Community structure and scale-free collections of Erdős–Rényi graphs

C. Seshadhri,* Tamara G. Kolda,† and Ali Pinar‡
Sandia National Laboratories, Livermore, CA, USA
(Dated: March 27, 2012)

Community structure plays a significant role in the analysis of social networks and similar graphs, yet this structure is little understood and not well captured by most models. We formally define a community to be a subgraph that is internally highly connected and has no deeper substructure. We use tools of combinatorics to show that any such community must contain a dense Erdős–Rényi (ER) subgraph. Based on mathematical arguments, we hypothesize that any graph with a heavy-tailed degree distribution and community structure must contain a scale free collection of dense ER subgraphs. These theoretical observations corroborate well with empirical evidence. From this, we propose the Block Two-Level Erdős–Rényi (BTER) model, and demonstrate that it accurately captures the observable properties of many real-world social networks.

I. INTRODUCTION

Graph analysis is becoming increasingly prevalent in the quest to understand diverse phenomena like social relationships, scientific collaboration, purchasing behavior, computer network traffic, and more. We refer to graphs coming from such scenarios collectively as *interaction networks*. A significant amount of investigation has been done to understand the graph-theoretic properties common to interaction networks. Of particular importance is the notion of community structure. Interaction networks typically decompose into internally well-connected sets referred to as low conductance or high modularity cuts [1, 2]. Moreover, many graphs have high clustering coefficients [3], which is indicative of underlying community structure. Communities occur in a variety of sizes, though the largest community is often much smaller than the graph itself [4, 5]. Community analysis can reveal important patterns, decomposing large collections of interactions into more meaningful components.

A. A Theory of Communities

One metric of the quality of a community is the modularity metric [2]. There are other measures such as conductance [6], but they are equivalent to modularity in terms of our intentions. Consider a graph G (undirected) with n vertices and degrees $d_1, d_2, \ldots, d_n$. Let $m = \frac{1}{2}\sum_{i=1}^{n} d_i$ denote the number of edges. We say a subgraph S has high modularity if S contains many more internal edges than predicted by a *null* model, which says vertices i and j are connected with probability $d_i d_j/2m$. (Technically, this is only true if we assume $d_i^2 \leq 2m$ for all i. We keep the notation simple for now and handle the case of $d_i^2 > 2m$ explicitly in our discussion of the theoretical details.) We refer to the null model as the CL model, based on its formalization by Chung and Lu [7, 8]; see also Aiello et al. [9]. It is very similar to the edge-configuration model of Newman et al.[10].

Given a high modularity subgraph S, we say it is a *module* if it does not contain any further substructures of interest; in other words, it is internally well-modeled by CL. Formally, assume S has r nodes with *internal* degrees $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_r$ and let the number of edges in S be denoted by $s = \frac{1}{2}\sum_{i=1}^{r} \hat{d}_i$. Consider the CL model on S, where edge (i, j) occurs with probability $\hat{d}_i \hat{d}_j/2s$. We call S a *module* if the induced subgraph on S (the subgraph internal to S) is modeled well by this CL model. Looking at the contrapositive, if S is not a module, then S itself contains a subset of vertices that should be separated out. A module can be thought of as an "atomic" substructure within a graph. In this language, we can think of community detection algorithms as breaking a graph into modules. This discussion is not complete, however, since communities are not just modules, but also internally well-connected.

Interaction networks have an abundance of triangles, a fact that Watts and Strogatz [3] succinctly express through *clustering coefficients*. Barrat and Weigt [11] defined this as

$$C = \frac{3 \times \text{total number of triangles}}{\text{total number of wedges}}, \tag{1}$$

where a *wedge* is a path of length 2 [1, 3]. It has been observed that C "has typical values in the range of 0.1 to 0.5 in many real-world networks" [1]. Moreover, our own studies have revealed that the node-level clustering coefficient (first used in [3]), C_i, defined by

$$C_i = \frac{\text{number of triangles incident to node } i}{\text{number of wedges centered at node } i}, \tag{2}$$

is typically highest for small degree nodes. Large clustering coefficients are considered a manifestation of the community structures. Naturally, we expect the triangles to be largely contained within the communities due to their high internal connectivity.

We now formally define a *community* to be a module with a large internal clustering coefficient. Note that

* scomand@sandia.gov
† tgkolda@sandia.gov
‡ apinar@sandia.gov

this is different from many other definitions of community which generally define a community as being more internally than externally connected. More formally, we say a module is a community if the expected number of triangles is more than $(\kappa/3)$ times the total number of wedges, for some constant κ. This is consistent with the notion of *triangle modularity* introduced by Arenas et al. [12]. By our definition, a community is tightly connected internally and therefore contains many triangles. A graph has *community structure* if it (or at least a constant fraction of it) can be broken up into communities. The benefit of this formalism is that we can now try to understand what graphs with community structure look like.

Let us first begin by just focusing on a single community. It seems fairly intuitive that a community in a social network cannot be large while comprising only low degree vertices nor that it consists of a single high-degree node connected to degree-one vertices (a star). We can actually prove a structural theorem about a community, given our formalization. Recall that an Erdős–Rényi (ER) graph [13, 14] on n vertices with connection probability p is a graph such that each pair of vertices is independently connected with probability p. If p is a constant, we call this a dense ER graph; if $p = O(1/n)$, then we call this a sparse ER graph. Using triangle bounds from extremal combinatorics and some probabilistic arguments, we can prove the following theorem.

Theorem 1. *A constant fraction of the edges in a community are contained in a dense Erdős–Rényi graph. More formally, if the community has s edges, then there must be $\Omega(\sqrt{s})$ vertices with degree $\Omega(\sqrt{s})$.*

This theorem is interesting because even though it is well known that ER graphs are not good models for interaction networks, they nonetheless form an important building block for the communities. We interpret this theorem as saying that the simplest possible community is just a dense ER graph. Building on this simple intuition, we think of an interaction network as consisting of a large collection of dense ER graphs.

This leads naturally to a question about the distribution of sizes of these ER components. A consequence of our theory is that an ER community with $d+1$ vertices would have $\rho \cdot d^2$ edges for some constant $\rho \leq 1$. For the hypothesis on the expected number of triangles to hold, ρ is assumed to be close to one. For simplicity, we assume that the communities are homogeneous in degree. Now, consider the power law degree distribution observed by Barabási and Albert [15] and others. They show that interaction graphs exhibit heavy-tailed degree distributions such as

$$X_d \propto d^{-\gamma} \tag{3}$$

where X_d is the number of nodes of degree d and γ is the power law exponent. If we assume that all nodes in a community have the same degree, then nodes of degree d yield $X_d/(d+1)$ communities. Thus, if we let Y_d be the number of communities of size $(d+1)$ (with vertices of degree d), then

$$Y_d \propto X_d/(d+1) \propto d^{-(\gamma+1)}.$$

This forms a scale-free distribution of communities, exactly as observed by many studies on community structure [4, 5]. *Hence, we hypothesize that real-world interaction networks consist of a scale-free collection of dense Erdős–Rényi graphs.* This is consistent with most of the important observed properties of these networks.

