Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles

Gergana Bounova* and Olivier de Weck*

Engineering Systems Division Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

(Received 31 August 2010; revised manuscript received 26 October 2011; published 30 January 2012)

This study is an overview of network topology metrics and a computational approach to analyzing graph topology via multiple-metric analysis on graph ensembles. The paper cautions against studying single metrics or combining disparate graph ensembles from different domains to extract global patterns. This is because there often exists considerable diversity among graphs that share any given topology metric, patterns vary depending on the underlying graph construction model, and many real data sets are not actual statistical ensembles. As real data examples, we present five airline ensembles, comprising temporal snapshots of networks of similar topology. Wikipedia language networks are shown as an example of a nontemporal ensemble. General patterns in metric correlations, as well as exceptions, are discussed by representing the data sets via hierarchically clustered correlation heat maps. Most topology metrics are not independent and their correlation patterns vary across ensembles. In general, density-related metrics and graph distance-based metrics cluster and the two groups are orthogonal to each other. Metrics based on degree-degree correlations have the highest variance across ensembles and cluster the different data sets on par with principal component analysis. Namely, the degree correlation, the s metric, their elasticities, and the rich club moments appear to be most useful in distinguishing topologies.

DOI: 10.1103/PhysRevE.85.016117 PACS number(s): 89.75.Fb

I. INTRODUCTION

This paper presents an overview of network topology metrics and a computational approach to analyzing graph topology via multiple-metric analysis on graph ensembles. We review two studies that use this approach and build upon them by discussing a wider set of metrics and their correlations in more depth. Our goal is to present examples of why it is hard to generalize about many of the metrics studied in the literature. Namely, studying a single measure or pulling networks from different domains and topologies together for statistical analysis might provide incorrect conclusions. This is true for (at least) three reasons: (i) There often exists considerable diversity among graphs that share any given topology metric [1,2], (ii) patterns vary depending on the underlying graph construction model, and (iii) many real data sets might not represent actual statistical ensembles.

The metrics overview includes a general introduction and discussion of correlation patterns for random graph ensembles and real data examples. It is found that, in many cases, distance-based metrics correlate negatively with density or degree distribution moments. There are exceptions, in both classical ensembles and real data. We discuss the exceptions, the high variance metrics, and differences between ensembles. We show that some of our real data sets do not behave as statistical ensembles. This is an additional argument that even combining data from the same domain can be inconclusive.

Finally, among all topology descriptors, we find that degree-degree correlations show the most variation across different types of data and can be used for classification. We claim that a multiple-metric graph ensemble approach is essential for the basic exploration of any network topology problem. This is intended as a reminder and reference to the ambiguity that the networks field faces as a multidomain field with inherently high-dimensional problems. An extensive review of this topic can be found in [1].

A. Challenges in analyzing network topology

The definition of network topology used here is as follows: *the configuration by which the elements of a network are connected*. There are $2^{n\times n}$ ways to connect n nodes, from a set of nodes with no edges to a complete graph. Transition from one topology to another is a discrete process accomplished by series of edge removals and additions. While much research has been done on random rewiring, with preserving or targeting certain properties, there is limited work in defining distances between topologies other than trees. The properties of metrics defined on this space are not well understood either. A statistical approach, by studying metric correlations, is a way to approach this challenge computationally.

In the literature, there are generally two approaches to analyzing topology: a construction approach and a detection approach. The construction papers [3,4] propose algorithms for building graphs to achieve a certain topology, or match a set of metrics, or to replicate the topology of a real system. The detection approach [5,6] concentrates on comparing statistical metrics across topologies to extract similarities or claim resemblance. In general, the second approach suffers from nonuniqueness because many topologies can map to a set of metrics, while the construction approach is better at reproducing statistical properties but challenged by computation. For example, matching a graph topology by matching subgraph distributions [4] is very effective but increasingly hard combinatorially with increasing subgraph size. Finding graph isomorphisms is a subproblem of motif search which does not have a known polynomial time solution.

