so.

Unlike many other papers (eg. [33, 25]) which first give a definition of community and then develop algorithms to find subsets satisfying the definition, we study the *small community phenomenon* in various networks. This is motivated by the common experience that in many social networks, almost every node belongs to at least one small community. This intuition is to some extent confirmed by the work of Allen [2] who finds that on-line communities have around 60 members and some other evidence which supports Dunbar's theory on the limit size of a stable community. The observation (see the last paragraph) of the work of Leskovec et al. [31] also gives evidence that small communities not only exist but also have the best quality in many large-scale networks. On the other hand, as we mentioned, we use the conductance to measure the quality of a set. Though conductance has been used to characterize communities (e.g. [31]), we combine the conductance of a set and its size in a more refined way which has never been considered before. We investigate our definition on a variety of random graph models to check whether they have the small community phenomenon or not. Through this line of research we can both determine whether a given model is good or not to be used to validate real world networks and motivate to design more appropriate network models.

We believe that our results not only build a theoretical frame for the study of community structures, but have potential applications in understanding other structural properties and/or dynamic behaviors of networks in general. For example, Chierichetti, Lattanzi and Panconesi [8, 7] recently established the connection between the rumor spreading on graphs and its conductance. It is known that communities play important roles in rumor spreading (see e.g. [4]), reflecting the intuition that rumor spreads quickly within a community. This experiment needs a mathematical proof, for which, our definition for community may as well be used, which is an interesting open question.

In Section 2, we will give some basic definitions on good communities, some corresponding quantities and formulate the concept of small community phenomenon. In Section 3 we will investigate the small community phenomenon on a set of classical network models, including the Erdös-Rényi model [14], the geometric preferential attachment model [17, 18], the hierarchical model [39], and in Section 4 we consider the community structure of some perturbed graphs, including the small world model, and show that small community phenomenon in a graph may be viewed as a slightly dual property of being an expander.

2. BASIC DEFINITIONS

Given a simple graph $G = (V, E)$, let d_v denote the degree of a vertex $v \in V$. The *volume* $\mathrm{vol}(S)$ of a subset $S \subseteq V$ is the sum of degrees of vertices in it, i.e., $\mathrm{vol}(S) = \sum_{v \in S} d_v$. Noting that the volume of V is twice of the number of edges, we denote it by $\mathrm{vol}(G) = \mathrm{vol}(V) = 2|E|$. For any two vertex subsets $S, T \subset V$, denote $e_G(S, T)$ to be the number of edges with one endpoint in S and the other in T, and denote $e_G(S)$ to be the number of edges with both endpoints in S. When it is clear from the context, we will abbreviate $e_G(S, T)$ and $e_G(S)$ as $e(S, T)$ and $e(S)$, respectively. Then obviously, $\mathrm{vol}(S) = 2e(S) + e(S, \bar{S})$. For $S \subseteq V$ and $\mathrm{vol}(S) \leq \frac{1}{2}\mathrm{vol}(G)$, the conductance of S is defined as:

$$\Phi(S) = \frac{e(S, \bar{S})}{\mathrm{vol}(S)}.$$

For S with $\mathrm{vol}(S) > \frac{1}{2}\mathrm{vol}(G)$, its conductance is defined to be the conductance of its complement, namely $\Phi(S) = \Phi(\bar{S})$.

Leskovec et al. [31] used the conductance of a set S to measure the goodness of the community S. As easily seen from the definition, conductance of a set S provides somehow a measure of the relative ratio between the number of edges incident to the set and the number of edges contained in the set. Thus, conductance is intuitively related to a community. More specifically, the smaller the conductance of a set S is, the more likely that S is a good community. Moreover, it is natural to require a community to be connected, which ensures that every two nodes in the community can establish a connection only through the nodes inside the community.

We will also require that the size of a meaningful community in a graph (or model) G will depend on the number

of vertices n in G, which means that we will not consider a set of constant size to be a proper community. This requirement can be seen as follows: on one hand, we are more interested in how communities change as the size of the network grows. On the other hand, a set of too small size can hardly be thought of a reasonable community [2]. Moreover, empirical results reveal that the sizes of many communities do scale with the sizes of the networks (see eg. [37]).

We now extend the idea of [31] to define a more refined way to measure the goodness of a community.

DEFINITION 1. *Given a graph $G = (V, E)$ and $\alpha, \beta > 0$, a connected set $S \subset V$ with $|S| = \omega(1)$[1] is a strong (α, β)-community[2] if*

$$\Phi(S) \leq \frac{\alpha}{|S|^{\beta}}.$$

Moreover, if $|S| = O((\ln n)^{\gamma})$, where $n = |V|$, then we say that S is a strong (α, β, γ)-community.

If the conductance satisfies some weaker condition, we can define a weak community. Formally,

DEFINITION 2. *Given a graph $G = (V, E)$ and $\alpha, \beta > 0$, a connected set $S \subset V$ with $|S| = \omega(1)$ is a weak (α, β)-community if*

$$\Phi(S) \leq \frac{\alpha}{(\ln |S|)^{\beta}}.$$

The weak (α, β, γ)-community can be defined similarly.

We call β as the *community exponent* of the graph. It is easily seen that $0 \leq \beta \leq 2$ in the definition of strong community. β captures the quality of a community. Intuitively, for a strong (α, β)-community S, if β is large, then the fraction of edges out of the S that cross the cut is low, which means that S is more community like. Thus, to some extent, we can say that if $\beta_1 > \beta_2 > 0$, strong (α_1, β_1)-community is better than strong (α_2, β_2)-community, which is again better than any weak community.

In many cases, we want to know whether a given random network model satisfies the small community phenomenon, i.e., every vertex in the graph is contained in some small communities.

DEFINITION 3. *Given a random graph G with vertex set V and $|V| = n$, if with high probability , almost every[3] vertex v is contained in a strong or weak (α, β, γ)-community, where $\alpha, \beta, \gamma > 0$ are some constants independent of n, then G is said to have the small community phenomenon.*

In the remaining sections, we will also use a quantity related to conductance, which is called *expansion*, and is introduced here.

DEFINITION 4. *In graph $G = (V, E)$, the expansion of a subset $S \subseteq V$ is*

$$\alpha(S) = \frac{e(S, \bar{S})}{\min(|S|, |\bar{S}|)}.$$

The expansion of the graph $\alpha(G)$ is $\min_{S \subseteq V, |S| \leq |V|/2} \alpha(S)$.

Flaxman et al. [16, 17, 18] studied the graph expansion of some network models.