Empirically, a variety of studies [4, 5, 16, 17] show the existence of a few "reasonably large" communities, a large number of small communities, and all scales in between. It has been observed for a large number of diverse graphs that the largest community seen is of the order of 100 vertices [4]. This is quite consistent with a calculation based on our hypothesis. If we assume $X_d = n/d^\gamma$ and that there exists a community of size d, then we require $n/d^{\gamma+1} \geq 1$. Hence, the maximum community size, $\bar{d}$, is $\bar{d} \approx n^{1/(\gamma+1)}$. For n being a million and $\gamma = 2$, we get an estimate at the order of a 100 nodes.

As an aside, Thm. 1 also proves that CL by itself is not a good model for interaction networks. Suppose the entire graph G (with m edges) can be modeled as a CL graph. Since G has a high clustering coefficient, then G itself is a module. Hence, G must have $\Omega(\sqrt{m})$ vertices with degree $\Omega(\sqrt{m})$, but this violates the tail behavior of the degree distribution.

B. The BTER model

Based on the idea of a graph comprising ER communities, we propose the Block Two-Level Erdős-Rényi model (BTER). The advantages of the BTER model are that it has community structure in the form of dense ER subgraphs and that it matches well with real-world graphs. We briefly describe the model here and provide a more detailed explanation and comparisons to real-world graphs in subsequent sections.

The first phase (or level) of BTER builds a collection of ER blocks in such a way that the specified degree distribution is respected. The BTER model allows one to construct a graph with *any* degree distribution. Real-world degree distributions might be idealized as power laws, but it is by no means a completely accurate description [18, 19]. When the degree distribution is heavy tailed, then the BTER graph naturally has scale-free ER subgraphs. The internal connectivity of the ER graphs is specified by the user and can be tuned to match observed data.

The second phase of BTER interconnects the blocks. We assume that each node has some *excess degree* after the first phase. For example, if vertex i should have d_i incident edges (according to the input degree distribution), and it has d'_i edges from its ER block, then the excess degree is $d_i - d'_i$. We use a CL model (which can

be considered as a weighted form of ER) over the excess degrees to form the edges that connect communities.

C. Previous models

There are many existing models for social networks and other real-world graphs. We give a short description of some important models; for more details, we recommend the survey of Chakrabarti and Faloutsos [20]. Classic examples include preferential attachment [15], small-world models [3], copying models [21], and forest fire [22]. Although these models may produce heavy-tailed degree distributions, their clustering coefficients of the former three models are often low [23]. Even for models that give high clustering coefficients, it is difficult to predict their community structure in advance. Because of their unpredictable behavior, it is not possible to match real data with these graphs. This makes it difficult to validate against real-world interaction networks. Moreover, none of these models explain community structure, one of the most striking features of interaction graphs.

A widely used model for generating large graphs is the Stochastic Kronecker Graph model (known as R-MAT in an early incarnation) [24, 25]. Notably, it has been selected as the generator for the Graph 500 Supercomputer Benchmark [26]. Though it has some desirable properties [25], it can only generate lognormal tails (after suitable addition of random noise [27]) and does not produce high clustering coefficients [23, 28]. Multifractal networks are closely related to the SKG model [29]. The random dot product model [30, 31] can be made scalable but has never been compared to real social networks. There have been successful dendogram based structures that perform community detection and link prediction in real graphs [32, 33]. The recent hyperbolic graph model [34, 35] is based on hyperbolic geometry and has been used to perform Internet routing.

The stochastic block model [36] has been used to generate better algorithms for community detection. A degree corrected version [37] has been defined to deal with imprecisions in this model. A key feature of these models is that they break the graph into a *constant* number of relatively large blocks, and our theory shows that this model does not give a satisfactory explanation of the clustering coefficients of low degrees (which constitute a majority of the graph). The LFR community detection benchmark [38] is also somewhat connected to this model, since it defines a set of communities and has probabilities of edges within and between these communities. We stress that these models do not attempt to match real graphs, nor do they explain the scale-free nature of communities [4, 5]. Our hypothesis and model are very different from these results, because we use a mathematical formalization to prove the existence of a scale-free dense ER collection, and the BTER model follows this theory. Nonetheless, our model can be seen as an extension of these block models, where the number and sizes of blocks form a scale-free behavior. Implicitly, our model can be seen to use a labeling scheme for vertices that depends on the degrees, and connecting vertices with probabilities depending on the labels (thereby related to the degree corrected framework of [37]).

II. MATHEMATICAL OVERVIEW

We provide a sketch of the proof for Thm. 1 to give an intuitive explanation; a complete proof is provided in the next section. Our analysis is fundamentally asymptotic, so for ease of notation we use the $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ to suppress constant factors. The notation $A \ll B$ indicates that there exists some absolute constant c such that $A \leq cB$. We let S denote the community of interest and assume that the internal degree distribution of the community S is $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_r$ and ordered so that $\hat{d}_i \leq \hat{d}_j$ for all $i < j$. We denote the number of edges in S by $s = \frac{1}{2}\sum_{i=1}^{r} \hat{d}_i$.

Based on the given distribution, let T denote the expected number of triangles in S. Since this is a community, we demand that T be at least $\kappa/3$ times the expected number of wedges, for some constant κ. This means that

$$T \geq (\kappa/3) \sum_i \binom{\hat{d}_i}{2}. \tag{4}$$

(For convenience, we will assume that $\forall i, \hat{d}_i > 1$, since degree-1 vertices do not participate in triangles.) A key fact we use is the *Kruskal-Katona theorem* [39–41] which states that if a graph has T triangles and s edges, then $T \leq s^{3/2}$. Combining, we have

$$\sum_i \hat{d}_i^2 \ll s^{3/2}. \tag{5}$$

Now, let us count the expected number of triangles based on the CL distribution. For any triple (i, j, k), let X_{ijk} be the indicator random variable for (i, j, k) being a triangle. This occurs when all the edges (i, j), (j, k), and (k, i) are present, and by independence, this probability is $p_{ij}p_{jk}p_{ki}$, where $p_{ij} = \hat{d}_i\hat{d}_j/2m$. The expected number of triangles T can be expressed as $\mathbf{E}[\sum_{i<j<k} X_{ijk}]$, which (by linearity of expectation) is $\sum_{i<j<k} \mathbf{E}[X_{ijk}]$. Therefore,

$$T = \sum_{i<j<k} \frac{\hat{d}_i\hat{d}_j}{2s} \cdot \frac{\hat{d}_j\hat{d}_k}{2s} \cdot \frac{\hat{d}_i\hat{d}_k}{2s} \leq \frac{\left(\sum_i \hat{d}_i^2\right)^3}{8s^3}. \tag{6}$$