This paper aims to enrich the detection approach by analyzing correlations in a large set of metrics. These metrics

*gergana@mit.edu, deweck@mit.edu

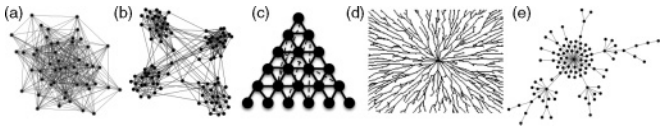

FIG. 1. Examples of network topologies discussed in the literature: (a) Erdös-Rényi graph [7], (b) general random modular graph [10], (c) a randomized hierarchy [11], (d) spatial distribution tree graph [12], (e) preferential attachment graph [3].

are not independent and their relationships vary depending on the graph topology.

B. Topology notions in the literature

While there is no general classification of network topologies, there are a few popular notions. The ones presented here are used for generating random graph ensembles for our metrics study. Examples of their graphical representation are shown in Fig. 1.

Random graph studies mark the beginnings of network theory [7]. Random topologies are often used for benchmarking against real data. These are topologies in which edges are added randomly, or with additional criteria, such as having the same degree distribution or preferentially.

Modular networks are of interest in product and process design, and social networks. Modularity is studied in the context of robustness, complexity and operability, and community dynamics [8]. If a system performs better by some parameter when designed modularly rather than integrally, then the modularity aspect can be a design consideration [9]. We analyze a random modular graph ensemble model by Clauset [10].

Hierarchy in networks is often studied in the context of organizations and communication in teams. Hierarchy is used in both construction approaches [11] as a model for building networks and in detection approaches, where the goal could be to find evidence for hierarchical patterns. Our statistical ensemble comes from the randomized hierarchies parametrized model by Dodds *et al.* [11].

Trees and lattice structures are the backbone of hierarchies. Trees are important in engineering applications where minimal linking is desired. This is true for sparse grids, transportation networks, distribution systems, and communication networks. The example used here is of spatial distribution trees [12] where a tree is linked to a source with varying degree of centralization versus optimal local linking.

Scale-free topologies are originally termed networks with power-law degree distributions ($P[X > x] \approx cx^{-\alpha}$, as $x \to \infty$), but the name generates some controversy, as power-law degree distributions can correspond to various types of network structures [13]. We use the basic preferential attachment model [3] to study an ensemble of preferential attachment trees, as well as a general "scale-free" random graph model by Catanzaro *et al.* [14].

This is not an exhaustive list of topology notions or models in the literature. There are variations on the above, for example, nested hierarchies or preferential attachment models with different fitness functions. For engineering applications, there are models for network growth optimized for specific performance and subject to specific constraints. Relevant terms are *highly optimized tolerance* (HOT) and *robust but fragile* that come from claims [15] that complex systems evolve to perform certain functions robustly, and efficiently, but that in itself makes them susceptible in very specific ways.

In addition to the synthetic data sets described above, we also analyze ensembles of airline data sets and language Wikipedias to enrich the metric analysis with real data. These are assumed to be proper ensembles as their individual networks grow under similar economic conditions and technological constraints.

C. Paper outline

With these notions of topology in mind, we review two papers that present multiple-metric analysis with linear correlation heat maps (Sec. II). Then, we present a wider spectrum of metrics (Sec. III) and compute them for the above classical topology ensembles as well as for five airline data sets and one Wikipedia ensemble (Sec. IV). For each data set, we discuss correlation patterns and how they differentiate its particular topology. Further discussion of metric statistics and selection of high-variance metrics is presented in Sec. V. Conclusion and further work remarks follow in Sec. VI.

II. LITERATURE REVIEW

There are many single-metric or distributions-based network topology studies. Extensive reviews of the field are available in [16–18]. We review two relevant papers that adopt a comprehensive metric approach.

Roy *et al.* [19] use a similar multiple-network-metrics statistical approach to understand the structure of a few biological data sets. They study 11 metrics across 32 data sets. The metrics include number of nodes and edges and the first three moments of the degree distribution, the betweenness distribution, and the geodesic (shortest path) distribution. They also create a heat map based on paired metrics correlations. The authors conclude that the correlations are not very strong overall, which allows them to decouple the metrics in their model fitting. The 11 network metrics are then examined with respect to the biological phenotypes using hierarchical linear regression, and the most relevant network metrics are extracted (as many as nine out of 11 for some phenotypes).

Filkov *et al.* [20] use a heat map and multiple-network-metric correlations to compare networks of various topologies. The primary goal of the paper is to test the end result of a network growth algorithm in comparison to other well-known topologies. The authors correlate 15 metrics, derived from seven original metrics, including second and third distribution moments. These are correlated across 113 real data sets which represent systems from social, technical, and biological domains. The authors find that the 15 metrics are not coupled strongly and select 11 that are least correlated for further analysis. They use principal component analysis to project both real and synthetic data onto the first three principal components based on these 11 metrics.

Both papers use the statistical approach of correlating metrics across ensembles for the purposes of validation or support of another model, but do not perform a full exploration

of patterns in these correlations. Our intent is to study a wider set of metrics and correlate them across larger ensembles. Our claim is that the correlation heat map is different depending on the type of ensemble studied. Random graphs should differ from preferential attachment graphs, from trees or hierarchies. Real networks should exhibit a set of unique patterns as well, depending on their structure and function. Therefore, using the correlations or lack thereof, across diverse data sets might not be valid. In addition, we view topology as independent of network size, and consider only normalized metrics.

III. THE TOPOLOGY METRICS

A. Metric selection

We analyze 21 metrics from the literature, but compute statistics for a total of 30 metrics including second moments. Most of these come from social science [21], while some are from the networks literature such as degree correlations. They are organized in six groups, which represent different properties of the graph: degree-related, degree correlations, geodesics-related (shortest paths), modularity, motif-related, and spectral properties. Table I shows all metrics with their description, nomenclature, and their normalization.

The list in Table I is not comprehensive. Other examples are the *in-* and *out-degree* distribution moments, the *radius* of a graph: the minimum eccentricity across all vertices (the eccentricity is the maximum distance to any other vertex), as well as third moments of all distributions. Motifs (stars and cliques) with up to six nodes showed correlations almost identical to the four-node motifs, so these were excluded. Degree distribution exponents, if they exist, are not in this list. In addition, for the real data sets, we did not consider application-specific metrics. Depending on the objective, these can contain more information about a network than any of the pure graph theoretic measures. For example, for airline data, traffic-related measures such as load factor are essential, but are not part of this analysis.

B. Normalization

One of the key assumptions of this paper is that topology is independent of size. A diameter of 5 can have different implications for network of size 10 or 10 000. This is why all metrics are normalized by some function of graph size. For example, when normalized by the maximum possible degree $(n-1)$, average nodal degree $\langle k\rangle$ is the same quantity as link density Eq. (1). In Eq. (1) n is the number of nodes, and m is the number of edges. This formulation is valid for undirected graphs only.

$$\frac{\langle k\rangle}{n-1}=\frac{2m/n}{n-1}=\frac{2m}{n(n-1)}. \tag{1}$$

Most natural and engineered networks are sparse, hence the average degree can probably be normalized by a factor much smaller than $(n-1)$. Since there is no general theoretical threshold for sparseness, and higher densities are observed for some systems (up to 20% even for airlines), we keep to this classic definition.

Normalization for all metrics is explained in Table I. Variances or second moments of metric distributions are the variances of the normalized metrics.

The need to normalize arises from the need to analyze an ensemble of graphs with varying sizes. Suppose that the graphs in the ensemble have number of nodes $x_1, x_2, \ldots, x_N$, where N is the size of the ensemble. Then the set of measurements for the i^{th} and j^{th} metric, $(\mu_1^i, \ldots \mu_N^i; \mu_1^j, \ldots \mu_N^j)$ are normalized by $(\frac{\mu_1^i}{f_{\mu^i}(x_1)}, \frac{\mu_2^i}{f_{\mu^i}(x_2)}, \ldots, \frac{\mu_N^i}{f_{\mu^i}(x_N)})$ and $(\frac{\mu_1^j}{f_{\mu^j}(x_1)}, \frac{\mu_2^j}{f_{\mu^j}(x_2)}, \ldots, \frac{\mu_N^j}{f_{\mu^j}(x_N)})$, where f_{μ^i} and f_{μ^j} are normalization functions that depend on μ^i and μ^j. Suppose all the x's are equal, that is, the graphs in the ensemble all have the same size. Then, the correlation between μ^i and μ^j is independent of the normalization functions: $\frac{1}{f_{\mu^i}(x)}(\mu_1^i, \ldots, \mu_N^i)$ and $\frac{1}{f_{\mu^j}(x)}(\mu_1^j, \ldots, \mu_N^j)$. If the x's are different, but with a small variance, the relationships between metrics will be affected slightly. On the other hand, if the x's vary significantly, not normalizing will distort the correlations. For the synthetic data we present, all ensembles have graphs of the same size, so there are no size effects. In real data, however, graphs with similar topology will not always have the same size. These correlations are discussed in Sec. IV B.

C. Constructing correlation heat maps

Hierarchically clustered heat maps are used to represent the pairwise correlations between metrics for every ensemble. One ensemble, in general, is represented by a (ensemble size) $\times$ (number of descriptors) matrix, X, which is usually 100×30. The columns are then converted to standard units and correlated. The correlation matrix, which is essentially the covariance matrix, is $P = \text{cov}(X^T, X)$ is 30×30, $P_{i,j} = \text{cov}(X_i, X_j)$, where X_i and X_j are columns of X. This matrix is used to plot a heat map in which the rows and columns are sorted via hierarchical clustering. We use a generic hierarchical clustering algorithm with a (negative) correlation distance, as we expect that two metrics will behave the same way if their correlation profiles are similar.

An example is shown in Fig. 2. This is an ensemble of 100 Erdös-Rényi graphs, each with 1000 nodes and density of 0.0069 ($0.0069 \approx \log(1000)/1000$, for an ER graph a density of $\log(n)/n$ is necessary to be almost surely connected). The heat map reveals two clusters of metrics, densities, and distances which are negatively correlated with each other. Degree correlations are more decoupled but they also have clustering patterns.

This heat-map representation is used to compare ensembles of other graph models in Sec. IV.

D. Linearity

Most of the 30 metrics are extensively discussed in the networks literature. The relationships between some of them are either derived explicitly, bounded, or derived empirically. For example, it is known that the average path length and the diameter are related, as well as the diameter and the radius $(r_G \leqslant d \leqslant 2r_G)$. There is a direct relationship between degree correlation and s metric [Eq. (2), [2]], as well as between average node betweenness and average path length, derived

TABLE I. Metric classes: degree-related, degree correlations, geodesics, modularity-related, motif counts, and spectral properties; n is the number of nodes; m is the number of edges. All of these refer to undirected graphs only.

	Metric	Description	Normalized by
1	$\langle k \rangle$, density	$2m/[n(n-1)]$	Maximum possible degree $(n-1)$
2	var(k), degree variance [2]	Variance of the normalized degree sequence	–
3	$\langle k_n \rangle$, average neighbor degree	$\frac{1}{n}\sum_i$ {average of neighbor degrees of node i}	Maximum possible degree $(n-1)$
4	var(k_n), average neighbor degree variance	Variance of the average neighbor degree sequence	–
5	$\langle C \rangle$, average clustering coefficient [22]	Average clustering coefficient (per node)	–
6	var(C), clustering coefficient variance	Variance of the clustering coefficient (per node)	–
7	r, degree correlation [23]	Pearson correlation coefficient of degrees between pairs of linked nodes	–
8	$r_e = \lvert r_{\max}\text{-}r_{\min} \rvert$, degree correlation elasticity [2]	The difference between the maximum and minimum degree correlation obtained by rewiring (with preserving degrees)	–
9	$\langle r_c \rangle$, rich club metric (average) [24]	The average of the rich club metric with respect to threshold degrees from 1 to $n-1$	Normalized by the corresponding random graph rich club metric
10	var(r_c), rich club metric variance	Variance of the rich club distribution (of rich club metrics with threshold degrees from 1 to $n-1$)	–
11	$s_{\max}$ ($s/s_{\max}$), [13]	The ratio between the s metric of the given graph to the maximum possible s metric with the given degree distribution (the s metric is the sum of the product of nodal degrees across edges, $\sum_{i,j\in E} k_i k_j$)	Normalized by $s_{\max}$
12	$\lvert s_{\max}\text{-}s_{\min} \rvert$, [2]	The difference between the maximum possible s metric and the minimum possible s metric, under degree-preserving rewiring	Normalized by $s_{\max} \Rightarrow \lvert 1\text{-}s_{\min}/s_{\max} \rvert$
13	μ,number of modules [8]	The number of modules according to the Newman eigenvector algorithm	Average number of nodes per module/total number of nodes $= 1/\mu$
14	$\langle w_n \rangle$, node betweenness [21]	Average node betweenness	Divided by the total possible number of paths (between i and j through k)
15	var(w_n), node betweenness variance	Variance of the nodal betweenness (across nodes)	–
16	$\langle w_e \rangle$, edge betweenness [25]	Average edge betweenness	Divided by the total possible number of paths (between i and j through edge e)
17	var(w_e), edge betweenness variance	Variance of the edge betweenness (across edges)	–
18	$\langle c_l \rangle$, average closeness	The closeness of a node is the sum of reciprocal distances to all other nodes	–
19	var(c_l), closeness variance	The variance of the closeness (across all nodes)	–
20	$\langle l \rangle$, average path length	The average shortest path across all pairs of nodes	Divided by the longest possible path $(n-1)$
21	d, diameter	The maximum shortest path across all pairs of nodes	Divided by the longest possible path $(n-1)$
22	$\langle d_d \rangle$, distance distribution mean	The mean of the frequency distribution of shortest paths	–
23	var(d_d), distance distribution variance	The variance of the frequency distribution of shortest paths	–
24	l_3, loops 3	Number of loops of size 3 (triangles)	Divided by the total number of triples, $\binom{n}{3}$
25	c_4, 4-cliques	Number of cliques of size 4 (complete subgraphs with 4 nodes)	Divided by the total number of 4-tuples, $\binom{n}{4}$
26	s_4, 4-stars	The number of all star motifs with four nodes (one hub and three spokes)	Divided by the total number of 4-tuples
27	G, graph energy [26]	The sum of the absolute values of the eigenvalues of the adjacency matrix	Normalized by $n^{1.5}$ [27]
28	$\langle e_C \rangle$, average eigencentrality [21]	The average of the maximum-eigenvalue eigenvector (of the adjacency)	–
29	var(e_C), eigencentrality variance	The variance of the maximum-eigenvalue eigenvector	–
30	a, algebraic connectivity [28]	The second smallest eigenvalue of the Laplacian of the adjacency	–

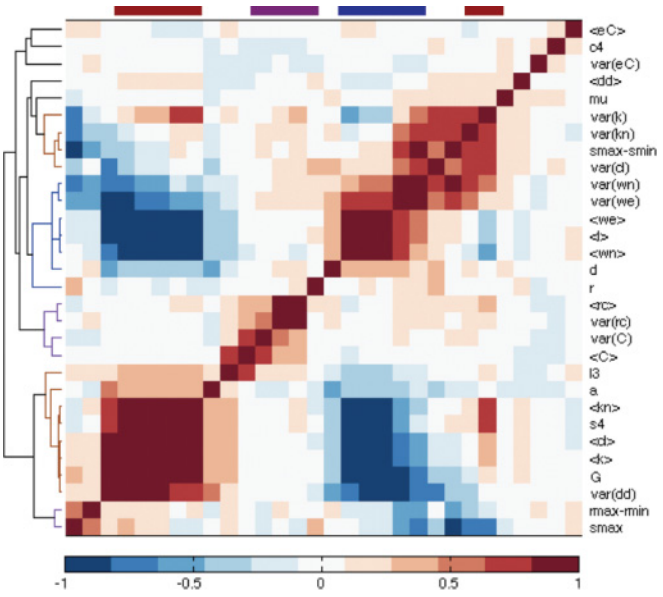

FIG. 2. (Color) Correlations heat map. Every metric is correlated (via the Pearson correlation) with all others, using measurements from an ensemble of graphs (in this case 100 Erdös-Rényi graphs with 1000 nodes and link density of 0.0069). The rows and columns are hierarchically clustered. Clusters of densities, distances, and degree correlations are annotated with red, blue, and purple, respectively. For metrics nomenclature see Table I.

from the definition of node betweenness [Eq. (3); see Table I].

$$r(g) = \frac{s(g) - s(g_c)}{s_{\max}^{G(D)} - s(g_c)}, \tag{2}$$

$$\langle w_n \rangle = (n-1)(\langle l \rangle - 1). \tag{3}$$

A direct result from Eq. (3) is that the average nodal betweenness is always positively correlated with the average path length. In the normalized sense,

$$\langle w_n \rangle' = \frac{\langle w_n \rangle}{n(n-1)}, \quad \langle l \rangle' = \frac{\langle l \rangle}{n-1}.$$

Substituting these in Eq. (3) gives

$$\begin{aligned} \langle w_n \rangle' n(n-1) &= (n-1)((n-1)\langle l \rangle' - 1) \\ &\Rightarrow \langle w_n \rangle' = (1 - 1/n)\langle l \rangle' - 1/n. \end{aligned}$$

So for large n, the normalized average nodal betweenness is linearly proportional to the normalized average path length. It is known that algebraic connectivity is bounded from below ($a \geqslant \frac{4}{nd}$), and that the value in practice stays close to the lower bound. Motif counts are proportional with density. The higher the number of edges, the higher is the probability of forming cliques, loops, and other motifs. Most of these facts are confirmed by our statistical results. In Sec. IV we discuss only the patterns at large. The more interesting pairwise relationships are discussed in Sec. V.

Studying linear correlations of largely nonlinear relationships is a major assumption. Some of the relationships are known to be linear Eq. (3), while others are known to be nonlinear. For example, graph energy is quadratic with link density, but linear for small densities [27]. Not all correlations are explained as intuitively as the relationship between density and motif counts. Some zero correlations hide nonlinear relationships. While this is true in general, for our data sets the pairwise scatter plots of all metrics show that strong correlations correspond to linear relationships, and lack of correlation corresponds to random scatter. For the sake of conciseness, these plots are not included.

E. Graph size, ensemble size

All graphs in this analysis are generated based on different stochastic models, with fixed rules over repeated trials. The fidelity of the models gets better with increasing graph size, while the metric distribution moments converge with ensemble size. If we assume that the first and second moments exist, which for finite distributions (finite graph size) is true, then by the law of large numbers the means will converge with increasing ensemble size. Similarly, the variances converge. The sample correlation coefficient can be expressed in terms of the means and variances, and as such, it also converges with increasing ensemble size.

As fidelity increases with graph size, so does computational hardness. The most computationally intensive metrics in Table I are the degree correlation elasticity and building a corresponding graph with a maximum s metric, based on the same degree distribution [13]. These have theoretical values that are easy to compute but less precise in practice.

The ensembles in this study are comprised of 1000-node graphs. There is no inherent limitation in producing correlation heat maps for larger networks. If multimetric ensemble-based analysis is relevant, for very large graphs, sample averages, approximations, or theoretical values would have to be used. This is especially true for motifs, if the motif size is large as well.

IV. CORRELATION HEAT MAPS

This section discusses the heat maps of the covariance matrices for the set of metrics in Table I, computed for both synthetic ensembles and real data. The synthetic graph models are inspired by those from Sec. I B. Real data comes from various airline networks data sets and a set of graphs based on different language Wikipedias. The aim is to emphasize that while there are some general patterns, every model is different. Real data-set heat maps do not resemble the classical models at all, and the metrics are never independent, as seen in previous work [20]. This is expected, because if real data sets approximate classical models at all, they will be either noisier instantiations of pure topologies or hybrids of canonical graph types.

A. Ensembles of random graphs

Figure 3 shows clustered heat maps for six families of random graphs: the classical Erdös-Rényi model, a modular random graph model [10], a general random graph model with a scale-free degree distribution [14], pure hierarchies with random cross links [11], spatial distribution trees [12], and preferential attachment trees.

Erdös-Rényi [Fig. 3(a)] is the most studied ensemble due to its simplicity and asymptotic properties. The ER heat map has two clusters of high correlation: (i) between nodal degree distribution moments, such as density, degree variances, motif

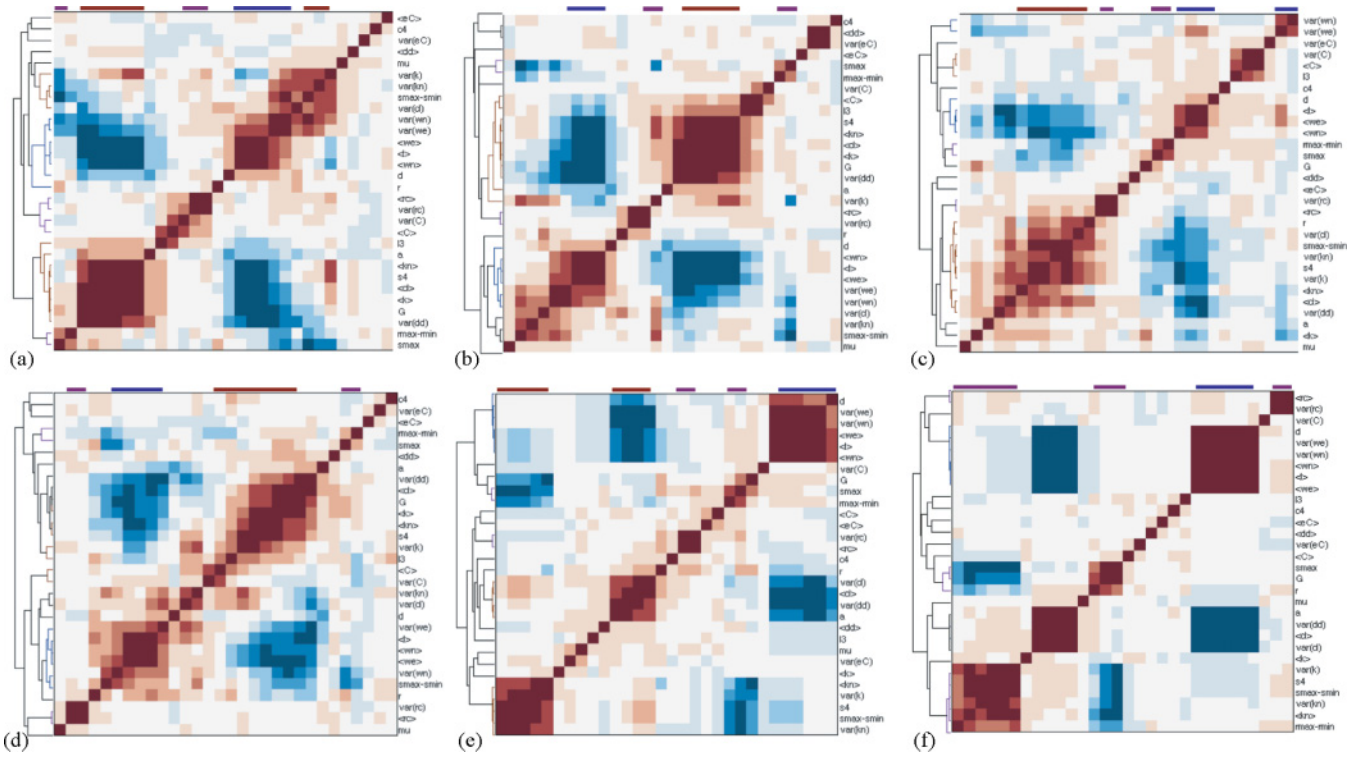

FIG. 3. (Color) Clustered heat maps of random graph ensembles; (a) the classic ER random graph; 100 graphs, 1000 nodes, $p = 0.0069$, also in Fig. 2; (b) random modular ensemble; 100 graphs, 1000 nodes, four clusters, overall density 0.0069; (c) randomized hierarchies; 100 graphs, 1000 nodes, density 0.0025; (d) general preferential attachment model; 100 graphs, 1002 nodes, $\gamma = 2.5$, density 0.0096; (e) preferential attachment trees; 100 graphs, 1000 nodes; (f) spatial distribution trees; 100 graphs, 1000 nodes, $\alpha = 0.5$ [12]. For metrics nomenclature see Table I. Clusters are annotated with red for densities, blue for distances, and purple for degree-degree correlations.

counts, and graph energy; and (ii) distance-related metrics, such as diameter and betweenness. The two clusters are strongly negatively correlated with each other, which is a general pattern for most data sets in this study. Intuitively, at higher density, distances are shorter, and vice versa. Densification and shrinking distances are discussed in [29]. Motifs are naturally abundant with higher densities. