

Diversity of graphs with highly variable connectivity

David L. Alderson*

Operations Research Department, Naval Postgraduate School, Monterey, California 93943, USA

Lun Li[†]

Engineering and Applied Science, California Institute of Technology, Pasadena, California 91125, USA

(Received 24 August 2006; published 3 April 2007)

A popular approach for describing the structure of many complex networks focuses on graph theoretic properties that characterize their large-scale connectivity. While it is generally recognized that such descriptions based on aggregate statistics do not uniquely characterize a particular graph and also that many such statistical features are interdependent, the relationship between competing descriptions is not entirely understood. This paper lends perspective on this problem by showing how the degree sequence and other constraints (e.g., connectedness, no self-loops or parallel edges) on a particular graph play a primary role in dictating many features, including its correlation structure. Building on recent work, we show how a simple structural metric characterizes key differences between graphs having the same degree sequence. More broadly, we show how the (often implicit) choice of a background set against which to measure graph features has serious implications for the interpretation and comparability of graph theoretic descriptions.

DOI: 10.1103/PhysRevE.75.046102

PACS number(s): 89.75.Hc, 89.20.Ff

INTRODUCTION

The recent use of network models to describe complex systems has emphasized the study of graph theoretic properties as a means to characterize the similarities and differences in structure and function of systems across a variety of domains [1–7]. Considerable effort has been directed both at the empirical analysis of graph theoretic properties of real systems and at the development of generative models that attempt to explain such properties. An implicit assumption in much of this work is that graph theoretic properties adequately capture key system features in order to serve as a basis for comparison and contrast.

Notwithstanding the potential pitfalls of reducing a complex system (e.g., one that may involve heterogeneous components, layered architectures, and feedback dynamics) to a simple graph [8,9], there exists the practical problem that many descriptions based on aggregate statistics do not uniquely characterize the system of interest. In fact, there often exists considerable diversity among graphs that share any single statistical feature, particularly when viewed through the lens of a specific application domain. For example, recent work on the router-level Internet has shown that there is enough diversity among graphs having the same power-law node degree distribution that, although indistinguishable when viewed by this aggregate statistic, these graphs can actually be interpreted as “opposites” when viewed from an engineering perspective that incorporates technology constraints and is motivated by throughput performance [8,10,11].

The purpose of this paper is to explore this notion of graph diversity and characterize more completely the way in which the degree sequence of a particular graph dictates many popular graph features, including its correlation struc-

ture. Furthermore, this paper emphasizes the importance of choosing an appropriate “background set” when evaluating a graph, as well as the importance of making sure that the comparative analysis of two graphs is conducted with respect to an appropriate reference. In this regard, we show that not all graph theoretic measures have an obvious interpretation or are directly comparable.

DEGREE SEQUENCE AND GRAPH DIVERSITY

For a graph with n vertices, let d_i denote the degree (i.e., number of connections) of vertex i , $1 \leq i \leq n$, and call $D = \{d_1, d_2, \dots, d_n\}$ the *degree sequence* of the graph, assumed without loss of generality always to be ordered $d_1 \geq d_2 \geq \dots \geq d_n$. Within the space of all graphs having n vertices, let $\mathcal{G}(D)$ denote the considerably smaller subset of graphs having particular degree sequence D .

Not all sequences of integers D correspond to realizable graphs. One well-known characterization of whether or not a sequence D corresponds to a simple, connected graph is due to Erdős and Gallai [12], who observed that a sequence of positive integers d_1, d_2, \dots, d_n with $d_1 \geq d_2 \geq \dots \geq d_n$ is *graphical* if and only if $\sum_{i=1}^n d_i$ is even and for each integer k , $1 \leq k \leq n-1$,

$$\sum_{j=1}^k d_j \leq k(k-1) + \sum_{j=k+1}^n \min(k, d_j).$$

Recent work has further reduced the number of sufficient conditions to be checked [13], and several algorithms have been developed to test for the existence of a graph satisfying a particular degree sequence D [14].

The restriction to graphs having a particular degree sequence has been considered previously in the context of graph generation mechanisms [2,15]. In particular, the configuration model (CM) [2,16,17] often serves as the null hypothesis of networks having a particular degree sequence, since it yields graphs that are maximally random (in the

*Electronic address: dlalders@nps.edu

[†]Electronic address: lun@cds.caltech.edu

sense of maximum entropy) while conforming to a specified degree sequence D . In what follows, we will always restrict attention to graphs with a specified D .

In considering the structural features of a particular graph, we leverage previous work [18] and define, for any graph g having fixed degree sequence D , the s metric

$$s(g) = \sum_{(i,j) \in \mathcal{E}} d_i d_j = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}} \frac{1}{2} d_i a_{ij} d_j, \quad (1)$$

where $A = [a_{ij}]$ is the vertex adjacency matrix for the graph, and \mathcal{V} and \mathcal{E} denote the sets of all vertices and edges in the graph, respectively. Accordingly, we assume without loss of generality that the number of vertices and edges in the graph are represented by $n = |\mathcal{V}|$ and $l = |\mathcal{E}|$, respectively. Note that the summation in (1) is easily computed for any graph and does not depend on the process by which it was constructed. Implicitly, the metric $s(g)$ measures the extent to which the graph g has a hublike core and is maximized when high-degree vertices are connected to other high-degree vertices.

In general, the set $\mathcal{G}(D)$ will have many elements exhibiting a range of s values. Within this space, we define the s_{\max} and s_{\min} graphs within $\mathcal{G}(D)$ as those having the maximum and minimum $s(g)$ values, respectively. To facilitate the derivation of these values, we introduce the vector

$$\mathbf{Z} \equiv \left\{ \underbrace{\overbrace{d_1, \dots, d_1}^{d_1 \text{ elements}}, \underbrace{\overbrace{d_2, \dots, d_2}^{d_2 \text{ elements}}, \dots, \underbrace{\overbrace{d_n, \dots, d_n}^{d_n \text{ elements}}}_{\sum_{i=1}^n d_i \text{ elements}}} \right\} \quad (2)$$

which is simply derived from the original degree sequence D . The s_{\max} and s_{\min} values within $\mathcal{G}(D)$ can be described in terms of Z in the following manner. Since $\mathcal{G}(D)$ only requires its elements to satisfy the degree sequence D (and ignores issues such as connectedness, multiple edges, etc.) it is easy to show that within $\mathcal{G}(D)$ one has

$$s_{\max} \leq \frac{1}{2} \mathbf{Z} \mathbf{Z}^T, \quad (3)$$

with equality achieved in practice only under certain circumstances (e.g., when the elements of D are all even or there is an even number of elements having any particular odd value). This observation follows from the rearrangement inequality [19], which states that if $a_1 \geq a_2 \geq \dots \geq a_n$ and $b_1 \geq b_2 \geq \dots \geq b_n$, then for any permutation $(a'_1, a'_2, \dots, a'_n)$ of (a_1, a_2, \dots, a_n) , we have

$$\begin{aligned} a_1 b_1 + a_2 b_2 + \dots + a_n b_n &\geq a'_1 b_1 + a'_2 b_2 + \dots + a'_n b_n \\ &\geq a_n b_1 + a_{n-1} b_2 + \dots + a_1 b_n. \end{aligned}$$

Accordingly, it follows that

$$s_{\min} \geq \frac{1}{2} \hat{\mathbf{Z}} \hat{\mathbf{Z}}^T, \quad (4)$$

where $\hat{\mathbf{Z}}$ is simply the vector \mathbf{Z} with elements in reverse order. However, unlike the case in (3) where equality is achieved in practice only sometimes and where the actual

value may deviate considerably from the upper bound, the relationship in (4) holds with approximate equality and typically the s_{\min} value deviates from the lower bound by only a single pair of edges, if at all.