3. RESULTS ON CLASSICAL NETWORK MODELS

In this section, we investigate the community structure (based on our definition) on several classical network models. We will see that some models do capture the small community structure, while others do not.

3.1 Erdös-Rényi model

Erdös-Rényi model [14] is one of the most basic network models. It is also called the $G(n, p)$ model, in which each potential edge appears with probability p, independently of other edges. We will see that for p large enough (in which case the graph is connected with high probability) this model does not have the small community phenomenon.

THEOREM 1. *If $p = \omega(n) \ln n/n$, where $\omega(n) \to \infty$ arbitrarily slowly, then with high probability, a random graph G in $G(n, p)$ does not even contain a weak (α, β, γ)-community for every $\beta > 0$ and all $\gamma > 0$.*

PROOF. It is well known that for $p = \omega(n) \ln n/n$, with high probability, every vertex in G has degree around $(n-1)p \sim \omega(n) \ln n$ (see p.129 of [3]), i.e., $deg(v) \sim \omega(n) \ln n$ for all vertices v. We will assume this property to hold in the following proof.

Now we consider a subset $S \subset V$ with $|S| = k \leq n/2$. We will show that with high probability, every such S has conductance $\Phi(S)$ at least δ, for a sufficiently small constant δ.

The expected number of edges $e(S, \bar{S})$ between S and its complementary $\bar{S}$ is

$$E[e(S, \bar{S})] = k(n-k)p \geq k\omega(n) \ln n/2.$$

If the conductance $\Phi(S)$ is smaller than δ, then $e(S, \bar{S}) < \delta vol(S) \sim \delta k\omega(n) \ln n$. Using the Chernoff bound (see eg. [35]), we have

$$\Pr[e(S, \bar{S}) < \delta k\omega(n) \ln n] \quad \leq \quad e^{-c_1 k\omega(n) \ln n},$$

[1] $\omega(1)$ means any slowly growing function. This condition ensures that a meaningful community can not have too small size.

[2] We note that Mishra et al. [33] have also given a definition to measure clustering and they also used the notation of (α, β)-clusters. Their definition needs precise bounds on both the number of intra- and inter- edges of a set, and thus is very different from ours.

[3] *with high probability* means that some event occurs with probability at least $1 - o(1)$; *almost every* means that at least $1 - o(1)$ fraction of vertices.

for some constant c_1.

The probability that there exists a subset $S \subset V$ with $|S| = k \leq n/2$ and $\Phi(S) < \delta$ is at most

$$\sum_{k=1}^{n/2} \binom{n}{k} e^{-c_1 k\omega(n)\ln n} \leq \sum_{k=1}^{n/2} e^{(1-c_1\omega(n))k\ln n} = o(1).$$

Therefore, for each set S with size at most $n/2$, the conductance $\Phi(S)$ is no less than δ. In particular, for any $\gamma > 0$, a set of size $O((\ln n)^\gamma)$ has conductance no smaller than some constant, which concludes the proof. □

3.2 Preferential Attachment model

Barabási and Albert proposed the preferential attachment (PA) scheme to reproduce the property that the vertex degrees follow a power law distribution in many real networks. This model has since then been extensively studied. In particular, Mihail, Papadimitriou and Saberi [32] showed that with high probability the graph from the preferential attachment model (which is a small variant of the original model) has constant expansion and constant conductance.

The model in [32] is based on the following random graph process. At time $t = 1$, the graph G'_1 equals a mini-vertex x_1 with a self-loop. At time $t \geq 2$, a new mini-vertex x_t arrives and chooses a mini-vertex $x_{t'}$ $(t' < t)$ in G'_{t-1} with probability proportional to the degree of $x_{t'}$. G'_t is constructed by adding edge $(x_t, x_{t'})$ to G'_{t-1}. Now if we stop at time dn (for some parameter d) and obtain G'_{dn}, then we contract every d consecutive mini-vertices $x_{d\tau-i}, 0 \leq i \leq d-1$ into a corresponding vertex x_τ. The final graph is denoted by $G_{d,n}$.

The following result is an immediate corollary of the main theorem in [32].

THEOREM 2. *[32] With high probability, for a graph $G_{d,n}$ in the preferential attachment model and $d \geq 2, 0 < \beta \leq 2$, there is no strong (or weak) (α, β)-community in $G_{d,n}$.*

3.3 Geometric Preferential Attachment model

As we have seen, the preferential attachment scheme generates graphs with constant expansion with high probability, which is indeed the case in many real networks. However, Bladford et al. [6] and Estrada [15] provide evidence that in some real networks the expansion (in many cases, also the conductance) is not bounded below by a constant. This motivates Flaxman et al. [17, 18] to define a class of geometric preferential attachment (GPA) models, which does not only contain sets with small expansion but also preserve the power law degree distribution. We will show that the GPA model also contains good communities.

The model is defined on the surface S of a sphere in R^3 of radius $\frac{1}{2\sqrt{\pi}}$ (so that $area(S) = 1$). Let $B_d(u) = \{x \in S : |x - u| \leq d\}$, where $|\cdot|$ denote the angular distance between two points on the surface of the sphere, i.e. $B_d(u)$ denotes the spherical cap of angular radius d around u on S. Let $A_d = area(B_d(u))$, for any $u \in S$.

At time 0, the initial graph G_0 equals the empty graph. At time $t \geq 1$, a vertex x_t is generated uniformly at random in S. Then x_t chooses m neighbors $\{y_i\}, 1 \leq i \leq m$ according to some distribution on the set of vertices near x_t. G_t is formed by adding these m new edges $(x_t, y_i), 1 \leq i \leq m$ to G_{t-1}. Specifically, let $V_{t-1}(x_t)$ be the set of vertices that are in G_{t-1} and within angular distance at most $r = n^{\rho-1/2}$ (here $0 < \rho < 1/2$) from x_t, and let $D_{t-1}(x_t) = \sum_{v \in V_{t-1}(x_t)} \deg_t(v)$. Then, for any vertex $u \in V_{t-1}(x_t)$, the probability that y_i (for $1 \leq i \leq m$) equals u is

$$\Pr[y_i = u] = \frac{deg_{t-1}(u)}{\max\{D_{t-1}(x_t), \alpha m A_r t\}},$$

and y_i may also equals x_t, with probability

$$\Pr[y_i = x_t] = 1 - \frac{D_{t-1}(x_t)}{\max\{D_{t-1}(x_t), \alpha m A_r t\}}.$$

Flaxman et al. showed that with high probability, the graph G_n generated from the above process has a power law degree distribution and contains some large set with small expansion. They also showed that when $m \geq K \ln n$ for K sufficiently large, the graph is connected.

