Internet Mathematics Vol. 7, No. 2: 81–106

Community Structures in Classical Network Models

Angsheng Li and Pan Peng

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 as a subset of nodes whose vertices are well connected with other vertices in the set and have relatively fewer connections with vertices outside the set. In contrast to traditional research that focuses mainly on detecting and/or testing such 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 exhibit the *small-community phenomenon*, 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, the small-community phenomenon exists, while in some other models, it does not.

1. Introduction

There are quite a few interesting phenomena that arise in the study of large-scale networks. For example, the degree sequences in many networks obey power-law distributions [Barabási and Albert 99, Mitzenmacher 04], 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* [Watts and Strogatz 98, Kleinberg et al. 00], namely, almost every

ISSN: 1542-7951 print / 1944-9488 online

pair of nodes in the graph are connected by a short path through the network, and in some cases, we can find such a short path efficiently using only local information. Other typical characteristics include the "triadic closure" property [Rapoport 53], the "densification and shrinking diameters" in the evolution of networks [Leskovec et al. 07], and the property of community structures [Fortunato 10], which is the main focus of this paper.

Communities (also called "clusters" or "modules") are naturally thought of as cohesive subgraphs in a network. Informally, vertices in a community are well interconnected with fellow members of the community and have relatively fewer connections with vertices outside the community. Communities appear in a wide range of applications. For instance, in protein–protein interaction (PPI) networks, groups of proteins sharing the same or similar functions are clustered together [Jonsson et al. 06]; in society, the communities may correspond to groups of friends or coworkers [Granovetter 73]; in scientific collaboration networks, scientists who investigate similar research topics or use similar methodologies group together to form communities [Girvan and Newman 02].

Previous research has focused 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 [Hopcroft et al. 03, Girvan and Newman 02]. Due to the many similar characteristics between clustering and graph partitioning, in which spectral techniques work particularly well, spectral algorithms are also used to find clusterings [von Luxburg 07]. Modularity-based methods have been very influential in recent research [Newman and Girvan 04, Danon et al. 05]. Other works may first treat communities from some specific perspective and then utilize that to achieve their specific goals. For example, [Palla et al. 05] views communities as a chain of adjacent cliques, and using this, the authors can find overlapping and/or nesting communities. Testing the quality of a community has also been studied [Lancichinetti et al. 08, Lancichinetti and Fortunato 09]. For more applications and experimental results on community detection, see the recent survey [Fortunato 10].

Though there are extensive studies on finding and testing communities, there is no uniform or standard definition of the term. In fact, many papers on finding clusters do not give a precise definition (mathematically) but do give algorithms that will output cohesive subsets of the nodes of graphs, and then these sets are treated as the communities (e.g., [Girvan and Newman 02, Ahn et al. 09]). Traditional definitions vary greatly depending on the field of research and the investigators' goals. Some definitions involve the global structure of the community—for instance, 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 corresponding quantities in a random model preserving the degree sequence of the original network [Newman and Girvan 04]—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 [Kumar et al. 00, Palla et al. 05]).

In this paper, we introduce a new definition of community. In our definition, a community is allowed to overlap and/or nest other communities. Furthermore, our definition provides a quantitative way to compare the quality of two communities (by comparing the *community components*; see Section 2). This definition uses the concept of the conductance of a subset of the graph, which measures somehow the 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 [Chung 97, Hoory et al. 06]. Some conductance results of random graphs have also been investigated [Durrent 07].

Several papers have appeared that connect conductance to clustering. In [Kannan et al. 04], the authors 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 intercluster edges takes only a small portion of the total edge weight. Their main goal is to analyze a spectral algorithm that gives a good approximation solution to the clustering problem under their definition.

In [Leskovec et al. 2008], the authors use the conductance directly to measure the goodness of a community. A good community is supposed to have low conductance. They considered the quality of network communities at different size scales. Specifically, they studied a quantity that 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 number [Dunbar 96], which predicts that a stable community has size bounded above by approximately 150.

In contrast to many other papers (e.g., [Mishra et al. 08, Kannan et al. 04]) that 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 [Allen 04], which finds that online communities have around 60 members and some other evidence that supports Dunbar's theory on the limit size of a stable community.

The observation of the work of [Leskovec et al. 2008], cited above, 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., [Leskovec et al. 2008]), we combine the conductance of a set and its size in a more refined way that has never been considered before.

We test our definition on a variety of random graph models to check whether they exhibit the small-community phenomenon. Through this line of research we can both determine whether a given model is suitable for validating real-world networks and provide motivation to design more appropriate network models.