It is easy to see that the s_{\max} value can be rewritten as

$$s_{\max} \approx \sum_{i=1}^n (d_i/2)(d_i)^2 = \sum_{i=1}^n (d_i)^3/2, \quad (5)$$

which is achieved in effect by creating primarily self-loops among the vertices in the network and then connecting the remaining stubs in order of decreasing d_i (see Appendix A of [18] for details). To the best of our knowledge, there does not exist a comparable analytic formula (or interpretation) for the s_{\min} graph in $\mathcal{G}(D)$.

Many graphs of practical interest have additional conditions imposed by functional or domain constraints, such as a requirement to be connected or a restriction against self-loops or multiple connections. Thus, in our investigation we also consider the restricted set of all simple and connected graphs having the same degree sequence D , which we denote as $G(D)$. Note that $G(D) \subset \mathcal{G}(D)$ and that most randomly generated graphs with particular D will be neither simple nor connected, so this is an important and nontrivial restriction. From these definitions it follows that

$$\frac{1}{2} \mathbf{Z} \hat{\mathbf{Z}}^T \leq s_{\min}^{G(D)} \leq s_{\min}^{G(D)} \leq s_{\max}^{G(D)} \leq s_{\max}^{G(D)} \leq \frac{1}{2} \mathbf{Z} \mathbf{Z}^T.$$

Although bounding values for the minimum and maximum elements of $\mathcal{G}(D)$ can be directly obtained from Eqs. (3) and (4), obtaining s_{\max} and s_{\min} values within the restricted space $G(D)$ is more complicated.

Given a particular degree sequence D , it is possible to use a deterministic procedure in order to construct the s_{\max} graph in $G(D)$. The details of this construction procedure are presented in [18], but the basic idea is to order all potential links (i,j) for all $i, j \in \mathcal{V}$ according to their weights $d_i d_j$ and then add them one at a time in a manner that results in a simple, connected graph having degree sequence D . While simple enough in concept, this type of “greedy” heuristic procedure may have difficulty achieving the intended sequence D due to the global constraints imposed by connectivity requirements, but it works well in practice for most graphs (again, see [18] for details). Obtaining the s_{\min} value is less exact, and it is easy to show that the s_{\min} graph is not unique. Whitney and Alderson [20] have recently used a heuristic approach, originally proposed by Maslov and Sneppen [21], which employs a Metropolis-like algorithm based on successive rewiring to obtain s_{\min} values within $G(D)$. Unfortunately, this method is inefficient and does not reliably obtain the actual $s_{\min}^{G(D)}$ value. However, in practice one finds that $\frac{1}{2} \mathbf{Z} \hat{\mathbf{Z}}^T \approx s_{\min}^{G(D)} \approx s_{\min}^{G(D)}$, so in the remainder of this paper we use the $s_{\min}^{G(D)}$ value defined in (4), as an approximate (and more conservative) bounding value for $s_{\min}^{G(D)}$.

As a measure of graph structure, the s metric provides a simple means for contrasting the differences between graphs having the same degree sequence, and in this paper we use it exclusively as a means for measuring the diversity within a

particular space of graphs. In particular, the extreme points s_{\max} and s_{\min} serve as meaningful reference points for individual graphs and the space as a whole, and for a given D the difference $s_{\max} - s_{\min}$ provides a measure of how different the absolute extremes are. Using this perspective, it is not hard to see that the amount of diversity for graphs having a particular D is related to the amount of variability within the sequence D itself. Here, we characterize variability with the standard measure of (sample) coefficient of variation (C_V), which for a given sequence $D = (d_1, d_2, \dots, d_n)$ is defined as

$$C_V(D) = \sigma(D)/\langle d \rangle, \quad (6)$$

where $\langle d \rangle = n^{-1} \sum_{k=1}^n d_k$ is the average vertex degree, and we measure deviations of the d_i from its average $\langle d \rangle$ using the sample standard deviation $\sigma(D) = [\sum_{k=1}^n (d_k - \langle d \rangle)^2 / (n-1)]^{1/2}$.

For graphs with regular structure that have low variability in their degree sequence D , there is typically very little diversity in the corresponding space of graphs $G(D)$. Consider as an extreme example, a one-dimensional lattice (i.e., a chain) with the degree sequence $D_{\text{chain}} = \{2, 2, 2, \dots, 2, 1, 1\}$. One can easily show that for a chain consisting of n nodes

$$C_V(D_{\text{chain}}) = \frac{n^{1/2}(n-2)^{1/2}}{2^{1/2}(n-1)^{3/2}}$$

and thus $C_V(D_{\text{chain}}) \rightarrow 0$ as $n \rightarrow \infty$. It is easy to see that there is no diversity among graphs having degree sequence D_{chain} , since all n -node chains are isomorphic to one another in $G(D)$ and thus $s_{\min} = s_{\max}$.

For sequences D with increasing $C_V(D)$, graph diversity as measured by the range $s_{\max} - s_{\min}$ also increases. Here, we leverage two classes of graphs as reference points. For graphs with a degree sequence having an exponential form, $ke^{\lambda d_k} \approx c$ for constant $c > 0$ (denoted here as D_{exp}), one observes that $C_V(D_{\text{exp}}) \rightarrow \kappa$ (a constant) as $n \rightarrow \infty$. In contrast, the scale-free graphs [22]—so called because their degree sequences exhibit a scaling relationship of the form $kd_k^\alpha = c$, for all $1 \leq k \leq n_s$, where $c > 0$ and $\alpha > 0$ are constants, and where n_s determines the range of scaling [23]—exhibit divergent C_V . It is easy to show that degree sequences D_{scaling} with $\alpha < 2$ follow $C_V(D_{\text{scaling}}) \rightarrow \infty$ as $n \rightarrow \infty$. As we will show below, these classes of graphs yield degree sequences with measurably different levels of diversity.

Although one might expect that graph diversity simply increases with $C_V(D)$, this need not be the case. Consider a star consisting of a single central node that connects to all others and having degree sequence $D_{\text{star}} = \{n-1, 1, 1, \dots, 1\}$. One can similarly show that

$$C_V(D_{\text{star}}) = \frac{n^{1/2}(n-2)}{2(n-1)},$$

and thus $C_V(D_{\text{star}}) \rightarrow \infty$ as $n \rightarrow \infty$. However, as for the chain, there is no diversity among graphs having degree sequence D_{star} [i.e., all stars are isomorphic to one another in $G(D)$ and $s_{\min} = s_{\max}$].

In order to make the previous discussion more concrete, we now consider a simple experiment to investigate the role of $C_V(D)$ in determining the diversity for graphs having par-

ticular D . For purposes of exposition, we begin with a study of acyclic graphs (i.e., trees) and then later comment on how our results apply to general graphs. Our experiment uses incremental growth via preferential attachment as described in [24], in which each newly added node connects to an existing node k with probability

$$\Pi(k) = b \frac{(d_k)^p}{\sum_j (d_j)^p}, \quad (7)$$

where d_k is again the degree of node k , and p is a parameter that tunes the attachment mechanism. The resulting graph is simple and connected, and thus an element of $G(D)$, although the degree sequence D that is realized will vary from trial to trial. Clearly, $p=0$ is equivalent to uniform attachment (resulting in D_{exp}), while $p=1$ is equivalent to linear preferential attachment used in the Barabási-Albert model [3] (resulting in D_{scaling}). A similar type of model was also considered in [25]. Note also that as $p \rightarrow \infty$ each newly added node attaches to the maximum degree node (resulting essentially in D_{star}), while as $p \rightarrow -\infty$ each newly added node attaches to the minimum degree node (resulting essentially in D_{chain}). In what follows, we first restrict attention to the case where $b=1$ (i.e., we generate acyclic graphs) and consider a range of values for p in order to generate graphs having a variety of degree sequences. We defer results on general graphs until the end.

Figure 1 shows the result of an experiment in which for each trial we generate a tree having $n=100$ nodes using preferential attachment rule given by (7). That is, each trial results in a tree having its own degree sequence D and s value. In generating these graphs, we use various attachment exponents p , but only for the purpose of realizing graphs with a diversity of degree sequences. In what follows we focus primarily on the degree sequence D and the constraints it places on the space of graphs, not the attachment exponent p that led to D . For each degree sequence D , we then calculate $C_V(D)$ as well as the corresponding s_{\max} and s_{\min} values as described above. The resulting picture in Fig. 1(a) shows a striking relationship between $C_V(D)$ and the range of possible s -values. One observes that, while the s_{\max} and s_{\min} values increase with $C_V(D)$ for both the unconstrained space $\mathcal{G}(D)$ and the constrained space $G(D)$, the differences given by $s_{\max} - s_{\min}$ for each space behave differently at the maximal values of $C_V(D)$. Specifically, this difference within the unconstrained space $\mathcal{G}(D)$ increases with $C_V(D)$, but it is zero at both extremes of $C_V(D)$ for the simple, connected graphs in $G(D)$ (again, the limiting cases of a chain and a star). It is also worth noting that the values for $s_{\min}^{\mathcal{G}(D)}$ and $s_{\max}^{\mathcal{G}(D)}$ are so close as to be indistinguishable, further supporting our choice to treat these values as equivalent. Figure 1(b) presents the same information for s_{\max} and s_{\min} within $G(D)$, but normalizes the s values for each graph against its respective s_{\max} value, thus resulting in a feasible range [0, 1] for each graph. Collectively, this suggests that for a given degree sequence one needs enough variability to enable diversity among simple, connected graphs but that too much variabil-

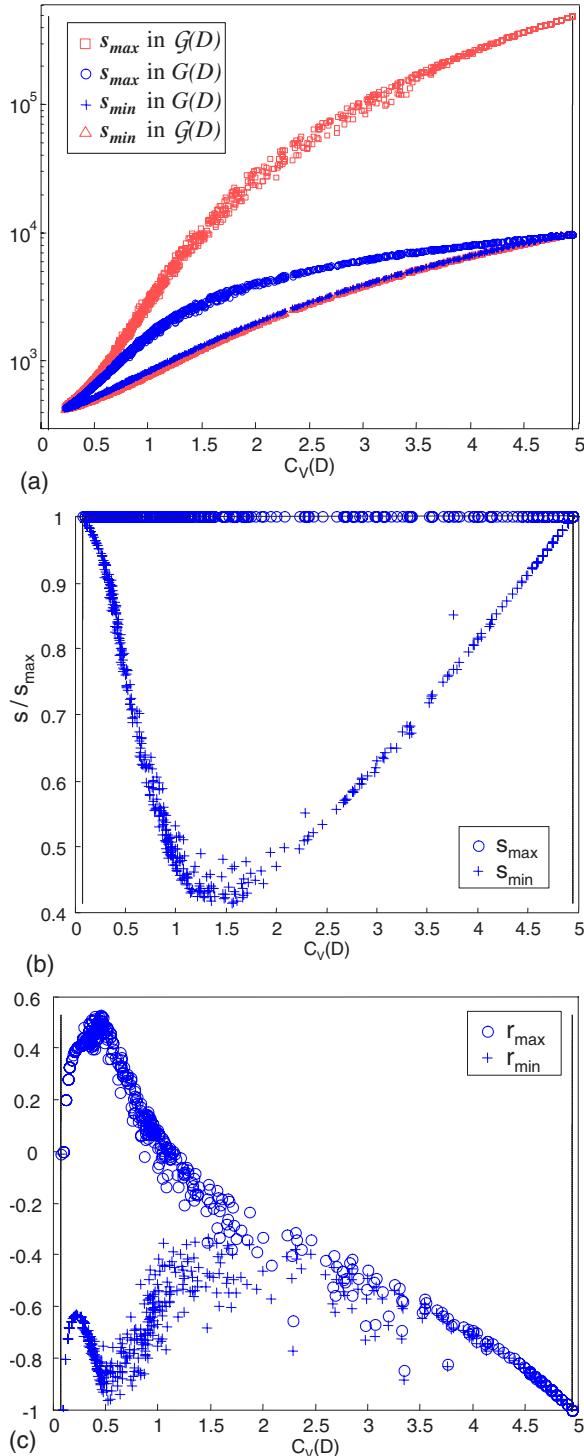


FIG. 1. (Color online) Three views of graph diversity. In this experiment trees of size $n=100$ were generated according to attachment rule (7) for different values of p . (a) For each resulting tree, we plot s_{\min} and s_{\max} values in both $\mathcal{G}(D)$ and $G(D)$ versus the $C_V(D)$ of the corresponding degree sequence. Note that $s_{\min}^{G(D)} \approx s_{\min}^{\mathcal{G}(D)}$. (b) The s_{\min} and s_{\max} in $G(D)$, each normalized by their respective s_{\max} . (c) The corresponding r_{\min} and r_{\max} values in $G(D)$. In all cases, the vertical lines correspond to the upper and lower limits of C_V for an acyclic graph having 100 nodes [i.e., $C_V(D_{\text{chain}})=0.0711$ and $C_V(D_{\text{star}})=4.9495$].

ity actually becomes a constraint within the space $G(D)$, something that Maslov *et al.* [26] have described as essentially a finite-size effect.

Although it is now well understood that there can be many graphs having the same degree sequence and that these graphs may have considerable structural differences, quantifying these differences and their implications in terms of real systems remains the topic of active research. Previous work by Li *et al.* [18] has shown that the s metric, and in particular the s_{\max} graph within $G(D)$, is relevant for many commonly studied graph properties. First, high-degree nodes in the s_{\max} graph have high centrality, and for trees this relationship was shown to be monotonic. Second, s_{\max} graphs are self-similar under appropriately defined operations of trimming and coarse graining. Finally, the s_{\max} graph has the highest likelihood of being generated by the generalized random graph (GRG) model [15]. As already noted, other work by Li *et al.* [10] has shown that, in modeling the router-level Internet, the observed degree sequences in real networks allow for dramatic diversity in candidate models, particularly when measured in terms of throughput performance. A previously unanswered question was whether this diversity is inherent in all networks, and here we have shown that it depends to some extent on the degree sequence of the network in question.

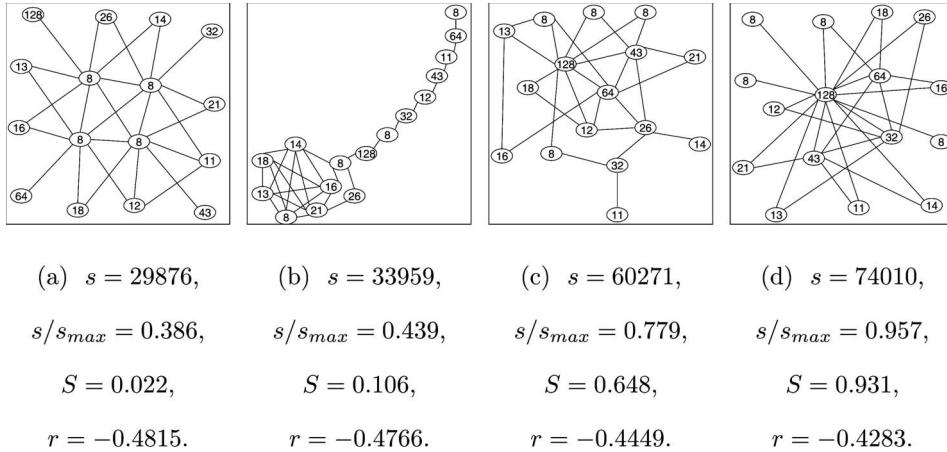
Taken by itself, this observation is neither groundbreaking nor surprising. For some time, there has been a general recognition in the literature that the degree sequence of a graph can provide only a simplistic characterization of its properties, and this has led many researchers to consider more sophisticated descriptions of graph structure. Most notable has been an emphasis on various forms of correlation in network connectivity, ranging from simple notions of network clustering (i.e., connectivity correlations between vertex triplets) to more general degree-degree correlations [also called the joint degree distribution (JDD)] and spectral methods. There is now a growing literature on the importance of correlation structure in networks [2,27–31] and how to generate networks having particular correlation structure [25,32–34]. A simple measure of correlation structure that has appeared extensively in the literature is the Pearson coefficient r (known more generally as the correlation coefficient [35]) which is used to quantify the average tendency of vertices to connect to others having similar degree. It turns out that there is an inherent relationship between the Pearson coefficient and the s metric, and a closer look at this relationship yields considerable insight into both the diversity within the background set $G(D)$ as well as the interpretation of r itself.

GRAPH ASSORTATIVITY RECONSIDERED

Recently, Newman [36] introduced the following sample-based measure of graph assortativity as defined by the Pearson coefficient:

$$r(g) = \frac{\left(\sum_{(i,j) \in \mathcal{E}} d_i d_j / l \right) - \left(\sum_{(i,j) \in \mathcal{E}} \frac{1}{2} (d_i + d_j) / l \right)^2}{\left(\sum_{(i,j) \in \mathcal{E}} \frac{1}{2} (d_i^2 + d_j^2) / l \right) - \left(\sum_{(i,j) \in \mathcal{E}} \frac{1}{2} (d_i + d_j) / l \right)^2}. \quad (8)$$

This relationship can be written as



$$r(g) = \frac{\left(\sum_{(i,j) \in \mathcal{E}} d_i d_j \right) - \left(\sum_{i \in \mathcal{V}} \frac{1}{2} d_i^2 \right)^2 / l}{\left(\sum_{i \in \mathcal{V}} \frac{1}{2} d_i^3 \right) - \left(\sum_{i \in \mathcal{V}} \frac{1}{2} d_i^2 \right)^2 / l}, \quad (9)$$

where the first term of the numerator is exactly $s(g)$. Although the Pearson coefficient is only a summary statistic for the correlation profile of the graph as a whole, it provides interesting information nonetheless and is often cited as a key feature distinguishing various classes of complex networks [4,27,36,37].

Here, we argue that $r(g)$ has a natural interpretation as a centered and normalized version of $s(g)$. In particular, observe that the first term of the denominator in (9) is exactly the s_{\max} value within the space $\mathcal{G}(D)$ as defined in (5). Accordingly, one can rewrite the Pearson coefficient as

$$r(g) = \frac{s(g) - s(g_c)}{s_{\max}^{\mathcal{G}(D)} - s(g_c)}, \quad (10)$$

where we refer to g_c as the center of the space $\mathcal{G}(D)$.

The reason that g_c can be viewed as the center of this space of graphs is discussed in the online supplement to our previous work [18]. The key idea is that a deterministic graph in $\mathcal{G}(D)$ with zero assortativity has exactly the same s value as $s(g_c)$, equal to $l^{-1}(\sum_{i \in \mathcal{V}} \frac{1}{2} d_i^2)^2$. More specifically, constructing such a deterministic graph with zero assortativity means connecting a vertex to any other vertex in a manner that is proportional to each vertex's degree. This can be realized using a pseudograph \tilde{g}_A in which the elements of the adjacency matrix $A = [a_{ij}]$ are non-negative real numbers representing link weights and satisfying

$$a_{ij} = \frac{d_i d_j}{\sum_{k \in \mathcal{V}} d_k} = a_{ji}.$$

By extension, the s metric for the pseudograph \tilde{g}_A is calculated as

$$s(\tilde{g}_A) = \sum_{j \in \mathcal{V}} \sum_{i \in \mathcal{V}} \frac{1}{2} d_i a_{ij} d_j = \frac{\left(\sum_{j \in \mathcal{V}} \frac{1}{2} d_j^2 \right)^2}{l},$$

showing that $s(\tilde{g}_A) = s(g_c)$. Note that the GRG method [15] can be interpreted as a stochastic procedure that generates real graphs from the zero-assortativity pseudograph \tilde{g}_A , with the one important difference that the GRG method always results in simple (but not necessarily connected) graphs. It has recently been shown that the statistical ensemble of graphs resulting from the stochastic GRG method has zero assortativity [39].

Thus, the Pearson coefficient r (as a summary statistic of graph assortativity) captures a fundamental feature of graph structure, one that is closely related to our s metric.¹ That r reflects s is obvious from its definition, but the question is whether a consideration of s by itself provides insight. The key observation is that the existing notion of assortativity for an individual graph g is implicitly measured against a background set of graphs $\mathcal{G}(D)$ that is *not* constrained to be either simple or connected. As we show next, because r is computed relative to an unconstrained background set, in some cases this normalization (against the unconstrained s_{\max} graph) and centering (against the \tilde{g}_A pseudograph) does a relatively poor job of distinguishing among graphs having the same degree sequence, particularly when that degree sequence exhibits high variability. Figure 2 shows four graphs having the same degree sequence, but with very different connectivity patterns. These graphs were originally constructed as contrasting representations of the router-level Internet (see [18], Fig. 5), but are presented here in a manner that highlights their diversity. Specifically, one observes that although they have nearly the same assortativity as defined by r , their structural differences are highlighted by s and its

¹Indeed, the Pearson coefficient is typically viewed as simply the correlation coefficient of the joint distribution $P(k, k')$ that a randomly selected link in the network will connect vertices having degree values k and k' . In this context, the “centering term” is simply the squared average of the marginal distribution of $P(k, k')$, and the denominator of (8) is the square root of the standard deviation.

TABLE I. Sensitivity of assortativity among graphs having low $C_V(D)$. Each graph shown has n nodes and $n-1$ links. In the limit where $n \rightarrow \infty$, minimal differences in graph structure, as measured by $C_V(D)$ and the ratio $s/s_{\max}^{G(D)}$, translate to large differences as measured by the Pearson coefficient r .

	$C_V(D)$	s / s_{\max}	r
	≈ 0	≈ 1	≈ 0
	≈ 0	≈ 1	≈ -1
	≈ 0	≈ 1	≈ 0

normalized values $s/s_{\max}^{G(D)}$ and $S(g)$ defined as

$$S(g) = \frac{s(g) - s_{\min}}{s_{\max} - s_{\min}}. \quad (11)$$

In cases where network performance is measured by the maximum throughput under fixed node capacities, these structural differences translate to big differences in performance [18].

For additional insight into the way in which differences in s translate to differences in r , we extend the previous computational experiment to values of r_{\max} and r_{\min} within the constrained background set $G(D)$. Note that these values can be computed directly from the corresponding values of s_{\max} and s_{\min} . In Fig. 1(c) we show these values for each of the generated graphs in our experiment. There are several striking features of this plot. The first is that the normalization of the s metric in the calculation of the Pearson coefficient r dramatically changes the sense of graph diversity among graphs having a particular D . For values of relatively high $C_V(D)$, $r < 0$ and seems largely independent of any diversity as measured by the range in allowable s . In other words, a second important conclusion is that all networks with high $C_V(D)$ have $r < 0$ and this seems largely a function of D and not any particular feature of the graph or whether it is a technological or social network as argued in [37]. This idea has been made previously in [7,26,29,33,38] and has also been recently argued [20] based largely on empirical observations of real networks having a range of r values. A third important result is that for small values of $C_V(D)$ one observes that small diversity as measured by $s_{\max} - s_{\min}$ translates to a large range of $r_{\max} - r_{\min}$. One can see this more clearly with the simple example in Table I, which illustrates the sensitivity of r to small changes in topology. Thus, for graphs that are simple and connected, the Pearson coefficient r can both hide structural diversity as well as display false diversity.

It is worth noting that although $r(g)=1$ is achieved approximately by the s_{\max} graph within $G(D)$ for all graphical D , it is only in very special instances of D where the s_{\min} graph is obtained. Specifically, when $s_{\min} = \mathbf{Z}\hat{\mathbf{Z}}^T$, then it follows that $r(g)=-1$ if and only if $z_k + \hat{z}_k = z$ (a constant) for each of the k pairs of elements. In other words, although it is

true that $r_{\max}=1$ for arbitrary D , one often observes that $r_{\min} \gg -1$ simply because of the degree sequence D itself. A proof of this appears in the Appendix.

Based on this analysis, one might reasonably conclude that the Pearson coefficient r is not a suitable metric for comparing the correlation structure of graphs from different domains. Indeed, it is well understood that a more accurate approach is to consider higher-order forms of correlation. Yet the deeper question relates to how one should evaluate any observed correlation structure. Recent efforts by several authors have warned against graph theoretic analysis of networks in isolation. For example, Maslov *et al.* [21,26] have argued that a real assessment of a network's correlation structure makes sense only when compared against its randomized counterpart. In the context of rich-club ordering in complex networks (i.e., the tendency of high-degree vertices to connect to one another), Colizza *et al.* [40] have also argued that the presence of high-degree vertices in a given network is enough to ensure that high-degree vertices are connected, and they similarly argue for the need to compare the features of any subject network to a randomized baseline. Thus, important questions include: What is the appropriate baseline against which to compare graphs? and how does this relate to the background set of graphs, as defined by $G(D)$ or $\mathcal{G}(D)$?

MEASURING AGAINST BACKGROUND SETS

The previous sections provide enhanced understanding of the way in which a given D constrains the possible s and r values a graph can have, and they also suggest that when making statements about a graph based on these graph properties one must consider the background set against which these properties are being evaluated. Here, we expand this viewpoint by considering the way in which a graph with given D compares within the space of graphs bounded by s_{\min} and s_{\max} values. We furthermore consider the location of randomized graphs within this space.

As above, our approach here is largely empirical, and we again leverage our previous numerical experiment in generating graphs via incremental growth according to an attachment exponent p . For a given value of p , we generate a graph having n vertices with resulting degree sequence D . Then, for that particular D we construct the s_{\min} and s_{\max} graphs

within $G(D)$. We also compute the theoretical upper bound (3) and lower bound (4) on s within $\mathcal{G}(D)$. We then obtain appropriately randomized graphs having degree sequence D in two ways. First, we generate $m=500$ new graphs according to the configuration method. Also, we consider the process of degree-preserving rewiring on the original graph.

Graph rewiring is effective as a conceptual, as well as computational, means for exploring the space of graphs having the same degree sequence D . Since exchanging the end points of any two links does not alter the degrees of the affected vertices (and hence leaves the overall degree sequence unchanged), this approach has been a popular tool for investigating the effects of local topological changes on global graph properties [18,21,26] as well as a means for generating graphs having a specified degree sequence and additional properties (i.e., connectedness) [41,42]. Here, we consider degree-preserving rewiring as a means for moving within the space of graphs having degree sequence D . In previous work [18], we have used the number of successive rewiring steps between two graphs as a measure of distance in the space $G(D)$; however, in this study we restrict attention to the distribution of s values within the possible range $s_{\max} - s_{\min}$ for both $G(D)$ and $\mathcal{G}(D)$. In the aforementioned extreme examples of a chain and star, any degree-preserving rewiring operation that precludes disconnection or self-loops yields a graph that is isomorphic to the original, and again shows that there is no diversity in either case.

Figure 3 shows the results of three representative numerical experiments exploring the distribution of graphs having particular s -values for a specified D . Figure 3(a) resulted from uniform attachment (i.e., $p=0$) and corresponds to the case of D_{exp} having low variation [here, $C_V(D)=0.6380$ within the possible range [0.0711, 4.9495] for acyclic graphs having $n=100$ nodes]. Figure 3(b) resulted from linear preferential attachment ($p \approx 1$) and corresponds to the case of D_{scaling} [here, $C_V(D)=1.4121$]. Figure 3(c) resulted from superlinear preferential attachment (i.e., $p > 1$) and corresponds to a case with high variability [here, $C_V(D)=2.5141$]. For each case, the s_{\max} graph within $G(D)$ was obtained by the construction mechanism described previously, while the s_{\min} value was obtained from (4). The leftmost graph for each case corresponds to an approximate s_{\min} graph obtained heuristically. From these results, several observations are immediately clear.

(1) For each particular D , there are considerable differences between the s_{\min} and s_{\max} graphs. In all cases, the s_{\min} graph looks very chainlike and the s_{\max} graph looks very starlike.

(2) The range of feasible s values for graphs in $\mathcal{G}(D)$ is considerably larger than the range for $G(D)$, and this difference increases with greater $C_V(D)$.

(3) The differences between the graphs in each case are less obvious when evaluated using the Pearson coefficient r [normalized against the graphs in the unconstrained space $\mathcal{G}(D)$] but are emphasized when evaluated using normalized s values (i.e., either s/s_{\max} or S). Thus, when comparing among elements of $G(D)$, the Pearson coefficient sometimes tends to hide the structural differences rather than highlight them. Similar observations were made previously in [18,20].

(4) Although rewiring within the space $G(D)$ yields a distribution of graphs that theoretically span the entire space, using rewiring to obtain graphs having extreme s values is difficult to achieve in practice. The implications for using rewiring as a means to obtain an ensemble of graphs is unclear. Moreover, it is unclear what, if anything, one can say about the original graph for each case based on its placement within the feasible range of graphs for $G(D)$.

(5) As expected, there is good correspondence in all cases between the distribution of graphs resulting from rewiring in the unconstrained space $\mathcal{G}(D)$ and those generated from the configuration method. Furthermore, the distribution of these graphs appears largely centered on $r=0$, as would be predicted since it was shown that the CM approach results in zero-assortativity graphs (in expectation).

(6) The distribution of graphs in $\mathcal{G}(D)$ is consistently shifted toward larger s values than those in $G(D)$. As C_V increases, the differences between the distribution of graphs in $G(D)$ and $\mathcal{G}(D)$ becomes more extreme, to the point where all of the graphs generated within $\mathcal{G}(D)$ have s values larger than can be achieved by the s_{\max} graph of $G(D)$. In other words, for large- C_V degree sequences, none of the graphs generated by the CM or resulting from rewiring within $\mathcal{G}(D)$ correspond to simple, connected graphs [i.e., elements in $G(D)$].

In practice, when considering graphs having high C_V , we advocate the use of $s/s_{\max}^{G(D)}$ or S as measures of diversity when considering graphs that are simple and connected. For graphs that are not simple or connected, the Pearson coefficient r provides insight into the diversity within $\mathcal{G}(D)$.

These observations yield several important conclusions.

First, graphs that arise from different contexts may not be directly comparable using structural metrics that are inherently computed against different background sets. In considering the above examples, one observes that the approximate s_{\min} graph in Fig. 3(b) [i.e., $C_V(D)=1.41$] translates to $r=-0.45$ while the s_{\max} graph for Fig. 3(c) [i.e., $C_V(D)=2.51$] translates to $r=-0.43$. A naive look at the Pearson coefficient suggests that they are similarly assortative, although the graph in Fig. 3(b) has the minimal r value and the graph in Fig. 3(c) has the maximal r value.

Second, the differences between the unconstrained space $\mathcal{G}(D)$ and the space of simple, connected graphs $G(D)$ may be more important in determining graph properties than other features as measured by aggregate statistics. Specifically, the use of graph generation techniques such as the configuration method, even if they replicate the measured degree sequence of a real network, may be entirely inappropriate if the domain under study requires simple and connected graphs. This strengthens previous results on the importance of these additional restrictions as reported in [26,43].

Third, while it is clear that the evaluation of a graph based on its structural properties may be appropriate only in relation to the corresponding background set, understanding the implication of those structural features (e.g., in terms of function) remains an open question. For example, it remains unclear what, if anything, the relative placement of a graph within the range $[s_{\min}, s_{\max}]$ actually says about the graph itself.

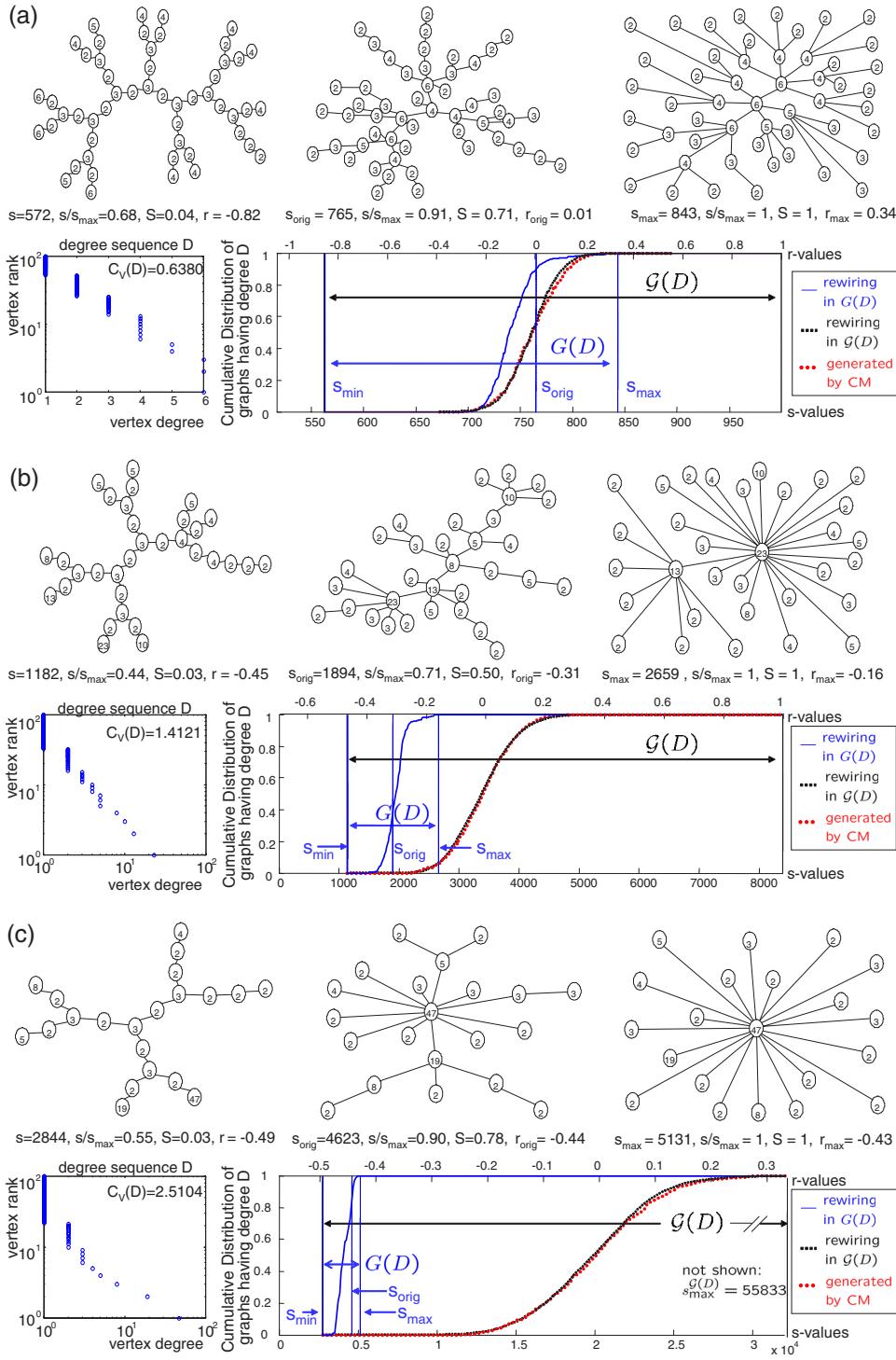


FIG. 3. (Color online) Diversity among graphs having the same degree sequence: (a) uniform attachment ($p=0$), (b) approximate linear attachment ($p \approx 1$), (c) superlinear attachment ($p > 1$). In each case, a single graph with $n=100$ vertices was generated using a different preferential attachment exponent and results in a different degree sequence D . The corresponding s_{\min} and s_{\max} graphs were also obtained for both $G(D)$ and $\mathcal{G}(D)$. Each node is labeled with its degree, with degree-1 nodes omitted for simplicity. Also shown for each is the distribution of graphs within the space $G(D)$ (from rewiring) and within $\mathcal{G}(D)$ (from rewiring and generated via the CM).

DISCUSSION

An inherent challenge in the study of graph diversity is that the combinatorics of even relatively small networks typically result in a space of graphs that is incredibly large. In this study, we have focused on graphs having $n=100$ (which are about the largest that can be visualized easily) for purposes of exposition, and even here a comprehensive analysis of the elements in $\mathcal{G}(D)$ and $G(D)$ is challenging. In choosing preferential attachment as our primary means for graph generation, we have tried to keep our methods closely

tied to the literature so that they may be easily replicated. An alternate approach could have been to identify specific degree sequences D for which graph isomorphism reduces the number of unique graphs to a small handful and the entire space of graphs (not just s_{\max} and s_{\min}) is easily visualized. Identifying and exploring such examples may represent an important step in future work.

The overall message of the results here is that one must carefully consider the inherent diversity of graphs sharing a particular statistical measure when making claims based on

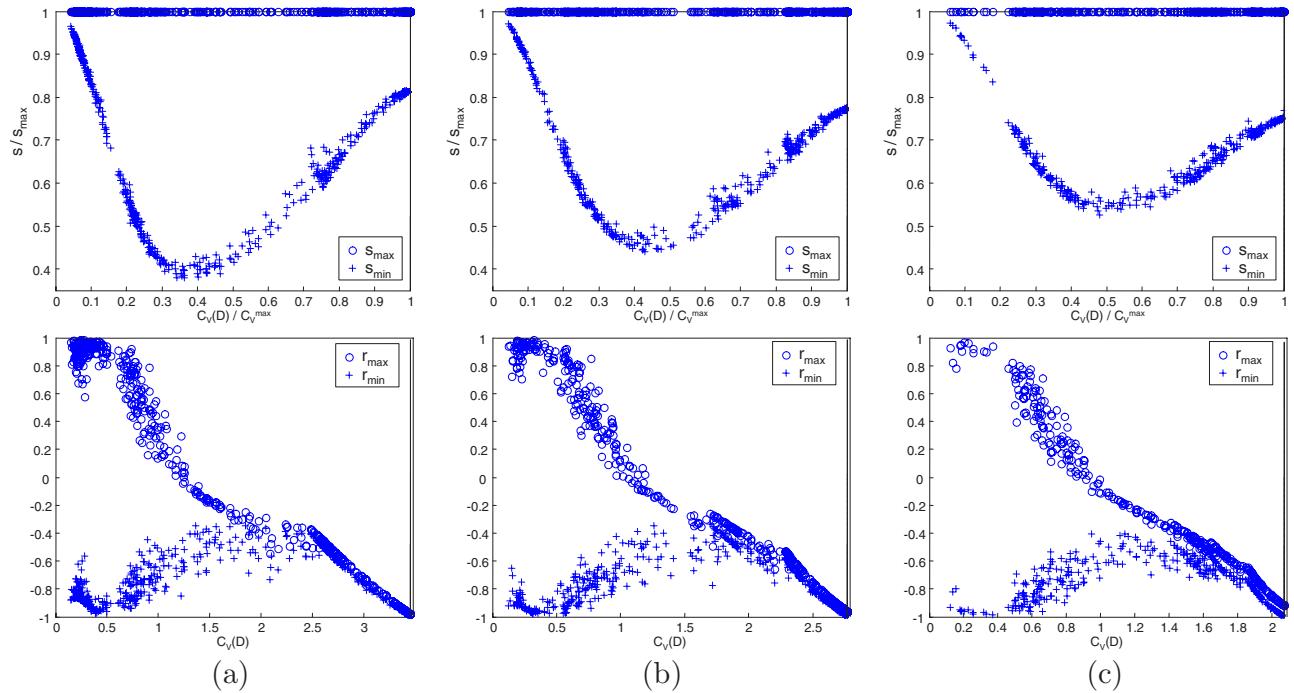


FIG. 4. (Color online) Graph diversity among nontrees. In this experiment, an additional $k(n-1)$ links were added to initial trees of size $n=100$. (a) $k=1$, $\langle d \rangle = 3.96$, $C_V^{\max} = 3.4451$. (b) $k=2$, $\langle d \rangle = 5.94$, $C_V^{\max} = 2.7672$. (c) $k=4$, $\langle d \rangle = 9.9$, $C_V^{\max} = 2.0701$. In the bottom graphs, variation is measured with $C_V(D)$ while in the top graphs it is represented as the normalized $C_V(D)/C_V^{\max}(D)$.

any such statistic. Nonetheless, additional work is required to understand fully the way in which graph diversity affects such characterizations. While others have argued for the need to compare against a randomized version of the graph, here we have compared against the entire feasible region, as measured by the range $[s_{\min}, s_{\max}]$. The examples here seem to suggest that the distribution of graphs within either $G(D)$ or $\mathcal{G}(D)$ is not uniform, and a general characterization of these distributions is unknown. Ideally, one would like to know more about where the randomized graph sits within the overall space (i.e., is it the center of this space?). Moreover, there may be important differences between graph properties that are imposed by structural constraints (e.g., by the degree sequence D) and those relative to what has been randomized.

Although this study provides additional insight into the way in which graph diversity affects one's ability to use aggregate statistics for characterizing complex networks, it has done so primarily for acyclic graphs (i.e., trees), and more work is required to understand the extent to which these same results hold for more general network structures. However, we now present preliminary empirical evidence that suggests the story for nontrees is qualitatively the same.

In Fig. 4, we show the results of a final experiment in which we again generate trees having $n=100$ nodes according to attachment rule (7) for a range of exponents p . However, to each tree having an initial $l=n-1$ links we then add an additional kl links by choosing end points probabilistically in correspondence with (7). In this manner, we generate graphs having n nodes and a degree sequence D satisfying $\sum_i d_i = 2(k+1)(n-1)$ [i.e., the average degree is $\langle d \rangle \approx 2(k+1)$]. Empirical evidence [4] suggests that, for many real networks, $\langle d \rangle < 10$. For each degree sequence D , we then

compute the corresponding s_{\min} , s_{\max} , r_{\min} , and r_{\max} values as was done previously. Figure 4 shows these values plotted against the variation of D , represented again as $C_V(D)$ and also now normalized as $C_V(D)/C_V^{\max}(D)$ for purposes of comparison.

One observes for graphs with increasing average degree [$\langle d \rangle \approx 4, 6, 10$ in Figs. 4(a)–4(c), respectively] that $C_V(D)$ decreases overall but the relative shape of the space of graphs within $G(D)$, as defined by the range $[s_{\min}, s_{\max}]$, remains qualitatively consistent with that of trees. However, the total variation as measured by the distance between $(s_{\max} - s_{\min})/s_{\max}$ decreases with increasing link density. At the same time, for graphs with increasing link density and having degree sequence with $C_V^{\max}(D)$, the difference $s_{\max} - s_{\min}$ is no longer zero in general, indicating inherent diversity even at higher levels of variation.² Graph assortativity as measured by the range $[r_{\min}, r_{\max}]$ is also qualitatively the same as for trees, in that high $C_V(D)$ is enough to dictate that $r < 0$ but considerable diversity exists for low values of $C_V(D)$. Although such results are not conclusive, we view them as generally supportive of graph diversity as we have discussed it here.

Finally, while this paper has focused on degree sequences and has used the s metric to highlight the differences in graphs sharing the same D , we conjecture that a similar story is apt to apply to other graph metrics (even higher-order ones

²However, when the degree sequence D corresponds to a multistar (e.g., double star, triple star), the overall picture in the upper row of Fig. 4 looks the same, except that the s_{\min}/s_{\max} values jump abruptly to 1 at $C_V^{\max}(D)$, since all multistars are isomorphic to one another in $G(D)$.

like the JDD). A detailed exploration of these issues for other metrics will be important in the development of new graph analysis and generation techniques.

ACKNOWLEDGMENTS

The authors thank Daniel Whitney for the use of his implementation of a rewiring algorithm to obtain s_{\min} values. The authors gratefully acknowledge John Doyle, Walter Willinger, and Daniel Whitney for many stimulating and insightful discussions. They also thank Aaron Clauset and two anonymous referees for comments that helped to improve the presentation of this work. Both authors were supported at Caltech by Boeing, AFOSR Grant No. URI 49620-01-1-0365 “Architectures for Secure and Robust Distributed Infrastructures,” the Army Institute for Collaborative Biotechnologies, AFOSR Grant No. FA9550-05-1-0032 “Bio Inspired Networks,” and Caltech’s Lee Center for Advanced Networking. D.A.’s work at NPS was supported by Grant No. NIFR-RIP-BORYB.

APPENDIX

In order to see when a degree sequence D can achieve $r(g)=-1$, we introduce a simplified version of the Cauchy-Schwarz-Burnyakovskii inequality, which states that for any vector $\{b_1, b_2, \dots, b_n\}$, it must be that

$$\sum_{i=1}^n b_i^2 \geq \frac{1}{n} \left(\sum_{i=1}^n b_i \right)^2,$$

with the equality holding if and only if $b_1=b_2=\dots=b_n$.

Applying this inequality to a graph with l links, it follows that

$$\sum_{(i,j) \in \mathcal{E}} (d_i + d_j)^2 \geq \frac{1}{l} \left(\sum_{(i,j) \in \mathcal{E}} (d_i + d_j) \right)^2.$$

Expanding the squared term on the left-hand side and dividing both sides by 2, we have from relations (8) and (9) that

$$\sum_{(i,j) \in \mathcal{E}} 2d_i d_j / 2 + \sum_{(i,j) \in \mathcal{E}} (d_i^2 + d_j^2) / 2 \geq \frac{1}{2l} \left(\sum_{(i,j) \in \mathcal{E}} (d_i + d_j) \right)^2,$$

$$s(g) + s_{\max}^{\mathcal{G}(D)} \geq 2s(g_c), \quad \frac{s(g) - s(g_c)}{s_{\max}^{\mathcal{G}(D)} - s(g_c)} \geq -1,$$

which is simply another way of showing that $r(g) \geq -1$, but it proves that $r(g) = -1$ if and only if $d_i + d_j = d$ (a constant) for all $(i,j) \in \mathcal{E}$.

Recall that within $\mathcal{G}(D)$ one has $s_{\min} = Z + \hat{Z}$ as defined by (4), and thus this s_{\min} graph corresponds to $r = -1$ if and only if for each element k one has $z_k + \hat{z}_k = z$ (a constant).

-
- [1] S. N. Dorogovtsev and J. F. F. Mendes, *Adv. Phys.* **51**, 1079 (2002).
 - [2] M. E. J. Newman, in *Handbook of Graphs and Networks: From the Genome to the Internet*, edited by S. Bornholdt and H. G. Schuster (Wiley-VCH, Berlin, 2003).
 - [3] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
 - [4] M. E. J. Newman, *SIAM Rev.* **45**, 167 (2003).
 - [5] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, UK, 2003).
 - [6] *Handbook of Graphs and Networks: From the Genome to the Internet* (Ref. [2]).
 - [7] J. Park and M. E. J. Newman, *Phys. Rev. E* **70**, 066117 (2004).
 - [8] J. C. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 14497 (2005).
 - [9] R. Tanaka, *Phys. Rev. Lett.* **94**, 168101 (2005).
 - [10] L. Li, D. Alderson, J. Doyle, and W. Willinger, in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (Association of Computing Machinery, New York, 2004), pp. 3–14.
 - [11] D. Alderson, L. Li, W. Willinger, and J. C. Doyle, *IEEE/ACM Trans. Netw.* **13**, 1205 (2005).
 - [12] P. Erdős and T. Gallai, *Mat. Lapok* **11**, 264 (1960) (in Hungarian).
 - [13] A. Tripathi and S. Vijay, *Discrete Math.* **265**, 417 (2003).
 - [14] S. S. Skiena, *The Algorithm Design Manual* (Springer-Verlag, New York, 1997).
 - [15] F. Chung and L. Lu, *Internet Math.* **1**, 91 (2003).
 - [16] E. A. Bender and E. R. Canfield, *J. Comb. Theory, Ser. A* **24**, 296 (1978).
 - [17] M. Molloy and B. Reed, *Random Struct. Algorithms* **6**, 161 (1995).
 - [18] L. Li, D. Alderson, J. Doyle, and W. Willinger, *Internet Math.* **2**, 431 (2006).
 - [19] http://en.wikipedia.org/wiki/Rearrangement_inequality
 - [20] D. E. Whitney and D. Alderson, in *Proceedings of the International Conference on Complex Systems* (New England Complex Systems Institute, Cambridge, MA, 2006).
 - [21] S. Maslov and K. Sneppen, *Science* **296**, 910 (2002).
 - [22] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
 - [23] B. B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk* (Springer-Verlag, New York, 1997).
 - [24] A.-L. Barabási, R. Albert, and H. Jeong, *Physica A* **272**, 173 (1999).
 - [25] P. L. Krapivsky and S. Redner, *Phys. Rev. E* **63**, 066123 (2001).
 - [26] S. Maslov, K. Sneppen, and A. Zalanyzk, *Physica A* **333**, 529 (2004).
 - [27] M. E. J. Newman, *Phys. Rev. E* **67**, 026126 (2003).
 - [28] S. N. Dorogovtsev, *Phys. Rev. E* **69**, 027104 (2004).
 - [29] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, *Eur. Phys. J. B* **38**, 204 (2004).
 - [30] M. A. Serrano and M. Boguñá, e-print cond-mat/0603353.
 - [31] I. J. Farkas, I. Derenyi, A. L. Barabási, and T. Vicsek, *Phys. Rev. E* **64**, 026704 (2001).

- [32] M. A. Serrano and M. Boguñá, Phys. Rev. E **72**, 036133 (2005).
- [33] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
- [34] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, in *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (Association of Computing Machinery, New York, 2006).
- [35] E. W. Weisstein, <http://mathworld.wolfram.com/CorrelationCoefficient.html>
- [36] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
- [37] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
- [38] Z. Burda and A. Krzywicki, Phys. Rev. E **67**, 046118 (2003).
- [39] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
- [40] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, Nat. Phys. **2**, 110 (2006).
- [41] C. Gkantsidis, M. Mihail, and E. Zegura, in *Proceedings of the 5th Workshop on Algorithm Engineering and Experiments*, edited by Richard E. Ladner (SIAM, Philadelphia, 2003).
- [42] R. Xulvi-Brunet and I. M. Sokolov, Phys. Rev. E **70**, 066102 (2004).
- [43] S. N. Dorogovtsev, J. F. F. Mendes, A. M. Povolotsky, and A. N. Samukhin, Phys. Rev. Lett. **95**, 195701 (2005).