Concerning the community structure, we have the following result.

THEOREM 3. *If $m \geq K \ln n$, where K is some sufficiently large constant, for G_n generated from the GPA model, with high probability, each vertex in G_n is contained in a strong (α, β)-community of size n^ϵ, where $0 < \beta, \epsilon < 1/2$.*

PROOF. Since $m \geq K \ln n$, using Lemma 6 in [17], we can guarantee that the community is connected.

Let $G_n = (V_n, E_n)$ and for each vertex $v \in V_n$, let $C_d(v)$ be the set of all vertices in V_n within angular distance at most d from v. Namely, $C_d(v) = V_n \cap B_d(v)$. We will show that for suitable choice of d, $C_d(v)$ is a good community with high probability. Here, we will assume $r \leq d = o(1)$.

For any u, the area of $B_d(u)$ is

$$A_d = 2\pi * \frac{1}{2\sqrt{\pi}} * \frac{1 - \cos d}{2\sqrt{\pi}} \sim \frac{d^2}{4}.$$

Note that a vertex $u \in C_d(v)$ can only connect vertices within angular distance r from it. Therefore, the neighbors of $C_d(v)$ belong to the strip within distance r of the boundary of $B_d(v)$. Let $Str1 = B_{d+r}(v) \backslash B_d(v), Str2 = B_d(v) \backslash B_{d-r}(v)$ and $T1 = V_n \cap Str1, T2 = V_n \cap Str2$. Then the edges between $T1$ and $T2$ form the edge set between $C_d(v)$ and $V_n \backslash C_d(v)$.

The areas of the two strips are

$$area(Str1) = A_{d+r} - A_d \sim \frac{r^2 + 2rd}{4},$$
$$area(Str2) = A_d - A_{d-r} \sim \frac{2rd - r^2}{4},$$

respectively.

Now let $d = n^{\delta-1/2}, \rho < \delta < 1/2$. Since each vertex u is generated uniformly and independently on S, the probability that $u \in B_d(v)$ is $A_d \sim \frac{d^2}{4} = \frac{n^{2\delta-1}}{4}$ (note that the area of S is 1). Therefore, $E[|C_d(v)|] \sim \frac{n^{2\delta}}{4}$. Using the Chernoff bound, we have that, with probability at least $1 - n^{-3}$,

$$(1-\sigma)\frac{n^{2\delta}}{4} \leq |C_d(v)| \leq (1+\sigma)\frac{n^{2\delta}}{4},$$

where σ is an arbitrarily small constant.

Similarly, we can bound the number of vertices in $T1$ and $T2$ to ensure that, with probability at least $1 - 2n^{-3}$,

$$|T1| \leq (1+\sigma)\frac{3n^{\rho+\delta}}{4}, |T2| \leq (1+\sigma)\frac{3n^{\rho+\delta}}{4}.$$

The number of edges between $T1$ and $T2$ is at most $m(|T1|+|T2|)$. Therefore, with probability at least $1 - 3n^{-3}$, the set $C_d(v)$ contains about $c_0\frac{n^{2\delta}}{4}$ vertices, where $1-\sigma \leq c_0 \leq 1+\sigma$, and the number of edges $e(C_d(v), V_n \backslash C_d(v))$ between $C_d(v)$ and $V_n \backslash C_d(v)$ is at most $2m(1+\sigma)\frac{3n^{\rho+\delta}}{4}$. Noting that $\mathrm{vol}(C_d(v)) \geq m|C_d(v)|$, we have

$$\begin{aligned}\Phi(C_d(v)) &\leq \frac{2m(1+\sigma)\frac{3n^{\rho+\delta}}{4}}{mc_0\frac{n^{2\delta}}{4}} \\ &= \Theta(\frac{1}{n^{\delta-\rho}}).\end{aligned}$$

Now if we set $\delta = \frac{\epsilon}{2}, \rho = (\frac{1}{2} - \beta)\epsilon$, then with probability at least $1 - 3n^{-3}$, both $|C_d(v)| = \Theta(n^\epsilon)$ and $\Phi(C_d(v)) = \frac{1}{|C_d(v)|^\beta}$ hold.

By the union bound, with probability at least $1 - 1/n$, every vertex v in V_n is contained in a community $C_{n^{\frac{\epsilon}{2}-\frac{1}{2}}}(v)$, which has size $\Theta(n^\epsilon)$ and community exponent $\beta < 1/2$. □

The geometric preferential attachment model has been extended to general models in which all the nice properties, that is, the small diameter property, the power law degree distribution, and the small community phenomena are satisfied simulatuously[4].

3.4 Ravasz-Barabási Hierarchical model

Ravasz and Barabási [39] constructed a model which not only has the power law degree distribution, but satisfies the property that the clustering coefficient decays in a characteristic manner. The later property characterizes the hierarchical feature of networks. The model (we call it Ravasz-Barabási Hierarchical model) is introduced as follows:

Initially at time $t = 1$, the graph G_1 equals a complete graph K_5, in which one of the nodes is marked *center* and the other four nodes are marked *peripheral nodes*. At time $t > 1$, suppose that we have constructed G_{t-1}, denoted by O_{t-1}. Then we first create four new copies of G_{t-1}, say $N^i_{t-1}, 1 \leq i \leq 4$, and then connect all the peripheral nodes in $N_{t-1} = \cup_i N^i_{t-1}$ to the center in O_{t-1}. This finishes the construction of G_t. We define the center node of G_t to be the center node of O_{t-1}, and the peripheral nodes of G_t to be all the peripheral nodes in the N_{t-1}.

A stochastic version of the hierarchical model can also be defined (see also [39]) if we modify the above process in the following way. At time $t = 1$, the graph G_1 is again the complete graph K_5. At time $t > 1$, we also denote the obtained G_{t-1} by O_{t-1}. Then we first create four new copies of G_{t-1}, say $N^i_{t-1}, 1 \leq i \leq 4$ and from each copy we randomly pick a p^{t-1} fraction of nodes (without replacement) to be the peripheral nodes. Each of the peripheral nodes in N^i_{t-1} then independently chooses a neighbor in O_{t-1} and connects an edge to the neighbor. More specifically, for a peripheral node v, it connects a node u from O_{t-1} with probability proportional to the degree of u. This finishes the construction of G_t. We define the peripheral nodes of G_t to be all the peripheral vertices from $N_{t-1} = \cup_i N^i_{t-1}$.

Note that in the stochastic version of the model, if $p < \frac{1}{5}$, then the number of peripheral nodes in a given step is smaller than 1 and the graphs generated from the model are just unconnected pieces of the complete graph K_5, which gives a trivial case. Therefore, we will restrict to the case that $p > \frac{1}{5}$ in the following argument.

We can show that in the deterministic model, *almost every* vertex is contained in some small community and that in the stochastic version, *every* vertex is contained in a small community. Therefore, we conclude that the small community phenomenon appears in the Ravasz-Barabási Hierarchical model.

Now given a graph or module G, we will use $V(G)$ and $E(G)$ to denote the vertex set and edge set of G, respectively.

THEOREM 4. *For a graph G_t generated from the deterministic Ravasz-Barabási Hierarchical model, almost every node is contained in a strong (α, β, γ)-community for some constants $\alpha, \beta, \gamma > 0$.*

PROOF. Let v_i, e_i, pv_i denote the number of vertices, edges and peripheral vertices of graph G_i, respectively. From the definition, we have $n_i = 5^i, pv_i = 4^i, e_i = 5e_{i-1} + pv_i, e_1 = 10$. We can solve e_i to get $e_i = 5^i(2 + \frac{16(1-(\frac{4}{5})^{i-1})}{5}) = c_1 * 5^i$.

Now $|V(G_t)| = v_t = 5^t$. Let $t_0 = c\log_5 t$ for some sufficiently large constant $c > 1$. First, we show that for a vertex born before time t_0, all vertices other than those born too early in G_{t_0} are contained in their corresponding small communities. A key observation is that for a vertex $v \neq x_1$ born at time i, it will have no connection with vertices born after time i, where x_1 is the center of G_1.

Specifically, if a node v is born before time $\ln t_0$, then we will call v bad and we will not find a community for such a node. If a node v is born at time i such that $\ln t_0 \leq i \leq t_0$, then it must be contained in a new copy of G_{i-1}, i.e., $v \in N^j_{i-1}$ for some $j = 1, 2, 3$ or 4. We denote this copy by $C(v) = N^j_{i-1}$ and we will show that $C(v)$ is a good community containing v. We note that $|C(v)| = 5^{i-1} \leq 5^{t_0} = (\log_5 n)^c$ and that the number of edges inside and outside of $C(v)$ is $c_1 * 5^{i-1}$

[4]Angsheng Li and Pan Peng, The small community phenomena in networks, to appear.

and 4^{i-1}, respectively. Thus

$$\begin{aligned}\Phi(C(v)) &= \frac{4^{i-1}}{2*c_1*5^{i-1}+4^{i-1}}\\ &= O(\frac{1}{(\frac{5}{4})^{i-1}})\\ &= O(\frac{1}{|C(v)|^{\log_5 \frac{5}{4}}}).\end{aligned}$$

For $i > t_0$, each node $v \in V(G_i)$ is contained in a unique copy $C_1(v)$ of G_{t_0}. We will call such a copy the basic module in G_i. For a module $M = C_1(v)$, the number of vertices and edges in it are $|V(C_1(v))| = 5^{t_0}$ and $e(C_1(v)) = c_1 * 5^{t_0}$, respectively.

Given such a module M, we treat it as the product of a new process that starts at K_5 with the center node c of M, then new vertices and edges come in by exactly the same rule as the one of G_t. As a consequence, M is the graph obtained from the new process at time t_0. Similarly, we define the vertex that born before time $\ln t_0$ in the new process to be bad. Note that the number of bad vertices in M is $5 + 5^2 + \cdots + 5^{\ln t_0 - 1} < 5^{\ln t_0}$.

Now if the center c of M is connected to some other vertices outside of the M, then for $v \in V(M)$ that is not bad in M, we define $C(v) = C'(v)$, where $C'(v)$ is the analogous community as we defined for G_{t_0} as above. Namely, if we treat M as the output of the new process and v is born at time i such that $\ln t_0 \le i \le t_0$, then it must be contained in a new copy of G_{i-1}, i.e., $v \in N_{i-1}^j$ for some $j = 1, 2, 3$ or 4. We denote this copy by $C(v) = N_{i-1}^j$. By the above calculation, we know that $C(v)$ is a good community containing v.

If the center c of a given module M is not connected to any vertices outside of M, i.e. c only connects vertices inside M, then for $v \in V(M)$, we define $C(v) = M = C_1(v)$. Now the number of edges out of $C(v)$ is at most $(t - t_0)4^{t_0}$. Thus,

$$\begin{aligned}\Phi(C(v)) &= \frac{(t-t_0)4^{t_0}}{2*c1*5^{t_0}+(t-t_0)4^{t_0}}\\ &= O(\frac{1}{t^{c(1-\log_5 4-1/c)}})\\ &= O(\frac{1}{|C(v)|^{1-\log_5 4-1/c}}).\end{aligned}$$

We can always choose large constant $c > 1$ to ensure that $1-\log_5 4-1/c > 0$. Now the number $b(t)$ of bad nodes in G_t can be easily calculated by induction. Since $b(t_0+1) < 5^{\ln t_0}$ and $b(i+1) = 5b(i)$, we have $b(t) < 5^{t-t_0+\ln t_0}$. Thus, the fraction of bad nodes in G_t is $\frac{5^{t-t_0+\ln t_0}}{5^t} = 5^{-t_0+\ln t_0} = o(1)$.

Therefore, for $c > 1$ large enough, $1 - o(1)$ fraction of nodes in G_t are contained in their own corresponding small communities that are strong (α, β, γ)-communities, where $0 < \beta \le 1 - \log_5 4 - 1/c$ and $(\ln n)^\gamma = (\log_5 n)^c$. □

The above analysis can also be adapted to the stochastic version of the model.

THEOREM 5. *Assume that $\frac{1}{5} < p < 1$. For a graph G_t generated from the stochastic Ravasz-Barabási hierarchical model, with high probability, every node is contained in a strong (α, β, γ)-community for some constants $\alpha, \beta, \gamma > 0$.*

PROOF. As in the previous proof, if we let v_i, e_i and pv_i denote the numbers of vertices, edges and the peripheral vertices of the graph G_i, then $v_i = 5^i, pv_i = (5p)^i, e_i = 5e_{i-1} + 4*(5p)^{i-1}$, from which we have $e_i = 5^i(\frac{6}{5} + \frac{4(1-p^i)}{5(1-p)}) = c_1 * 5^i$.