We argued earlier that $T = \Omega(\sum_i \hat{d}_i^2)$. We can put this bound in (6) and rearrange to get

$$s^{3/2} \ll \sum_i \hat{d}_i^2. \tag{7}$$

This is the exact reverse of (5)! This means that these quantities are the *same* up to constant factors. When

can this be satisfied? If the community consists of $\sqrt{s}$ vertices all with degree $\sqrt{s}$, then $\sum_i \hat{d}_i^2 = \sum_i s = s^{3/2}$, and the conditions are exactly satisfied. Intuitively, to satisfy both (5) and (7), there have to be $\Theta(\sqrt{s})$ vertices of degree $\Theta(\sqrt{s})$. These vertices form a dense ER graph within the community proving that each community involves a constant fraction of the edges in an ER graph.

III. THEORETICAL DETAILS

A reader interested in only a general overview of the results can skip this section. The aim of this section is to prove Thm. 1, which we restate in slightly different wording for convenience. The proof is fairly involved mathematically, and formalizes the argument discussed in the previous section.

We use small Greek letters for constants less than 1, and small Roman letters for constants whose values may exceed 1. All constants are positive. We make no attempt to optimize various constant factors in the proof. The proof is asymptotic in s, the number of edges of our community. That means that the proof holds for any sufficiently large s.

We describe the specifics of the CL graph that is generated. For every ordered pair (i,j), let $p_{ij} = \hat{d}_i\hat{d}_j/(2s)^2$. Note that this creates a distribution over all pairs, since $\sum_{i,j} \hat{d}_i\hat{d}_j/(2s)^2 = (\sum_i \hat{d}_i)^2/(2s)^2 = 1$. We generate s independent samples from this distribution to get our graph. The final graph is made undirected and simple (so parallel edges are removed). This is one of the standard methods for the edge-configuration model.

Theorem 2 (Restatement of Thm. 1)**.** *Consider a CL graph with degree sequence $1 < \hat{d}_1 \le \hat{d}_2 \le \cdots \le \hat{d}_r$ and set $s = \sum_i \hat{d}_i/2$. The quantities $c > 0$ and $\kappa \in (0,1)$ are constants (independent of s).*

Assume the expected number of triangles in a CL graph generated with this degree sequence is at least $(\kappa/3)\sum_i \binom{\hat{d}_i}{2}$. Then for sufficiently large s, there exists a set of indices $U \subseteq \{1,\ldots,r\}$, such that $|U| = \Omega(\sqrt{s})$ and $\forall k \in U, \hat{d}_k = \Omega(\sqrt{s})$.

(The constants hidden in the $\Omega(\cdot)$ notation only hide a dependence on c and κ.)

The proof of this theorem requires some extremal combinatorics and probability theory. We first state some of these building blocks before describing the main proof. Henceforth, the assumptions stated in the theorem hold. An important tool is the Kruskal-Katona theorem that gives an upper bound on the number of triangles in a graph with a fixed number of edges.

Theorem 3 (Kruskal-Katona [39–41])**.** *If a graph has t triangles and m edges, then $t \le m^{3/2}$.*

The probability that i and j are connected is well approximated by $\hat{d}_i\hat{d}_j/2s$ when this is much smaller than 1. But we can always use it as an *upper bound*, as the following claim shows.

Claim 4. *Consider three vertices $i \ne j \ne k$. The probability that the triangle (i,j,k) forms in the CL graph is*

$$O\left(\min\left(\frac{\hat{d}_i\hat{d}_j}{2s},1\right)\cdot\min\left(\frac{\hat{d}_i\hat{d}_k}{2s},1\right)\cdot\min\left(\frac{\hat{d}_j\hat{d}_k}{2s},1\right)\right).$$

Proof. Consider the pair (i,j). For a single edge insertion, the probability that the edge (i,j) is inserted is $2\hat{d}_i\hat{d}_j/(2s)^2$. Define this to be q_{ij}. Note that $q_{ij} < 0.5$. The probability that the edge is never inserted over s edge insertions is $(1-q_{ij})^s \ge \exp(-q_{ij}s/(1-q_{ij}))$. Suppose $q_{ij}s \le 0.5$, so the term in the exponent is (strictly) at most 1. Then, this can be approximated as $(1-q_{ij})^s \ge 1 - q_{ij}s/(1-q_{ij}) \ge 1 - 2q_{ij}s$. Hence, the probability that this edge is inserted is at most $2q_{ij}s = O(\hat{d}_i\hat{d}_j/2s)$.

When $q_{ij}s \ge 0.5$, then $\hat{d}_i\hat{d}_j/2s = \Omega(1)$. The probability that an edge is inserted is trivially at most 1, so in this case as well, this probability is $O(\hat{d}_i\hat{d}_j/2s)$. Combining both case, the probability of this edge insertion is at most $O(\min(\hat{d}_i\hat{d}_j/2s, 1))$.

We note that the events corresponding to the appearance of edges (i,j), (j,k), and (k,i) are independent in the limit, as $s \to \infty$. □

We now prove some claims about the expected number of triangles and the degree sequence.

Claim 5. *Let T denote the expected number of triangles. There exist constants β and c', depending only on c and κ, such that*

1. *$T \ge \beta \sum_i \hat{d}_i^2$, and*
2. *$\sum_i \hat{d}_i^2 \le c' s^{3/2}$.*

Proof. By assumptions in Thm. 2, $T \ge (\kappa/3)\sum_i \binom{\hat{d}_i}{2}$. For $\hat{d}_i > 1$, $\binom{\hat{d}_i}{2} \ge \hat{d}_i^2/4$ (for large $\hat{d}_i$, it is actually much closer to $\hat{d}_i^2/2$). Hence, $T \ge (\kappa/12)\sum_i \hat{d}_i^2$, and setting $\beta = \kappa/12$ completes the proof of the first part.

Suppose we generate a random CL graph. Let t be the number of triangles and E be the number of edges (both random variables). By Thm. 3, $t \le E^{3/2}$. Taking expectations and noting that $E \le s$, $T \le \mathbf{E}[E^{3/2}] \le s^{3/2}$. Combining with the first part of this claim, $\sum_i \hat{d}_i^2 \le (1/\beta)s^{3/2}$. We set constant $c' = 1/\beta$. □

We come to the proof of the main theorem.

Proof. (of Thm. 2) We choose b to be a sufficiently large constant, and γ to be sufficiently small. Let ℓ be the smallest index such that $\hat{d}_\ell > b\sqrt{s}$. For a triple of vertices (i,j,k), let X_{ijk} be the indicator random variable for (i,j,k) forming a triangle. Note that $T = \mathbf{E}[\sum_{i<j<k} X_{ijk}]$. Then we have the following. We bound

$\mathbf{E}[X_{ijk}]$ using Claim 4 as follows. As mentioned earlier, we use $\ll$ as shorthand for the big-Oh notation:

$$\begin{aligned}
&\mathbf{E}[\sum_{i<j<k} X_{ijk}] = \sum_{i<j<k} \mathbf{E}[X_{ijk}] \\
&\ll \sum_{i<j<k} \min\left(\frac{\hat{d}_i\hat{d}_j}{2s}, 1\right) \min\left(\frac{\hat{d}_i\hat{d}_k}{2s}, 1\right) \min\left(\frac{\hat{d}_j\hat{d}_k}{2s}, 1\right) \\
&\leq \sum_{i<j<k} \frac{\hat{d}_i\hat{d}_j}{2s} \cdot \frac{\hat{d}_i\hat{d}_k}{2s} \cdot \min\left(\frac{\hat{d}_j\hat{d}_k}{2s}, 1\right) \\
&\leq \sum_i \hat{d}_i^2 \sum_{j<k} \frac{\hat{d}_j\hat{d}_k}{4s^2} \cdot \min\left(\frac{\hat{d}_j\hat{d}_k}{2s}, 1\right)
\end{aligned}$$

We now split the second sum based in the case $j \leq \ell$ and $j > \ell$. In the former case, we bound the min term by $\hat{d}_j\hat{d}_k/2s$ and in the latter case, by 1. Note that in the second sum (below), $k \geq \ell$, since $k > j$.

$$\begin{aligned}
\mathbf{E}[\sum_{i<j<k} X_{ijk}] &\ll \sum_i \hat{d}_i^2 \left[\sum_{\substack{j,k:\\ j\leq\ell}} \frac{\hat{d}_j^2\hat{d}_k^2}{8s^3} + \sum_{\substack{j<k:\\ j>\ell}} \frac{\hat{d}_j\hat{d}_k}{4s^2}\right] \\
&\leq (\sum_i \hat{d}_i^2)\left[\frac{(\sum_{j\leq\ell}\hat{d}_j^2)(\sum_k \hat{d}_k^2)}{8s^3} + \frac{(\sum_{j\geq\ell}\hat{d}_j)^2}{4s^2}\right].
\end{aligned}$$

By the first part of Claim 5, $T \geq \beta\sum_i \hat{d}_i^2$. For convenience, we replace all the independent indices above by i. Then

$$\begin{aligned}
\beta &\ll \frac{(\sum_i \hat{d}_i^2)(\sum_{i\leq\ell}\hat{d}_i^2)}{8s^3} + \frac{(\sum_{i\geq\ell}\hat{d}_i)^2}{4s^2} \\
\implies \beta' &\leq \frac{(\sum_i \hat{d}_i^2)(\sum_{i\leq\ell}\hat{d}_i^2)}{8s^3} + \frac{(\sum_{i\geq\ell}\hat{d}_i)^2}{4s^2} \qquad (8)
\end{aligned}$$

We use β' to denote some constant (that comes from $\ll$ in the previous inequality). By Claim 5, $\sum_i \hat{d}_i^2 \leq c's^{3/2}$. Furthermore, $\sum_i \hat{d}_i^2 \geq b\sqrt{s}\sum_{i\geq\ell}\hat{d}_i$ (since for $i \geq \ell$, $\hat{d}_i \geq b\sqrt{s}$). Combining the two bounds, $\sum_{i\geq\ell}\hat{d}_i \leq (c'/b)s$. Applying these bounds in (8) and setting constant τ appropriately,

$$\begin{aligned}
&\beta' \leq \frac{c'\sum_{i\leq\ell}\hat{d}_i^2}{8s^{3/2}} + (c'/2b)^2, \\
&\implies \sum_{i\leq\ell}\hat{d}_i^2 \geq (8/c')(\beta' - (c'/2b)^2)s^{3/2} = \tau s^{3/2}.
\end{aligned}$$

(By setting b to be sufficiently large, we can ensure that τ is a positive constant.) Let ℓ' be the smallest index such that $\hat{d}_{\ell'} \geq \gamma\sqrt{m}$ and set $s' = \sum_{\ell'\leq i\leq\ell}\hat{d}_i$. Then,

$$\begin{aligned}
\tau s^{3/2} &\leq \sum_{i\leq\ell} d_i^2 \leq \sum_{i<\ell'} \hat{d}_i^2 + \sum_{\ell'\leq i\leq\ell} \hat{d}_i^2 \\
&\leq \gamma\sqrt{s}\sum_{i<\ell'}\hat{d}_i + b\sqrt{s}\sum_{\ell'\leq i\leq\ell}\hat{d}_i \\
&\leq \gamma(s-s')\sqrt{s} + bs'\sqrt{s}, \\
&\implies \tau s \leq s'(b-\gamma) + \gamma s, \\
&\implies s' \geq s(\tau-\gamma)/(b-\gamma) = \Omega(s).
\end{aligned}$$

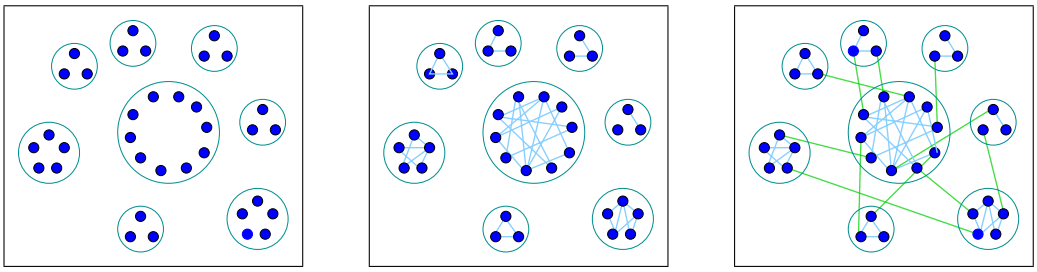
(Again, a sufficiently small γ ensures positivity.) This means that the vertices with indices in $[\ell', \ell]$ are incident to at least $\Omega(s)$ edges. All these vertices have degrees that are $\Theta(\sqrt{s})$, and hence there are $\Theta(\sqrt{s})$ such vertices. □

IV. THE BTER MODEL IN DETAIL

The BTER model comprises an interconnected scale-free collection of communities. Intuitively, short-range connections (Phase 1) tend to be dense and lead to large clustering coefficients. Long-range connections (Phase 2) are sparse and lead to heavy-tailed degree distributions. We describe the steps in detail below.

a. Preprocessing In the preprocessing step, each node of degree 2 or higher is assigned to a community. We assume the desired degree distribution $\{d_i\}$ is given where d_i denotes the desired degree of node i. Roughly speaking, $d+1$ vertices of degree d are assigned to a community, so that the induced subgraph on these edges is dense. We sort the vertices of degree at least 2 in increasing order of degree, so we have $d_1 \leq d_2 \leq d_3 \ldots$. Think of all vertices as being placed in a single stack (in this order), with vertex 1 at the head. We read the degree d of the head, and pop $d+1$ vertices from the stack. These vertices form a community. We repeatedly perform this popping operation, until the stack is empty. The vertices are all partitioned into these communities. Note that this process groups together vertices of the same degree except for the few instances where there is some cross-over (i.e., the last degree-2 vertex is groups with two degree-3 vertices) or for high-degree vertices where there are very few of each degree. If the degree distribution has a scale-free behavior, the number of communities of a given size is also scale-free. Since the degree distribution is an input to the model, this step is relatively straightforward and results in a structure as shown in Fig. 1a. We let $\mathcal{G}_r$ denote the rth community and u_i denote the community assignment for node i.

b. Phase 1 The local community structure is modeled as an ER graph on each community. This is illustrated in Fig. 1b. The connectivity of each community is a parameter of the model. By observing the clustering coefficient plots for real graphs, we can see that low degree vertices have a much higher clustering coefficient than higher degree ones. This suggests that small communities are much more tightly connected than larger



(a) Preprocessing: Distribution of nodes into communities

(b) Phase 1: Local links within each community

(c) Phase 2: Global links across communities

FIG. 1: (Color online) BTER Model Construction. In the preprocessing phase, the nodes are divided into communities. In Phase 1, within-community links are generated using the ER model. In Phase 2, across-community links are generated using the CL model on the *excess* degrees.

ones, and so we adjust the connectivity accordingly. Any formula may be used; we have found empirically that the following works well in practice. We let the edge probability for community r be defined as

$$\rho_r = \rho \left[1 - \eta \left(\frac{\log(\bar{d}_r + 1)}{\log(d_{\max} + 1)} \right)^2 \right], \tag{9}$$

where $\bar{d}_r = \min \{ d_i \mid i \in \mathcal{G}_r \}$, $d_{\max}$ is the maximum degree in the graph, and ρ and η are parameters that can be selected for the best fit to a particular graph. (These were selected by manual experimentation for our results, but more elaborate procedures could certainly be developed.) Why do we choose such a formula? We observe that in most real networks, the clustering coefficients of low degree vertices are quite large. As the degree increases, this decays and finally reaches very low values for large degree. This decay appears to happen in log-scale, and hence we use the above formula. In the next section, we show how well BTER matches the plots for clustering coefficients of real graphs.

c. Phase 2 The global structure is determined by interconnecting the communities. We apply a CL model to the *excess* degree, e_i, of each node, which is computed as follows:

$$e_i = \begin{cases} 1, & \text{if } d_i = 1, \\ d_i - \rho_{u_i}(|\mathcal{G}_{u_i}| - 1), & \text{otherwise,} \end{cases} \tag{10}$$

where $|\mathcal{G}_r|$ is the size of community r. Given the e_i's for all nodes, edges are generated by choosing two endpoints at random. Specifically, the probability of selecting node i is $e_i / \sum_j e_j$. It is possible to produce duplicate links or self-links, but these are discarded. Phase 2 is illustrated in Fig. 1c.

d. Reference implementation A MATLAB reference implementation of BTER is available at `http://www.sandia.gov/~tgkolda/bter_supplement/`. Scripts are also provided to reproduce the findings in this paper. In this implementation, we have taken some care to reduce the variance in the CL model with respect to degree-one nodes. We also generate extra edges in Phase 2 to account for expected repeats and self-loops that are removed. These details are described in the next section.

e. Assortativity Phase 1 of BTER is highly assortative, and Phase 2 is unassortative (being a CL graph). The ρ_r parameter controls the proportion of edges in Phase 1 versus Phase 2. In general, BTER is an appropriate model when the clustering coefficient of the real data is large, since that conforms to our theoretical framework which assumes many triangles. When the underlying clustering coefficient is extremely low (as in the case of internet backbone or peer-to-peer graphs [5]), then the communities are not expected to have many triangles. As a consequence, BTER is not a good model for graphs that are highly disassortative.

V. IMPLEMENTATION DETAILS

We give some specifics of our BTER implementation, but a reader interested in just a general overview may skip this section. In Phase 1, we have found it convenient to set $\rho_r = 0$ for the last community since it comprises just a few "leftover" nodes.

We split the calculation of the Phase 2 edges into three subphases so that we can specially handle the degree-1 edges. The variance for degree-1 vertices in the CL model is high, so we set aside a proportion of these vertices to be handled "manually." Let w denote the number of degree-1 vertices, and assume the vertices are indexed from least degree to greatest. By default, 75% of the degree-1 vertices are handled "manually" (the exact proportion is user-definable); let $p = \lfloor 0.75w \rceil$ denote this quantity where $\lfloor \cdot \rceil$ denotes nearest integer. We update e_i as follows:

$$e_i \leftarrow \begin{cases} 0, & \text{for } 1 \le i \le p, \\ 1.10, & \text{for } p+1 \le i \le w, \\ e_i, & \text{otherwise.} \end{cases}$$

This update removes the first p nodes from the CL part and also slightly raises the probability of an edge for the remaining $w - p$ degree-1 nodes. This modifications help

to balance out the fact that some nodes of degree greater than 1 (in expectation) become degree-1 nodes in the final graph, so we need some of the degree-1 nodes (in expectation) to become higher degree in the final graph.

In Phase 2a, we set aside $q \leq p$ (q even) degree-1 vertices to be connected to other degree-1 vertices. This value can be specified by the user or defaults to

$$q = 2 \left\lfloor \frac{p^2}{2\sum_i d_i} \right\rfloor,$$

which is the expected number of degree-1-to-degree-1 edges expected in the CL model. This can be accomplished by randomly pairing the selected vertices. In all of our experiments, we used $q = 0$.

In Phase 2b, we manually connect the remaining $(p-q)$ vertices to the rest of the graph. For each degree-1 vertex, we select an endpoint proportional to e_i.

In Phase 2c, we finally create the CL model. We modify the expected degrees to account for the edges used in Phase 2b and to account for duplicates. Thus, we update $e_i \leftarrow \eta e_i$ where

$$\eta = 1 - 2\frac{p-q}{p-q+\sum_i e_i} + \beta,$$

where β is the proportion of duplicates. We use $\beta = 0.10$ in our experiments. The total number of edges generated in Phase 2c (including repeats and self-edges, which are discarded) is $\lfloor \sum_i e_i/2 \rfloor$.

VI. RESULTS

We consider comparisons of the BTER model with four real-world data sets from the SNAP collection [42]. All the graphs are treated as undirected. The number of nodes, edges, and the clustering coefficient of each graph are shown in Table I. We compare BTER with the real

TABLE I: Data sets for empirical validation

	Vertices	Edges	C
ca-AstroPh [22]	18,772	396,100	0.32
soc-Epinions1 [43]	75,879	811,480	0.07
cit-HepPh [44]	34,546	841,754	0.15
ca-CondMat [22]	23,133	186,878	0.26

data as well as the corresponding CL model as a baseline. The data are briefly summarized as follows:

- ca-AstroPh: Fig. 2 shows results on a collaboration network on 124 months of data from the astrophysics section of the arXiv preprint server. Here, the edge probabilities in the communities are given by (9) with $\rho = 0.95$ and $\eta = 0.05$.
- soc-Epinions1: Fig. 3 shows results on a who-trusts-whom online social (review) network from the Epinions website. Here, the edge probabilities in the communities are given by (9) with $\rho = 0.70$ and $\eta = 1.25$.
- cit-HepPh: Fig. 4 shows results from a citation network on the high energy physics phenomenology section of the arXiv preprint server. In this case, we use an alternate formula for ρ_r as follows:

$$\rho_r = 0.7\left[1 - 0.6\left(\frac{\log(\bar{d}_r+1)}{\log(d_{\max}+1)}\right)^3\right].$$

This is to ensure a faster decay of clustering coefficients.
- ca-ContMat: Fig. 5 shows results on a collaboration network on 124 months of data from the condensed matter section of the arXiv server. Here, the edge probabilities in the communities are given by (9) with $\rho = 0.95$ and $\eta = 0.95$.

In the leftmost plots of Figures 2–5, we see the comparison of the degree distributions. For ease of visualization, we have binned them logarithmically. As expected, both BTER and CL match the degree distribution, as they have been constructed to do so. The degree distributions do not necessarily conform to a standard degree distribution such as lognormal or power law. For example, the degree distribution for ca-AstroPh has a slight "kink" mid-way.

The difference between BTER and CL is highlighted when we instead consider the clustering coefficient, shown in the center plots of Figures 2–5. (Again, we have binned logarithmically by degree.) As noted previously, CL cannot have both a high clustering coefficient and a heavy tail, and this is evident in these examples. BTER, on the other hand, has a very close match with the observed clustering coefficients for all of these graphs. The dense ER graphs ensure that all nodes have high clustering coefficient. Note that the real graphs arise from diverse settings, and yet BTER matches the clustering coefficients quite well. We stress that this is a semilog plot, so the matches are extremely close.

The importance of matching the clustering coefficients becomes apparent when considering other features of the graph such as the eigenvalues of the adjacency matrix, as shown in the rightmost plots of Figures 2–5. For both ca-AstroPh, cit-HepPh, and ca-cond-Mat, the BTER eigenvalues are a much closer match than the CL eigenvalues because the community behavior is significant (respectively C is $0.32, 0.15, 0.26$). Indeed, barring the first two or three eigenvalues, BTER matches the remaining eigenvalues extremely closely. For soc-Epinions1, the difference between the models in terms of the eigenvalues is less dramatic because the community behavior is much less evident ($C = 0.07$); nonetheless, BTER is still a reasonable match.

The experimental results on these graphs are consistent with with our theory. The leftmost plots show that both CL and BTER can match the degree distributions of the original graph, as expected. Again, the clustering coefficients plots in the middle highlight the strengths of BTER, and how it differs from CL: BTER matches the clustering coefficients closely, while CL does not produce any significant number of triangles. The rightmost column shows that the eigenvalues of the adjacency matrices

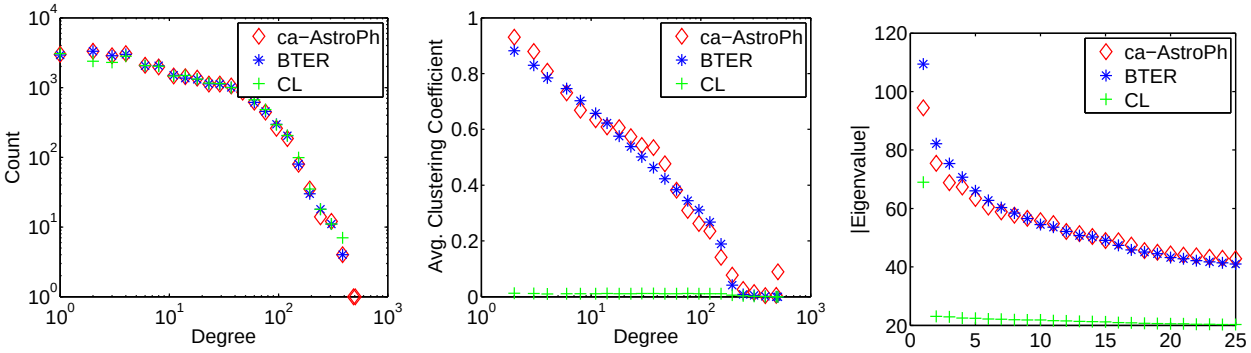


FIG. 2: (Color online) Properties of ca-AstroPh, a co-authorship network from astrophysics papers, compared with the BTER and CL models. Observe the close match of the clustering coefficients of the real data and BTER, in contrast to CL. Additionally, the eigenvalues of the BTER adjacency matrix are close to those of the real data.

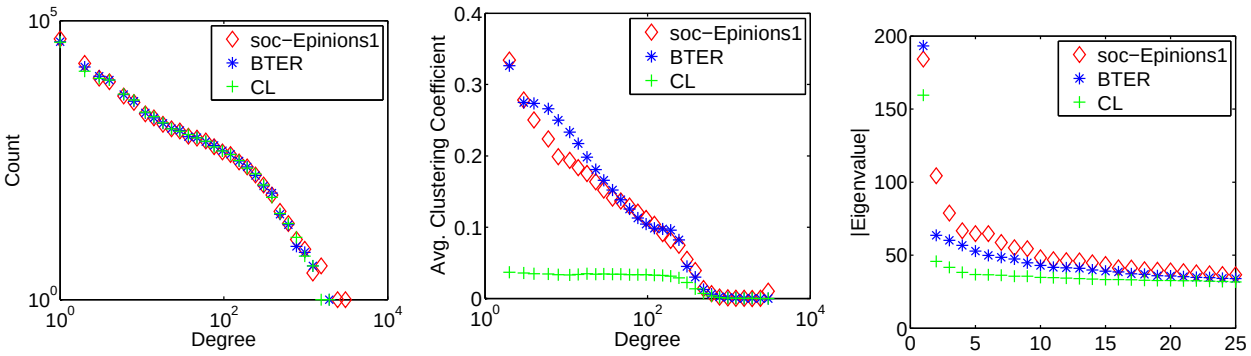


FIG. 3: (Color online) Properties of soc-Epinions1, a social network from the Epinions website, compared with the BTER and CL models. In this case, the clustering coefficients are much smaller overall, but the BTER model is still a closer match to the real data than CL in terms of both the clustering coefficient and the eigenvalues of the adjacency matrix.

of BTER are closer to those of the original graph than those produced by CL.

VII. DISCUSSION

We define a community to be a subgraph that is internally well-modeled by CL (and thus has no further substructure) and has many triangles. We prove that any community must contain a dense ER subgraph. Note that this automatically implies that these communities have a high density of links. Our alternative definition of community may help in simplifying and mathematically formalizing community structure.

Consider graphs with high clustering coefficients. Any graph model that captures community structure in these graphs must contain dense substructures in the form of dense ER graphs. This observation leads naturally to the BTER model, which explicitly builds communities of varying sizes and simultaneously generates a heavy tail.

Fitting the BTER model to real-world data is straightforward. The community sizes and composition in BTER are determined automatically according to the degree distribution. We currently assumes that all nodes in a community have (approximately) the same expected degree. Undoubtedly, this is an unrealistic assumption, but the variance of the model ensures that the degrees within a community vary considerably and Phase 2 adds connections between nodes of widely varying degrees. The connectivity of each ER block is a user-tunable parameter that can be adjusted to fit observed data. We currently prescribe a simple formula (9) and fit by trial and error, but the procedure could certainly be automated. Moreover, there is no particular requirement that ρ_r be exactly the same for all communities with the same $\bar{d}_r$ (minimum degree) nor that ρ_r be computed by a deterministic formula.

Our experimental results show that BTER-generated graphs have properties that are remarkably similar to real-world data sets. We contend that this makes BTER an appropriate model to use for testing algorithms and architectures designed for interaction graphs. In fact, BTER is even designed to be scalable. In particular, in Phase 2 we could compute the exact excess degree and

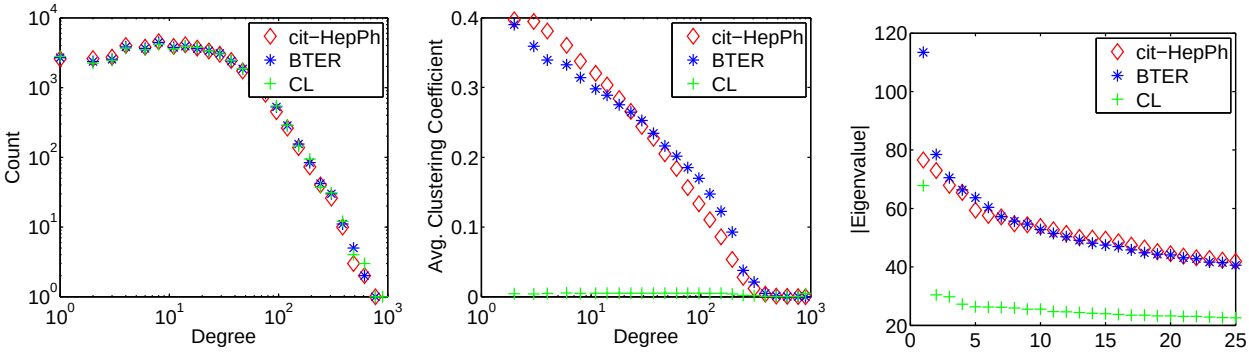


FIG. 4: (Color online) Properties of cit-HepPh, a citation network of High Energy Physics papers, compared with the BTER and CL models.

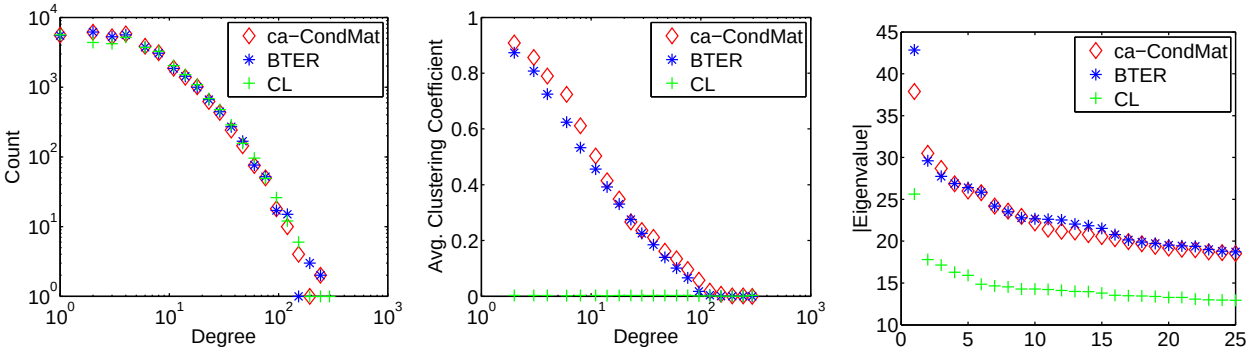


FIG. 5: (Color online) Properties of ca-CondMat, a co-authorship network of Condensed Matter physics, compared with the BTER and CL models.

use a matching procedure to complete the graph. The advantage of computing the excess degree in expectation is that it is more easily parallelized. In that case, the assignment to communities, the community connectivity, and the expected excess degree can all be computed in the preprocessing stage. Both Phase 1 *and* Phase 2 edges can be efficiently generated in parallel via a randomized procedure. Therefore, the BTER model is suitable for massive-scale modeling, such as that needed by Graph 500 [26]. The details of this implementation are outside the scope of the current discussion but will be considered in future work. As we mentioned earlier, BTER does not model disassortative graphs, and it would be very interesting to develop our theory for those graphs as well.

Another topic of future study is directly verify our theoretical hypothesis on real graphs by finding dense ER subgraphs. Our comparisons clearly show that BTER matches real graphs, so this suggests that dense ER subgraphs do exist as we propose. It is not obvious how to design algorithms for verifying the existence of dense ER graphs, but our proofs do suggest a scheme. In a community with s links, we expect $\sqrt{s}$ vertices of degree $\sqrt{s}$ (up to some constants) to form a dense subgraph. We may be able to empirically find such structures.

We feel that this work can play an important role in the algorithmic task of community detection, which is related to the problem of finding dense ER subgraphs. These dense ER subgraphs, according to our hypothesis, form the "heart" of communities. This could be a useful guide to current algorithms. For example, agglomerative community finding algorithms might want to use these dense subgraphs as seeds. More importantly, we may be able to use this theory to mathematically validate community detection algorithms.

Our formalism captures the more advanced notion of *link communities* [45] '(where edges, rather than vertices, form communities). This allows vertices to participate in many communities. The notion of communities uses modules over *internal degrees*, so one can easily imagine a vertex in many communities. Thm. 1 is still true, and we still get a scale-free collection of ER graphs which may share vertices. Thus, another interesting direction is to extend BTER to link (and hence overlapping) communities.

ACKNOWLEDGMENTS

This work was funded by the applied mathematics program at the United States Department of Energy and by

an Early Career Award from the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

We thank the anonymous reviewers for their helpful comments which greatly improved the presentation.

[1] M. Girvan and M. Newman, Proceedings of the National Academy of Sciences **99**, 7821 (2002).
[2] M. E. J. Newman, Proceedings of the National Academy of Sciences **103**, 8577 (2006).
[3] D. Watts and S. Strogatz, Nature **393**, 440 (1998).
[4] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Mathematics **6**, 29 (2009).
[5] A. Lancichinetti, M. Kivel, J. Saramki, and S. Fortunato, PLoS ONE **5**, e11976 (2010).
[6] F. R. K. Chung, *Spectral Graph Theory* (American Mathematical Society, 1992).
[7] F. Chung and L. Lu, Proceedings of the National Academy of Sciences **99**, 15879 (2002).
[8] F. Chung and L. Lu, Annals of Combinatorics **6**, 125 (2002).
[9] W. Aiello, F. Chung, and L. Lu, Experimental Mathematics **10**, 53 (2001).
[10] M. Newman, D. Watts, and S. Strogatz, Proceedings of the National Academy of Sciences **99**, 2566 (2002).
[11] A. Barrat and M. Weigt, The European Physical Journal B - Condensed Matter and Complex Systems **13**, 547 (2000).
[12] A. Arenas, A. Fernández, S. Fortunato, and S. Gómez, Journal of Physics A: Mathematical and Theoretical **41**, 224001 (2008).
[13] P. Erdös and A. Rényi, Publicationes Mathematicae **6**, 290 (1959).
[14] P. Erdös and A. Rényi, Publications of the Mathematical Institute of the Hungarian Academy of Sciences **5**, 17 (1960).
[15] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[16] A. Hernando, D. Villuendas, C. Vesperinas, M. Abad, and A. Plastino, European Physics Journal B **76** (2010), 10.1140/epjb/e2010-00211-5.
[17] B. Gonalves, N. Perra, and A. Vespignani, PLoS ONE **6**, e22656 (2011).
[18] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).
[19] A. Sala, H. Zheng, B. Y. Zhao, S. Gaito, and G. P. Rossi, in *PODC '10: Proceeding of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing* (ACM, 2010) pp. 400–401.
[20] D. Chakrabarti and C. Faloutsos, ACM Computing Surveys **38** (2006), 10.1145/1132952.1132954.
[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (2000) pp. 57–65.
[22] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Transactions on Knowledge Discovery from Data **1**, article 2 (2007).
[23] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao, in *WWW '10 Proceedings of the 19th International Conference on the World Wide Web* (ACM, New York, 2010) pp. 861–870.
[24] D. Chakrabarti, Y. Zhan, and C. Faloutsos, in *SDM '04: Proceedings of the 2004 SIAM International Conferen* (2004) pp. 442–446.
[25] D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, Journal of Machine Learning Research **11**, 985 (2010).
[26] Graph500, "Graph 500 benchmark," (2010), available at `http://www.graph500.org/Specifications.html`.
[27] C. Seshadhri, A. Pinar, and T. Kolda, in *ICDM'11: Proceedings of the 2011 IEEE International Conferen* (2011).
[28] A. Pinar, C. Seshadhri, and T. G. Kolda, in *SDM'11: SIAM Conference on Data Mining* (2012).
[29] G. Palla, L. Lovász, and T. Vicseka, Proceedings of the National Academy of Sciences **107**, 7640 (201
[30] S. J. Young and E. R. Scheinerman, in *WAW'07: Proceedings of the 5th International Conference on Al* Lecture Notes in Computer Science, Vol. 4863, edited by A. Bonato and F. Chung (Springer Berlin/Heidelberg, 2007) pp. 138–149.
[31] S. J. Young and E. R. Scheinerman, Internet Mathematics **5**, 91 (2008).
[32] A. Clauset, M. E. J. Newman, and C. Moore, Physical Review E **70**, 066111 (2004).
[33] A. Clauset, C. Moore, and M. E. J. Newman, Nature **453**, 98 (2008).
[34] M. Boguna, F. Papadopoulos, and D. Krioukov, Nature Communications **1**, 62 (2010).
[35] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguna, Physical Review E **82**, 036106 (2010).
[36] P. J. Bickel and A. Chen, Proceedings of the National Academy of Sciences **106**, 21068 (20
[37] B. Karrer and M. E. J. Newman, Physical Review E **83**, 21068 (2011).
[38] A. Lancichinetti, S. Fortunato, and F. Radicchi, Physical Review E **78**, 046110 (2008).
[39] J. B. Kruskal, in *Mathematical Optimization Techniques*, edited by R. Bellman (Cambridge University Press, London, 1963) pp. 251–278.
[40] G. Katona, in *Theory of Graphs: Proceedings of the Colloquium held at Tihany, Hungary, Sept. 1966* (Academic Press and Akadémiai Kaidó, Budapest, 1968) pp. 187–207.
[41] P. Frankl, Discrete Mathematics **48**, 327 (1984).
[42] SNAP, "SNAP: Stanford Network Analysis Project," Available at `http://snap.stanford.edu/`.
[43] M. Richardson, R. Agrawal, and P. Domingos, in *ISWC (2003). Proceedings of the Second International Semantic W* Lecture Notes in Computer Science, Vol. 2870 (Springer Berlin/Heidelberg, 2003) pp. 351–368.
[44] J. Gehrke, P. Ginsparg, and J. Kleinberg, ACM SIGKDD Explorations Newsletter **5**, 149 (2003).
[45] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature **466**, 761 (2010).