We believe not only that our results build a theoretical framework for the study of community structures, but that they have potential applications in understanding other structural properties and/or dynamic behaviors of networks in general. For example, [Chierichetti et al. 10a, Chierichetti et al. 10b] recently established the connection between rumor-spreading on a graph and its conductance. It is known that communities play an important role in rumor-spreading (see, e.g., [Ball et al. 97]), reflecting the intuition that rumor spreads quickly within a community. This experiment needs a mathematical proof, for which our definition of community might well be used.

In Section 2, we will give some basic definitions on good communities and some corresponding quantities, and we will formulate the concept of the small-community phenomenon. In Section 3, we will investigate the small-community phenomenon on a set of classical network models, including the Erdős–Rényi model [Erdős and Rényi 60], the geometric preferential attachment model [Flaxman et al. 07a, Flaxman et al. 07b], and the hierarchical model [Ravasz and Barabási 03]. In Section 4, we consider the community structure of some perturbed graphs, including the small-world model, and show that the 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 the number of edges, we denote it by $\mathrm{vol}(G) = \mathrm{vol}(V) = 2|E|$. For any two vertex subsets $S, T \subset V$, we let $e_G(S, T)$ denote be the number of edges with one endpoint in S and the other in T, while $e_G(S)$ denotes 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 $\Phi(S) = \Phi(\bar{S})$.

In [Leskovec et al. 2008], the conductance of a set S is used to measure the goodness of the community S. As easily seen from the definition, the conductance of a set S provides somehow a measure of the 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 it is that S is a good community. Moreover, it is natural to require a community to be connected, which ensures that every pair of 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 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 the one hand, we are more interested in how communities change as the size of the network grows. On the other hand, a set that is too small can hardly be thought of as a reasonable community [Allen 04]. Moreover, empirical results reveal that the sizes of many communities scale with the size of the associated network (see, e.g., [Palla et al. 05]).

We now extend the idea of [Leskovec et al. 2008] to give a more refined way to measure the goodness of a community.

Definition 2.1. Given a graph $G = (V,E)$ and $\alpha, \beta > 0$, a connected set $S \subset V$ with[1] $|S| = \omega(1)$ 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.*

[1] Here $\omega(1)$ means any slowly growing function. This condition ensures that a meaningful community cannot be too small.

[2] We note that [Mishra et al. 08] has also given a definition to measure clustering, and it also uses the notation of (α,β)-clusters. That definition needs precise bounds on both the number of intra- and interedges of a set, and thus is very different from ours.

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

Definition 2.2. 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 the *community exponent* of the graph. It is easily seen that $0 \leq \beta \leq 2$ in the definition of strong community. Here β captures the quality of a community. Intuitively, for a strong (α, β)-community S, if β is large, then the fraction of edges outside of S that cross the cut is low, which means that S is more like a community. Thus, to some extent, we can say that if $\beta_1 > \beta_2 > 0$, a strong (α_1, β_1)-community is better than a 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 exhibits the small-community phenomenon, i.e., whether every vertex in the graph is contained in some small community.

Definition 2.3. Given a random graph G with vertex set V and $|V| = n$, if with high probability, almost every 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 exhibit the *small-community phenomenon*.[3]

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

Definition 2.4. In a graph $G = (V, E)$, the *expansion* of a subset $S \subseteq V$ is

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

The expansion of the graph $\alpha(G)$ is $\min_{S \subseteq V,\, |S| \leq |V|/2} \alpha(S)$. The graph expansion of some network models is studied in [Flaxman 07, Flaxman et al. 07a, Flaxman et al. 07b] .

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

3. Results on Classical Network Models

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

3.1. The Erdös–Rényi Model

The Erdös–Rényi model [Erdős and Rényi 60] 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 exhibit the small-community phenomenon.

Theorem 3.1. *If $p = \omega(n)\ln n/n$, where $\omega(n) \to \infty$ arbitrarily slowly, then for every $\beta > 0$ and all $\gamma > 0$, a random graph G in $G(n,p)$ with high probability does not contain even a weak (α,β,γ)-community.*

Proof. It is well known that for $p = \omega(n)\ln n/n$, with high probability, every vertex in G has degree around $(n-1)p \approx \omega(n)\ln n$ (see [Alon and Spencer 08, p. 129]), i.e., $\deg(v) \approx \omega(n)\ln n$ for all vertices v. We will assume this property to hold in this proof.

Now we consider a subset $S \subset V$ with $|S| = k \le n/2$. 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 complement $\bar{S}$ is

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

If the conductance $\Phi(S)$ is smaller than δ, then $e(S,\bar{S}) < \delta\,\mathrm{vol}(S) \approx \delta k\omega(n)\ln n$. Using the Chernoff bound (see, e.g., [Mitzenmacher and Upfal 05]), we have

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

for some constant c_1.

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

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

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 [99] 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 et al. 06] shows 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 [Mihail et al. 06] is based on the following random graph process. At time $t = 1$, the graph G'_1 equals a minivertex x_1 with a self-loop. At time $t \geq 2$, a new minivertex x_t arrives and chooses a minivertex $x_{t'}$ $(t' < t)$ in G'_{t-1} with probability proportional to the degree of $x_{t'}$. Then 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 minivertices $x_{d\tau-i}$, $0 \leq i \leq d-1$, 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 [Mihail et al. 06].

Theorem 3.2. [Mihail et al. 06] *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, [Blandford et al. 03] and [Estrada 06] provide evidence that in some real networks, the expansion (in many cases, also the conductance) is not bounded below by a constant. This motivates the definition in [Flaxman et al. 07a, Flaxman et al. 07b] of a class of geometric preferential attachment (GPA) models that not only contains sets with small expansion but also preserves 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 $1/(2\sqrt{\pi})$ (so that area$(S) = 1$). Let $B_d(u) = \{x \in S : |x - u| \leq d\}$, where $|\cdot|$ denotes 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 $A_d = \text{area}(B_d(u))$, for every $u \in S$.

At time 0, the initial graph G_0 is 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. Then 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 equal 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 3.3. *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 [Flaxman et al. 07a, Lemma 6], 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

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

Note that a vertex $u \in C_d(v)$ can connect vertices only 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 $\text{Str}_1 = B_{d+r}(v) \setminus B_d(v)$, $\text{Str}_2 = B_d(v) \setminus B_{d-r}(v)$,

and $T_1 = V_n \cap \mathrm{Str}_1$, $T_2 = V_n \cap \mathrm{Str}_2$. Then the edges between T_1 and T_2 form the edge set between $C_d(v)$ and $V_n \setminus C_d(v)$.

The respective areas of the two strips are

$$\mathrm{area}(\mathrm{Str}_1) = A_{d+r} - A_d \approx \frac{r^2 + 2rd}{4},$$
$$\mathrm{area}(\mathrm{Str}_2) = A_d - A_{d-r} \approx \frac{2rd - r^2}{4}.$$

Now let $d = n^{\delta - 1/2}$, $\rho < \delta < 1/2$. Since each vertex u is generated uniformly and independently on S, the probability of $u \in B_d(v)$ is

$$A_d \approx \frac{d^2}{4} = \frac{n^{2\delta - 1}}{4}$$

(note that the area of S is 1). Therefore,

$$E[|C_d(v)|] \approx \frac{n^{2\delta}}{4}.$$

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 T_1 and T_2 to ensure that with probability at least $1 - 2n^{-3}$,

$$|T_1| \leq (1 + \sigma)\frac{3n^{\rho + \delta}}{4}, \quad |T_2| \leq (1 + \sigma)\frac{3n^{\rho + \delta}}{4}.$$

The number of edges between T_1 and T_2 is at most $m(|T_1| + |T_2|)$. Therefore, with probability at least $1 - 3n^{-3}$, the set $C_d(v)$ contains about $c_0 n^{2\delta}/4$ vertices, where $1 - \sigma \leq c_0 \leq 1 + \sigma$, and the number of edges $e(C_d(v), V_n \setminus C_d(v))$ between $C_d(v)$ and $V_n \setminus C_d(v)$ 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

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

Now if we set $\delta = \frac{\epsilon}{2}$, $\rho = (\frac{1}{2} - \beta)\epsilon$, then with probability at least $1 - 3n^{-3}$, we have both

$$|C_d(v)| = \Theta(n^{\epsilon}) \quad \text{and} \quad \Phi(C_d(v)) = \frac{1}{|C_d(v)|^{\beta}}.$$

By the union bound, with probability at least $1 - 1/n$, every vertex $v \in V_n$ is contained in a community $C_{n^{(\epsilon-1)/2}}(v)$, 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 phenomenon are satisfied simultaneously [Li and Peng 11].

3.4. The Ravasz–Barabási Hierarchical Model

In [Ravasz and Barabási 03], the authors construct a model that not only has the power law degree distribution, but satisfies the property that the clustering coefficient decays in a characteristic manner. The latter property characterizes the hierarchical feature of networks. The model (we call it the Ravasz–Barabási hierarchical model) is introduced as follows.

Initially, at time $t = 1$, the graph G_1 is a complete graph K_5 in which one of the nodes is marked *center* and the other four nodes are marked *peripheral*. 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}, $1 \le i \le 4$, 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 N_{t-1}.

A stochastic version of the hierarchical model can also be defined (see also [Ravasz and Barabási 03]) 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}, $1 \le i \le 4$, and from each copy we randomly pick a fraction p^{t-1} 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 $p > \frac{1}{5}$ 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 let $V(G)$ and $E(G)$ denote the vertex set and edge set of G, respectively.

Theorem 3.4. *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 for e_i to get

$$e_i = 5^i \left(2 + \frac{16(1 - (\frac{4}{5})^{i-1})}{5}\right) = c_1 \cdot 5^i.$$

Now we have $|V(G_t)| = v_t = 5^t$. 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., $v \in N_{i-1}^j$ for some $j = 1, 2, 3, 4$. We denote this copy by $C(v) = N_{i-1}^j$, 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 numbers of edges inside and outside of $C(v)$ are $c_1 \cdot 5^{i-1}$ and 4^{i-1}, respectively. Thus

$$\Phi(C(v)) = \frac{4^{i-1}}{2 \cdot c_1 \cdot 5^{i-1} + 4^{i-1}} = O\left(\frac{1}{(\frac{5}{4})^{i-1}}\right) = O\left(\frac{1}{|C(v)|^{\log_5 5/4}}\right).$$

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 numbers of vertices and edges in it are $|V(C_1(v))| = 5^{t_0}$ and $e(C_1(v)) = c_1 \cdot 5^{t_0}$, 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 that for G_t. As a consequence, M is the graph obtained from the new process at time t_0. Similarly, we define a vertex 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 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 defined above for G_{t_0}. 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., $v \in N^j_{i-1}$ for some $j = 1, 2, 3, 4$. We denote this copy by $C(v) = N^j_{i-1}$. 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 connects only vertices inside M, then for $v \in V(M)$, we define $C(v) = M = C_1(v)$. Now the number of edges outside of $C(v)$ is at most $(t - t_0)4^{t_0}$. Thus,

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

We can always choose a large constant $c \gg 1$ 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, a fraction $1 - o(1)$ 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 3.5. *Assume that $1/5 < p < 1$. 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 peripheral vertices of the graph G_i, then $v_i = 5^i$, $pv_i = (5p)^i$, $e_i = 5e_{i-1} + 4 \cdot (5p)^{i-1}$, from which we have

$$e_i = 5^i \left(\frac{6}{5} + \frac{4(1 - p^i)}{5(1-p)}\right) = c_1 \cdot 5^i.$$

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 $e(C(v)) = c_1 \cdot 5^{t_0}$ 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 $e(M, V(G_i) \setminus M)$ between the module M and $V(G_i) \setminus M$ may increase as i grows from $t_0 + 1$ to t. We will bound $e(M, V(G_t) \setminus M)$ by showing that at each step, the number of newly formed edges coming out of M is small.

Claim 3.6. *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*

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

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 peripheral modules and one central module. For a peripheral module M and central module M', we have

$$e(M, V(G_{t_0+1}) \setminus M) = (5p)^{t_0} \quad \text{and} \quad e(M', V(G_{t_0+1}) \setminus M') = 4 \cdot (5p)^{t_0},$$

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

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

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 $j + 1$, we see that $e(M, O_j) \approx H(5^j, 5^{t_0}, (5p)^j)$, where $H(A, B, C)$ is the hypergeometric distribution with parameters A, B, and C. Therefore, by the concentration inequality on the hypergeometric distribution (see, e.g., [Dubhashi A. Panconesi 09]), with probability at most $1/2^{c_2 (5p)^{t_0}}$,

$$e(M, O_j) \ge c_2 (5p)^{t_0}. \tag{3.1}$$

If c and c_2 are sufficiently large, then (3.1) holds with probability at most $1/\left(2 \cdot 4 \cdot 5^{j-t_0}(t^c)\right)$.

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

If M is contained in O_j, then

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

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 $j+1$, we have by induction that a selected node in N_j chooses a neighbor in M with probability at most

$$p_j = \frac{2 \cdot c_1 \cdot 5^{t_0} + c_2(j - t_0)(5p)^{t_0}}{2 \cdot c_1 \cdot 5^j} = (1 + o(1))5^{t_0 - j},$$

so $e(M, N_j)$ is dominated by $\mathrm{Bi}(4 \cdot (5p)^j, p_j)$, where $\mathrm{Bi}(n, p)$ denotes the binomial distribution with parameters n and p. Therefore, by the Chernoff bound,

$$e(M, N_j) \geq c_2(5p)^{t_0}, \tag{3.2}$$

with probability at most

$$\frac{1}{2^{c_2(5p)^{t_0}}}.$$

If c and c_2 are sufficiently large, then (3.2) holds with probability at most

$$\frac{1}{2 \cdot 5^{j-t_0}(t^c)}.$$

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

$$1 - \left(\frac{j}{t^c} + \frac{1}{t^c}\right) = 1 - \frac{j+1}{t^c}$$

that

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

□

Using the above claim, we have that with probability at least $1 - 1/t^{c-1}$, all basic modules $C(v)$ in G_t have conductance value

$$\Phi(C(v)) \leq \frac{c_2 t \cdot (5p)^{t_0}}{2c_1 \cdot 5^{t_0} + c_2 t \cdot (5p)^{t_0}} = O\left(\frac{t^{1+c\log_5 5p}}{t^c}\right)$$
$$= O\left(\frac{1}{|C(v)|^{\log_5 \frac{1}{p} - \frac{1}{c}}}\right).$$

Therefore, when c is large enough, with high probability, every vertex v in G_t is contained in a strong (α, β, γ)-community that 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 defined on the vertex set of $\bar{G}$. For example, the small-world model of [Kleinberg et al. 00] is a perturbed graph with $\bar{G}$ and R representing respectively the d-dimensional grid and a random graph on the grid. 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 that arises naturally from our definition of community. Intuitively, a graph exhibiting the 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 of interest to explore the relationship between these two properties. Here we show that for a particular model, with high probability it is an expander under certain conditions, while under some other conditions, it exhibits the small-community phenomenon.

4.1. The d-Dimensional Small-World Model

Edge expansion on several classes of perturbed graphs is studied in [Flaxman 07]. Particularly, it is shown 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 Flaxman's 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 4.1. (Threshold theorem for the small-community phenomenon.) *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 exist weak (α_1, β_1)-communities of size $n/(\ln n)^{c_1}$ for every node, where $\beta_1 < 1$, $c_1 > 0$, and there exist 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 exist 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 $d_{i,j}^{-r}/Z$, where $r > 0$ is the parameter of this model.

We divide the proof into two cases.

Case 1: $r < 1$. In this case, [Flaxman 07] proves that the expansion of G is greater than some small constant δ with high probability. Therefore, for every S satisfying $|S| \leq \frac{n}{2}$, we have $e_G(S, \bar{S}) \geq \delta |S|$. 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.

Case 2: $r \geq 1$. 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 respective events that i has chosen its out-contact in C and $\bar{C}$. For $j \in \bar{C}$, let $X_{j,C}$ and $X_{j,\bar{C}}$ denote respectively the indicator random variables of the events that j has chosen its out-contact in C and $\bar{C}$. 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_{R_1} 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_{R_2} be the number of random arcs from $\bar{C}$ to C. Thus, $e_R(C, \bar{C}) = e_{R_1} + e_{R_2}$. We analyze e_{R_1} and e_{R_2} separately.

For $r = 1$, we have $Z = \sum_{k \neq i} d_{i,k}^{-1} = \Theta(\ln n)$. We can calculate the expectation of e_{R_1} as follows:

$$\begin{aligned} E[e_{R_1}] &= \sum_{i \in C} E[X_{i,\bar{C}}] = \sum_{i \in C} \sum_{j \in \bar{C}} E[X_{ij}] \\ &= \sum_{i \in C} \sum_{j \in \bar{C}} \frac{d^{-1}(i,j)}{Z} \\ &= \Theta\left(\frac{1}{Z} \sum_{i=1}^{k} \left(\sum_{j=i}^{n/2} \frac{1}{j} + \sum_{j=2k+2-i}^{n/2} \frac{1}{j} \right) \right) \\ &= \Theta\left(\frac{1}{\ln n} \left(2k \ln \frac{n}{2} - \sum_{i=1}^{k} \ln i(2k+2-i) \right) \right) \\ &= O\left(\frac{k}{\ln n} \ln \frac{n}{2k} \right). \end{aligned} \tag{4.1}$$

Now since the random variables $\{X_{i,\bar{C}}\}_{i \in C}$ are independent 0, 1 random variables, by the Chernoff bound, we know that e_{R_1} concentrates around its expectation when k is large. Specifically, for any $c_1 > 0$, $0 < c_2 \leq \frac{1}{4}$, and $n/(\ln n)^{c_1} \leq k \leq c_2 n$, with probability at most $o(1/n)$,

$$e_{R_1} > \frac{c'k}{\ln n} \ln \frac{n}{2k},$$

for some constant c'.

Similar results for e_{R_2} can be obtained. Indeed, $E[e_{R_2}]$ is the same as $E[e_{R_1}]$ by the symmetry of $d(\cdot, \cdot)$. Then with probability at least $1 - o(1/n)$,

$$e_R(C, \bar{C}) = e_{R_1} + e_{R_2} \leq \frac{2c'k}{\ln n} \ln \frac{n}{2k}.$$

Now we know that

$$e_G(C) = \Theta(k), \quad e_G(C, \bar{C}) = O\left(\frac{k}{\ln n} \ln \frac{n}{2k} \right),$$

from which we can estimate the conductance of C by definition.

If $k = n/(\ln n)^{c_1}$, then with probability at least $1 - o(1/n)$,

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

where β is an arbitrary constant with $0 \leq \beta < 1$.

If $k = c_2 n$, then with probability at least $1 - o(1/n)$,

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

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

For $r > 1$, the calculations are almost the same. We need only notice that in this case, $Z = \sum_{k \ne i} d_{i,k}^{-r} = \Theta(1)$, and that in equation (4.1), we should replace $1/j$ by $1/j^r$. In so doing, we obtain

$$E[e_{R_1}] = O\Big(\sum_{i=1}^{k}\Big(\sum_{j=i}^{n/2}\frac{1}{j^r} + \sum_{j=2k+2-i}^{n/2}\frac{1}{j^r}\Big)\Big) = O\Big(\sum_{i=1}^{k}(i^{1-r} + (2k+2-i)^{1-r})\Big)$$
$$= \begin{cases} O(k^{2-r}) & \text{if } 1 < r < 2, \\ O(\ln k) & \text{if } r = 2, \\ O(1) & \text{if } r > 2. \end{cases}$$

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

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

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

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

Thus, with probability $1 - o(1)$, for $r > 1$, every vertex v in the graph is contained in a strong (α, β, γ)-community that 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 $d \ge 2$, the proof is almost the same as above: we also need to define $C_k(v) = \{u : d(u, v) \le k\}$ for appropriate k. Noting that in the 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

It is also shown in [Flaxman 07] that a perturbed graph G can be written as $G = \bar{G} + R$, with $\bar{G}$ 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 resultant 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 as q varies from 0 to 1, the structure of the network changes. Intuitively speaking, if q is small, then the conductance of a small subset of $G = \bar{G} + R$ is at least some constant and G does not have a small community; if q is large, then a small community appears.

Theorem 4.2. *If $q < 1/n_1^{\sigma}$ for some constant $\sigma_1 < 1$, then with high probability, every subset S of size $|S| \leq \epsilon n$, where ϵ is an arbitrarily small constant, has conductance larger than some constant; if $q > 1 - 1/\ln n$, then with high probability, every vertex is contained in a strong $(\alpha, 1, 1)$-community.*

Proof. Using the Chernoff bound, it is easy to see that with high probability, the degree of every vertex is bounded above 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 $\delta < \frac{1}{10}$.

It is shown in [Flaxman 07] that the probability that there exists some set S with $|S| = s$ and $e_R(S, \bar{S}) \leq \delta|S|$ is at most

$$P_1 = n \left(\frac{ne}{\delta s}\right)^{2\delta s} \Pr[e_R(S, \bar{S}) \leq \delta s].$$

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

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

If

$$s_0 = \frac{3}{1-3\delta} \le s \le \epsilon n,$$

then for

$$q < \frac{1}{n^{\sigma_1}}, \quad \text{where} \quad \sigma_1 = \frac{5-3\delta}{6(1-\delta)} < 1,$$

we have

$$\begin{aligned} P_1 &\le n \left(\left(\frac{ne}{\delta s}\right)^{2\delta} \left(\frac{e}{\delta}\right)^{\delta} \left(\frac{s-1}{n-1} + q\frac{n-s}{n-1}\right)^{1-\delta} \right)^s \\ &\le n \left(\left(\frac{ne}{\delta s}\right)^{2\delta} \left(\frac{e}{\delta}\right)^{\delta} \left(\frac{2}{n^{\sigma_1}}\right)^{1-\delta} \right)^s \\ &\le n \left(\frac{\left(\frac{e}{\delta}\right)^{3\delta} 2^{1-\delta}}{s^{2\delta} n^{\sigma_1(1-\delta)-2\delta}} \right)^s = n \left(\frac{\left(\frac{e}{\delta}\right)^{3\delta} 2^{1-\delta}}{s^{2\delta} n^{\frac{5-15\delta}{6}}} \right)^s. \end{aligned}$$

Let

$$f(s) = \left(\frac{c}{s^{2\delta} n^{\frac{5-15\delta}{6}}} \right)^s = e^{s\left(\ln c - 2\delta \ln s - \frac{(5-15\delta)\ln n}{6}\right)},$$

where

$$c = c(\delta) = \left(\frac{e}{\delta}\right)^{3\delta} 2^{1-\delta}.$$

For n sufficiently large and δ small enough, the derivative of f is

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

Therefore, we get

$$P_1 \le n f(s_0) = O\left(n \frac{1}{n^{5/2}}\right) = o\left(\frac{1}{n}\right).$$

Combining the fact that

$$e(S, \bar{S}) \ge \delta \frac{3}{1-3\delta}$$

for each S of size

$$|S| = s < \frac{3}{1-3\delta},$$

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 $q < 1/n_1^{\sigma}$, each small set has constant conductance with high probability.

If $q > 1 - 1/\ln n$, 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 in analogy to the proof for the small-world model, we can show that the number of edges $e_R(C(v), V \setminus C(v))$ between $C(v)$ and $V \setminus C(v)$ is concentrated around its expectation

$$E[e_R(C(v), V \setminus C(v))] = \Theta(k(n-k)p).$$

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

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

In particular, if $q = 1 - 1/\ln n$, then

$$\Phi(C(v)) \leq \Theta\left(\frac{1}{|C(v)|}\right).$$

For q larger than this probability, $C(v)$ has conductance even smaller, but the best possible $\Phi(C(v))$ is still of order

$$\Theta\left(\frac{1}{|C(v)|}\right).$$

□

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 based on graph-partitioning are very successful in finding large communities. However, experience in human society tells us that small communities exist almost everywhere, that small communities overlap, and that small communities play important roles in social organization.

Given that networks are natural mathematical models for describing relationships of massive objects in many different subjects of both the 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 society.

In this article, we have proposed a novel approach to defining communities in a network, allowing us to study the small-community phenomenon in some well-defined network models. We show that a number of natural network models

satisfy the small-community phenomenon, which can be regarded as a new feature for a number of networks. Not only do the results we have proved help us to explore and characterize some general properties of real-world networks; they also have potential applications in validation and control of networks.

On the other hand, in the definition of the small-community phenomenon, the requirement that almost every node belong to some community may be too stringent. It would be of interest to study cases in which only a constant fraction of nodes or even fewer belong to some community in both theoretical models and real-world networks.

Acknowledgments. This 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.

References

[Ahn et al. 09] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. "Link Communities Reveal Multi-Scale Complexity in Networks." arxiv:0903.3178, 2009.

[Allen 04] C. Allen. "Life with Alacrity: The Dunbar Number as a Limit to Group Sizes." Available at http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html, 2004.

[Alon and Spencer 08] N. Alon and J. H. Spencer. *The Probabilistic Method*, 3rd edition. New York: John Wiley, 2008.

[Ball et al. 97] F. G. Ball, D. Mollison, and G. Scalia-Tomba. "Epidemics with Two Levels of Mixing." *Ann. Appl. Prob.* 7 (1997), 46–89.

[Barabási and Albert 99] A.-L. Barabási and R. Albert. "Emergence of Scaling in Random Networks." *Science* 286 (1999), 509–512.

[Blandford et al. 03] 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*, pp. 679–688. Philadelphia: Society for Industrial and Applied Mathematics, 2003.

[Chierichetti et al. 10a] F. Chierichetti, S. Lattanzi, and A. Panconesi. "Rumour Spreading and Graph Conductance." In *SODA '10:ACM–SIAM Symposium on Discrete Algorithms*, 2010.

[Chierichetti et al. 10b] F. Chierichetti, S. Lattanzi, and A. Panconesi. "Almost Tight Bounds for Rumour Spreading with Conductance." *STOC '10:ACM Symposium on Theory of Computing*, 2010.

[Chung 97] F. R. K. Chung. "Spectral Graph Theory." *Regional Conference Series in Mathematics, American Mathematical Society* 92 (1997), 1–212.

[Danon et al. 05] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. "Comparing Community Structure Identification." *Journal of Statistical Mechanics: Theory and Experiment* 9 (2005), P09008.

[Dubhashi A. Panconesi 09] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. New York: Cambridge University Press, 2009.

[Dunbar 96] R. Dunbar. *Grooming, Gossip and the Evolution of Language*. Cambridge, MA: Harvard Univ. Press, 1996.

[Durrent 07] R. Durrent. *Random Graph Dynamics*. Cambridge, U.K.: Cambridge University Press, 2007.

[Erdős and Rényi 60] P. Erdős and A. Rényi. "On the Evolution of Random Graphs." *Publ. Math. Inst. Hungary. Acad. Sci.*, 5 (1960), 17–61.

[Estrada 06] E. Estrada. "Spectral Scaling and Good Expansion Properties in Complex Networks." *Europhysics Letters* 73 (2006), 649–655.

[Flaxman 07] A. D. Flaxman. "Expansion and Lack Thereof in Randomly Perturbed Graphs." *Internet Mathematics* 4:2 (2007), 131–147.

[Flaxman et al. 07a] A. D. Flaxman, A. Frieze, and J. Vera. "A Geometric Preferential Attachment Model of Networks." *Internet Mathematics* 3:2 (2007), 87–111.

[Flaxman et al. 07b] A. D. Flaxman, A. M. Frieze, and J. Vera. "A Geometric Preferential Attachment Model of Networks II." *Internet Mathematics* 4:1 (2007), 87–111.

[Fortunato 10] S. Fortunato. "Community Detection in Graphs." *Physics Reports* 486 (2010), 75–174.

[Girvan and Newman 02] M. Girvan and M. E. J. Newman. "Community Structure in Social and Biological Networks." *PNAS* 99:12 (2002), 7821–7826.

[Granovetter 73] M. Granovetter. "The Strength of Weak Ties." *The American Journal of Sociology* 78:6 (1973), 1360–1380.

[Hoory et al. 06] S. Hoory, N. Linial, and A. Wigderson. "Expander Graphs and Their Applications." *Bull. Amer. Math. Soc. (N.S* 43 (2006), 439–561.

[Hopcroft et al. 03] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. "Natural Communities in Large Linked Networks." In *Proc. of KDD'03*, 2003.

[Jonsson et al. 06] 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* 2006, 7:2.

[Kannan et al. 04] R. Kannan, S. Vempala, and A. Vetta. "On Clusterings: Good, Bad and Spectral." *J. ACM* 51:3 (2004), 497–515.

[Kleinberg et al. 00] J. Kleinberg. "The Small-World Phenomenon: An Algorithmic Perspective." In *Proceedings of the 32nd ACM Symposium on the Theory of Computing*, 2000.

[Kumar et al. 00] 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*, p. 57. Washington, DC: IEEE Computer Society, 2000.

[Lancichinetti and Fortunato 09] A. Lancichinetti and S. Fortunato. "Benchmarks for Testing Community Detection Algorithms on Directed and Weighted Graphs with Overlapping Communities." *Phys. Rev. E* 80:1 (2009), 016118.

[Lancichinetti et al. 08] A. Lancichinetti, S. Fortunato, and F. Radicchi. "Benchmark Graphs for Testing Community Detection Algorithms." *Physical Review E* 78:4 (2008), 46110.

[Leskovec et al. 07] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graph Evolution: Densification and Shrinking Diameters." *ACM Trans. Knowl. Discov. Data*, Vol. 1, No. 1 (2007), 2.

[Leskovec et al. 2008] 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.

[Li and Peng 11] Angsheng Li and Pan Peng. "The Small-Community Phenomenon in Networks." To appear, 2011.

[Mihail et al. 06] M. Mihail, C. Papadimitriou, and A. Saberi. "On Certain Connectivity Properties of the Internet Topology." *J. Comput. Syst. Sci.* 72:2 (2006), 239–251.

[Mishra et al. 08] N. Mishra, R. Schreiber, I. Stanton, and R. E. Tarjan. "Finding Strongly Knit Clusters in Social Networks." *Internet Mathematics* 5:1 (2008), 155–174.

[Mitzenmacher 04] M. Mitzenmacher. "A Brief History of Generative Models for Power Law and Lognormal Distributions." *Internet Mathematics* 1:2 (2004), 226–251.

[Mitzenmacher and Upfal 05] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. New York: Cambridge University Press, 2005.

[Newman and Girvan 04] M. E. J. Newman and M. Girvan. "Finding and Evaluating Community Structure in Networks." *Phys. Rev. E*, 69:2 (2004), 026113.

[Palla et al. 05] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society." *Nature* 435:7043 (2005), 814–818.

[Rapoport 53] A. Rapoport. "Spread of Information through a Population with Sociostructural Basis." *Bulletin of Mathematical Biophysics* 15 (1953), 523–543.

[Ravasz and Barabási 03] E. Ravasz and A.-L. Barabási. "Hierarchical Organization in Complex Networks." *Physical Review E* 67 (2003), 026112.

[von Luxburg 07] U. von Luxburg. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17:4 (2007), 395–416.

[Watts and Strogatz 98] D. J. Watts and S. H. Strogatz. "Collective Dynamics of "Small-World" Networks." *Nature* 393 (1998), 440–442.

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)

Received September 26, 2010; accepted December 20, 2010.