Now $n = |V(G_t)| = v_t = 5^t$. Let $t_0 = c \log_5 t$ for some sufficiently large constant $c > 1$. For $t_0 < i \le t$, we will again treat the copy of G_{t_0} as the basic module. Noting that each $v \in V(G_i)$ is contained in a unique module $C(v)$, in which the numbers of vertices and edges are $|V(C(v))| = 5^{t_0} = (\log_5 n)^c$ and $e(C(v)) = c_1 * 5^{t_0}$ respectively. We will show that $C(v)$ is a good community containing v.

Note that once a module $M = C(v)$ is formed in the process, then the edges inside the module will not change by definition. However, the number of edges $e(M, V(G_i)\backslash M)$ between the module M and $V(G_i)\backslash M$ may increase as i grows from $t_0 + 1$ to t. We will bound $e(M, V(G_t)\backslash M)$ by showing that in each step, the number of newly formed edges coming out of M is small.

CLAIM 1. *If c is large enough and $t_0 < i \le t$, then with probability $1 - \frac{i}{t^c}$, for every module M in G_i, we have that*

$$e(M, V(G_i)\backslash M) \le c_2(i - t_0)(5p)^{t_0},$$

for some large constant c_2.

PROOF. We prove the Claim by induction on i.

For $i = t_0 + 1$, G_{t_0+1} contains five modules, namely, four periphery modules and one central module. For a periphery module M and the central module M', $e(M, V(G_{t_0+1})\backslash M) = (5p)^{t_0}$ and $e(M', V(G_{t_0+1})\backslash M') = 4 * (5p)^{t_0}$, respectively. Thus, if $c_2 \ge 4$, then the Claim holds for $i = t_0 + 1$.

Suppose by induction that the Claim holds for all i with $t_0 < i \le j$. Let $i = j + 1$.

By definition, G_{j+1} is composed of four new copies $\{N_j^i\}_{i=1}^4$ and one old copy O_j of G_j. Assume that M is an arbitrary module in G_{j+1}.

If M is contained in N_j^i for some i with $1 \le i \le 4$, then

$$e(M, V(G_{j+1})\backslash M) = e(M, V(N_j^i)\backslash M) + e(M, O_j),$$

where $e(M, O_j)$ is the number of edges between M and O_j. Noting that $e(M, O_j)$ is also the number of nodes being chosen in M at time $j + 1$. Thus, $e(M, O_j) \sim H(5^j, 5^{t_0}, (5p)^j)$, where $H(A, B, C)$ is the hypergeometric distribution with parameters A, B and C. Therefore, by the concentration inequality on hypergeometric distribution (see eg. [11]), with probability at most $\frac{1}{2^{c_2(5p)^{t_0}}}$,

$$e(M, O_j) \ge c_2(5p)^{t_0}. \quad (1)$$

If c and c_2 are sufficiently large, then (1) holds with probability at most $\frac{1}{2*4*5^{j-t_0}(t^c)}$.

Thus, with probability at least $1-\frac{1}{2t^c}$, for every module M in $N_j = \cup_{i=1}^4 N_j^i$, $e(M, O_j) \leq c_2(5p)^{t_0}$ holds.

If M is contained in O_j, then

$$e(M, V(G_{j+1})\backslash M) = e(M, V(O_j)\backslash M) + e(M, N_j),$$

where $e(M, N_j)$ is the number of edges between M and N_j. Noting that $e(M, N_j)$ is also the number of nodes being chosen in M at time $j+1$. By induction, a picked node in N_j chooses a neighbor in M with probability at most $p_j = \frac{2*c_1*5^{t_0}+c_2(j-t_0)(5p)^{t_0}}{2*c_1*5^j} = (1+o(1))5^{t_0-j}$, so $e(M, N_j)$ is dominated by $\mathrm{Bi}(4*(5p)^j, p_j)$, where $\mathrm{Bi}(n, p)$ denotes the binomial distribution with parameters n and p. Therefore, by the Chernoff bound, with probability at most $\frac{1}{2^{c_2(5p)^{t_0}}}$,

$$e(M, N_j) \geq c_2(5p)^{t_0}, \tag{2}$$

if c and c_2 are sufficiently large, then (2) holds with probability at most $\frac{1}{2*5^{j-t_0}(t^c)}$.

Thus, with probability at least $1-\frac{1}{2t^c}$, for every module M in O_j, $e(M, N_j) \leq c_2(5p)^{t_0}$ holds.

Therefore with probability at least $1-\frac{1}{t^c}$, for every module M in G_{j+1}, the number of newly formed edges incident to M is no more than $c_2(5p)^{t_0}$. By induction, we have that with probability at least $1-(\frac{j}{t^c}+\frac{1}{t^c}) = 1-\frac{j+1}{t^c}$,

$$e(M, V(G_{j+1})\backslash M) \leq c_2(j-t_0)(5p)^{t_0}.$$

□

By using the above Claim, we have that with probability at least $1-\frac{1}{t^{c-1}}$, all basic modules $C(v)$ in G_t have conductance value:

$$\begin{aligned}\Phi(C(v)) &\leq \frac{c_2 t*(5p)^{t_0}}{2c_1*5^{t_0}+c_2 t*(5p)^{t_0}} \\ &= O(\frac{t^{1+c\log_5 5p}}{t^c}) \\ &= O(\frac{1}{|C(v)|^{\log_5 \frac{1}{p}-\frac{1}{c}}}).\end{aligned}$$

Therefore, when c is large enough, with high probability, every vertex v in G_t is contained in a strong (α, β, γ)-community which is also the basic module containing v, where $0 < \beta \leq \log_5 \frac{1}{p} - \frac{1}{c}$ and $(\ln n)^\gamma = (\log_5 n)^c$.

4. PERTURBED GRAPHS

In this section, we will consider the community structure of a graph (a perturbed graph) in which "randomness" and "structure" are combined in a more natural way. Specifically, a perturbed graph G is composed of a base graph $\bar{G}$ and a random graph R, which is defined on the vertex set of $\bar{G}$. For example, the small world model of Kleinberg [26] is a perturbed graph with $\bar{G}$ and R representing the d-dimensional grid and a random graph on the grid, respectively. Here R is constructed in the following way: let $d(u, v)$ denote the l_1 norm on the grid. Each vertex u chooses an out-contact v with probability proportional to $d(u, v)^{-r}$, where $r \geq 0$ is some parameter, and a directed edge from u to v is added to the graph.

We will also consider another question which naturally arises from our definition of community. Intuitively, a graph which has small community phenomenon contains many sets with small conductance. On the other hand, an expander is a graph with all sets having large conductance. Thus, it is interesting to explore the relationship between these two properties. Here we show that for some particular model, with high probability, under certain conditions it is an expander; while under some other conditions, it has the small community phenomenon.

4.1 d-dimensional small world model

Flaxman [16] studied the edge expansion on several classes of the perturbed graphs. Particularly, he showed that with high probability, for $r < d$, the expansion of the small world model is greater than some small constant; for $r = d$ the expansion is $o(1)$. We refine his analysis to show that as r changes, the small community phenomenon appears. In fact, there exists a threshold result of the small community phenomenon in the small world model.

THEOREM 6. *(Threshold Theorem of the Small Community Phenomena) In the d-dimensional small world model G, with high probability, when $r < d$, there is no proper community for an arbitrary node; when $r = d$, there exists weak (α_1, β_1)-communities of size $\frac{n}{(\ln n)^{c_1}}$ for every node, where $\beta_1 < 1, c_1 > 0$ and there exists weak $(\alpha_2, 1)$-communities of size $c_2 n$ for every node, where $0 < c_2 \leq \frac{1}{4}$; when $r > d$, there exists strong (α, β, γ)-communities for every node for some constants α, β, γ.*

PROOF. We first look at the 1-dimensional small world model. Namely, we consider the perturbed graph $G = \bar{G} + R$, where $\bar{G}$ is a cycle on n vertices, and R is the random graph on the same n vertices and each vertex chooses an out-contact j with probability proportional to $d_{i,j}^{-r}$. Specifically, if we set $Z = \sum_{k \neq i} d_{i,k}^{-r}$, then in R the probability that there is an arc from i to j is $\frac{d_{i,j}^{-r}}{Z}$, where $r > 0$ is the parameter of this model.

We divide the proof into two cases.

- $r < 1$.

  In this case, Flaxman [16] has proved that the expansion of G is greater than some small constant δ with high probability. Therefore for every S satisfies $|S| \leq \frac{n}{2}$, $e_G(S, \bar{S}) \geq \delta|S|$ holds. By using the fact that $e_G(S) \leq 3|S|$, we have $\Phi(S) \geq c_0$, for some constant c_0. Therefore, there is no proper community for an arbitrary node.

- $r \geq 1$.

  Now for a vertex v, we define $C_k(v)$ to be the set of vertices within distance at most k from v, i.e. $C_k(v) = \{j : d(v, j) \leq k\}$, where $k \leq \frac{1}{4}n$ will be specified later. We show that $C_k(v)$ is indeed a good community with respect to its size k. When there is no confusion, we will use C to denote $C_k(v)$ for simplicity.

  It is obvious that $e_{\bar{G}}(C) = 2k, e_{\bar{G}}(C, \bar{C}) = 2$ and $0 \leq e_R(C) \leq 2k+1$. We only need to estimate $e_R(C, \bar{C})$.

For $i \in C$, let $X_{i,C}$ and $X_{i,\bar{C}}$ denote the indicator random variables of the events that i has chosen its out-contact in C and $\bar{C}$, respectively. For $j \in \bar{C}$, let $X_{j,C}$ and $X_{j,\bar{C}}$ denote the indicator random variables of the events that j has chosen its out-contact in C and $\bar{C}$, respectively. In addition, let X_{ij} be the indicator random variable of the event that i has chosen j as its out-contact.

Now let e_{R1} be the number of random arcs from C to $\bar{C}$. Such an arc is formed by some vertex i in C choosing its out-contact j in $\bar{C}$. We also let e_{R2} be the number of random arcs from $\bar{C}$ to C. Thus, $e_R(C,\bar{C}) = e_{R1} + e_{R2}$. We analyze e_{R1} and e_{R2} separately.

1. For $r = 1$, $Z = \sum_{k\neq i} d_{i,k}^{-1} = \Theta(\ln n)$.
   Firstly. we can calculate the expectation of e_{R1} as follows:

$$\begin{aligned} E[e_{R1}] &= \sum_{i\in C} E[X_{i,\bar{C}}] = \sum_{i\in C}\sum_{j\in\bar{C}} E[X_{ij}] \\ &= \sum_{i\in C}\sum_{j\in\bar{C}} \frac{d^{-1}(i,j)}{Z} \\ &= \Theta(\frac{1}{Z}\sum_{i=1}^{k}(\sum_{j=i}^{n/2}\frac{1}{j} + \sum_{j=2k+2-i}^{n/2}\frac{1}{j})) \qquad (3) \\ &= \Theta(\frac{1}{\ln n}(2k\ln\frac{n}{2} - \sum_{i=1}^{k}\ln i(2k+2-i))) \\ &= O(\frac{k}{\ln n}\ln\frac{n}{2k}). \end{aligned}$$

   Now since the set of random variables $\{X_{i,\bar{C}}\}_{i\in C}$ are independent $0,1$ random variables, by the Chernoff bound, we know that e_{R1} concentrates around its expectation when k is large. Specifically, for any $c_1 > 0, 0 < c_2 \leq \frac{1}{4}$ and $\frac{n}{(\ln n)^{c_1}} \leq k \leq c_2 n$, with probability at most $o(1/n)$, $e_{R1} > \frac{c'k}{\ln n}\ln\frac{n}{2k}$ for some constant c'.
   Similar results for e_{R2} can be obtained. Actually, $E[e_{R2}]$ is the same as $E[e_{R1}]$ by the symmetry of $d(\cdot,\cdot)$. Then with probability at least $1 - o(1/n)$,

$$e_R(C,\bar{C}) = e_{R1} + e_{R2} \leq \frac{2c'k}{\ln n}\ln\frac{n}{2k}.$$

   Now we know that $e_G(C) = \Theta(k)$, $e_G(C,\bar{C}) = O(\frac{k}{\ln n}\ln\frac{n}{2k})$, by using which we can estimate the conductance of C by definition.
   If $k = \frac{n}{(\ln n)^{c_1}}$, then with probability at least $1 - o(1/n)$,

$$\begin{aligned} \Phi(C) &\leq \frac{e_G(C,\bar{C})}{2e_G(C) + e_G(C,\bar{C})} \leq \frac{\frac{2c'k}{\ln n}\ln\frac{n}{2k}}{2k + \frac{2c'k}{\ln n}\ln\frac{n}{2k}} \\ &= O(\frac{\ln\ln n}{\ln n}) = O(\frac{1}{(\ln|C|)^{\beta}}), \end{aligned}$$

   where β is an arbitrary constant with $0 \leq \beta < 1$.
   If $k = c_2 n$, then with probability at least $1 - o(1/n)$,

$$\Phi(C) = O(\frac{1}{\ln|C|}).$$

   Thus, with probability $1-o(1)$, every vertex in the graph is contained in a weak (α,β)-community of size $\frac{n}{(\ln n)^{c_1}}$, where $0 \leq \beta < 1$; it is also contained in a weak $(\alpha,1)$-community of size $c_2 n$.

2. For $r > 1$, the calculations are almost the same. We only need to notice that in this case $Z = \sum_{k\neq i} d_{i,k}^{-r} = \Theta(1)$ and that in equation (3) we should replace $\frac{1}{j}$ by $\frac{1}{j^r}$. In so doing,

$$\begin{aligned} E[e_{R1}] &= O\Big(\sum_{i=1}^{k}(\sum_{j=i}^{n/2}\frac{1}{j^r} + \sum_{j=2k+2-i}^{n/2}\frac{1}{j^r})\Big) \\ &= O\Big(\sum_{i=1}^{k}(i^{1-r} + (2k+2-i)^{1-r})\Big) \\ &= \begin{cases} O(k^{2-r}) & \text{if } 1 < r < 2 \\ O(\ln k) & \text{if } r = 2 \\ O(1) & \text{if } r > 2\ . \end{cases} \end{aligned}$$

   As a result, if $1 < r < 2$, and $k = c_3(\log n)^{\frac{1}{2-r}}$ for some large constant c_3, then with probability at least $1 - o(1/n)$,

$$\Phi(C) = O(\frac{\log n}{k}) = O(\frac{1}{(\log n)^{\frac{r-1}{2-r}}}) = O(\frac{1}{|C|^{r-1}}).$$

   If $r \geq 2$, and set $k = (\log n)^{c_4}$ for an arbitrary constant $c_4 > 1$, then with probability at least $1 - o(1/n)$,

$$\Phi(C) = O(\frac{\log n}{k}) = O(\frac{1}{(\log n)^{c_4-1}}) = O(\frac{1}{|C|^{1-\frac{1}{c_4}}}).$$

   Thus, with probability $1 - o(1)$, for $r > 1$, every vertex v in the graph is contained in a strong (α,β,γ)-community, which is the set of vertices not far from v.

In conclusion, with high probability, when r is in the range $[0,1)$, there is no proper community in the graph; when $r = 1$, every vertex is contained in some large and weak communities. Finally, when r grows to be larger than 1, small strong communities appear for every node.

For $d \geq 2$, the proof is almost the same as above: we also need to define $C_k(v) = \{u : d(u,v) \leq k\}$ for appreciate k. Noting that in the case of d dimensional model, $C_k(v)$ contains about $\Theta(k^d)$ nodes and the boundary of $C_k(v)$ contains about $\Theta(k^{d-1})$ nodes, we can easily verify the corresponding results. □

4.2 A generalized perturbed graph

Flaxman [16] also showed that a perturbed graph $G = \bar{G} + R$ with $\bar{G}$ being an arbitrary connected graph and R a uniformly random mapping on $V(\bar{G})$. Specifically, each $v \in V(\bar{G})$ independently chooses a neighbor u uniformly at random from the vertex set and connects to u. The resulted graph has constant edge expansion, with high probability.

Now we generalize the definition of R (generalized random mapping) in the following way. We introduce a parameter q which is the probability for a vertex to choose itself as its

neighbor (i.e. a loop is formed). We define the probability for a vertex u to choose a vertex v ($v \neq u$) as its neighbor to be $p = \frac{1-q}{n-1}$. It is easy to see that if $q = \frac{1}{n}$, then R corresponds to the uniformly random mapping.

In the following, we will specify the base graph $\bar{G}$ to be the n-node cycle and R to be the generalized random mapping on $\bar{G}$. We show that for q varying from 0 to 1, the structure of the network is changing. Intuitively speaking, if q is small, then the conductance of small subset of $G = \bar{G}+R$ is at least some constant and G does not have small community; if q is large, then small community appears.

THEOREM 7. *If $q < \frac{1}{n_1^{\sigma}}$ for some constant $\sigma_1 < 1$, then with high probability, every subset S with size $|S| \leq \epsilon n$, where ϵ is an arbitrarily small constant, has conductance larger than some constant; if $q > 1 - \frac{1}{\ln n}$, then with high probability, every vertex is contained in a strong $(\alpha, 1, 1)$-community.*

PROOF. By using the Chernoff bound, it is easy to see that with high probability, the degree of every vertex is upper bounded by a constant c', which means that for a set $S \subset V(G)$, the volume of S satisfies $|S| \leq \mathrm{vol}(S) \leq c'|S|$. Thus, to show that some set S has low conductance, it suffices to bound the probability that $e_R(S, \bar{S}) \leq \delta$, for some sufficiently small constant $\delta < \frac{1}{10}$.

Flaxman [16] showed that the probability that there exists some set S with $|S| = s$ and $e_R(S, \bar{S}) \leq \delta|S|$ is at most

$$P1 = n(\frac{ne}{\delta s})^{2\delta s} \Pr[e_R(S, \bar{S}) \leq \delta s].$$

Now we consider all sets of size no more than ϵn for a small constant ϵ. For a set S such that $|S| = s \leq \epsilon n$, we know that

$$\begin{aligned} \Pr[e_R(S, \bar{S}) \leq \delta s] &\leq \binom{s}{\delta s}(q + (s-1)p)^{s-\delta s} \\ &\leq ((\frac{e}{\delta})^{\delta}(\frac{s-1}{n-1} + q\frac{n-s}{n-1})^{1-\delta})^s. \end{aligned}$$

If $s_0 = \frac{3}{1-3\delta} \leq s \leq \epsilon n$, then for $q < \frac{1}{n^{\sigma_1}}$, where $\sigma_1 = \frac{5-3\delta}{6(1-\delta)} < 1$, we have

$$\begin{aligned} P1 &\leq n((\frac{ne}{\delta s})^{2\delta}(\frac{e}{\delta})^{\delta}(\frac{s-1}{n-1} + q\frac{n-s}{n-1})^{1-\delta})^s \\ &\leq n((\frac{ne}{\delta s})^{2\delta}(\frac{e}{\delta})^{\delta}(\frac{2}{n^{\sigma_1}})^{1-\delta})^s \\ &\leq n(\frac{(\frac{e}{\delta})^{3\delta}2^{1-\delta}}{s^{2\delta}n^{\sigma_1(1-\delta)-2\delta}})^s \\ &= n(\frac{(\frac{e}{\delta})^{3\delta}2^{1-\delta}}{s^{2\delta}n^{\frac{5-15\delta}{6}}})^s. \end{aligned}$$

Let $f(s) = (\frac{c}{s^{2\delta}n^{\frac{5-15\delta}{6}}})^s = e^{s(\ln c - 2\delta \ln s - \frac{(5-15\delta)\ln n}{6})}$, where $c = c(\delta) = (\frac{e}{\delta})^{3\delta}2^{1-\delta}$. For n sufficiently large and δ small enough, the derivative of f is

$$f'(s) = f(s)(\ln c - 2\delta \ln s - \frac{(5-15\delta)\ln n}{6} - 2\delta) < 0.$$

Therefore, we get

$$P1 \leq nf(s_0) = O(n\frac{1}{n^{5/2}}) = o(1/n).$$

Combining the fact that for each S with size $|S| = s < \frac{3}{1-3\delta}$, $e(S, \bar{S}) \geq \delta\frac{3}{1-3\delta}$, we know that the probability that there exists a set of size no more than ϵn and conductance less than δ is $o(1)$. Thus, for $q < \frac{1}{n_1^{\sigma}}$, each small set has constant conductance with high probability.

If $q > 1 - \frac{1}{\ln n}$, then for each vertex v, we define $C(v)$ to be the set of vertices within distance at most $k = \ln n$, where the distance is the l_1 norm on the 1-dimensional grid (i.e., the cycle). Then similar to the proof for the small world model, we can show that the number of edges $e_R(C(v), V\backslash C(v))$ between $C(v)$ and $V\backslash C(v)$ is concentrated around its expectation $E[e_R(C(v), V\backslash C(v))] = \Theta(k(n-k)p)$.

Therefore, we can estimate the conductance of $C(v)$ as

$$\Phi(C(v)) \leq \Theta(\frac{k(n-k)p}{k}) = \Theta((n-k)\frac{1-q}{n-1}) = \Theta(1-q).$$

In particular, if $q = 1 - \frac{1}{\ln n}$, then $\Phi(C(v)) \leq \Theta(\frac{1}{|C(v)|})$. For q larger than this probability, $C(v)$ has conductance even smaller but the best possible $\Phi(C(v))$ is still of order $\Theta(\frac{1}{|C(v)|})$. □

5. CONCLUSIONS

Intuitively speaking, a real large-scale network is a dynamic evolution of sparse graphs, in which a single node or edge is no longer essential. In this case, it is a challenge to define the "basic elements" of a network, leading to a wide range of research on communities in networks. Existing algorithms which are built based on graph partitioning are very successful in finding large communities. However, experience in the human society tells us that small communities exist almost everywhere, that small communities overlap and that small communities play important roles in our society. Given that networks are natural mathematical models for describing relationships of massive objects in many different subjects of both physical and social sciences, it is an important scientific problem to study the functions, roles and mechanisms of small communities of general networks in nature, in industry and in our society.

In this article, we propose a novel approach to define communities in a network, allowing us to study the small community phenomena in some well-defined network models. We show that a number of natural network models satisfy the small community phenomena, which can be regarded as a new feature for a number of networks. The results we proved not only help us to explore and characterize some general properties of real world networks but also have potential applications in validation and controlling of networks.

On the other hand, in the definition of the small community phenomena, the requirement that almost every node belongs to some community may be harsh. It is interesting to study the cases that only a constant fraction of nodes or less belong to some meaning communities in both theoretical models and real world networks.

6. REFERENCES

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multi-scale complexity in networks, 2009. cite arxiv:0903.3178.

[2] C. Allen. Life with alacrity: The dunbar number as a limit to group sizes. 2004.

[3] N. Alon and J. H. Spencer. *The probabilistic method, third edition*. John Wiley, 2008.

[4] F. G. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *Ann. Appl. Prob.*, 7:46–89, 1997.

[5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[6] D. K. Blandford, G. E. Blelloch, and I. A. Kash. Compact representations of separable graphs. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 679–688, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[7] F. Chierichetti, S. Lattanzi, and A. Panconesi. Almost tight bounds for rumour spreading with conductance. *STOC '10:ACM Symposium on Theory of Computing*, 2010.

[8] F. Chierichetti, S. Lattanzi, and A. Panconesi. Rumour spreading and graph conductance. *SODA '10:ACM-SIAM Symposium on Discrete Algorithms*, 2010.

[9] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

[10] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 9:P09008, September 2005.

[11] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[12] R. Dunbar. *Grooming, Gossip and the Evolution of Language*. Harvard Univ Press, Cambridge, MA, 1996.

[13] R. Durrent. *Random graph dynamics*. Cambridge University Press, Cambridge, U.K., 2007.

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

[15] E. Estrada. Spectral scaling and good expansion properties in complex networks. *Europhysics Letters*, 73:649–655, 2006.

[16] A. D. Flaxman. Expansion and lack thereof in randomly perturbed graphs. *Internet Mathematics*, 4(2):131–147, 2007.

[17] A. D. Flaxman, A. Frieze, and J. Vera. A geometric preferential attachment model of networks. *Internet Mathematics*, 3(2), 2007.

[18] A. D. Flaxman, A. M. Frieze, and J. Vera. A geometric preferential attachment model of networks II. *Internet Mathematics*, 4(1):87–111, 2007.

[19] S. Fortunato. Community detection in graphs. *Physics Reports*, 486, 2010.

[20] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, June 2002.

[21] M. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[22] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc. (N.S*, 43:439–561, 2006.

[23] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Natural communities in large linked networks. In *Proc. of KDD'03*, 2003.

[24] P. Jonsson, T. Cavanna, D. Zicha, and P. Bates. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7:2, 2006.

[25] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[26] J. Kleinberg. The small-world phenomenon: an algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on the Theory of Computing*. 2000.

[27] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 57, Washington, DC, USA, 2000. IEEE Computer Society.

[28] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1):016118, Jul 2009.

[29] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):46110, 2008.

[30] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[31] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *CoRR*, abs/0810.1355, 2008. informal publication.

[32] M. Mihail, C. Papadimitriou, and A. Saberi. On certain connectivity properties of the internet topology. *J. Comput. Syst. Sci.*, 72(2):239–251, 2006.

[33] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. Finding strongly knit clusters in social networks. *Internet Mathematics*, 5(1):155–174, 2008.

[34] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[35] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

[36] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb. 2004.

[37] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, jun 2005.

[38] A. Rapoport. Spread of information through a

population with socio-structural basis. *Bulletin of Mathematical Biophysics*, 15:523–543, 1953.

[39] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67:026112, 2003.

[40] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

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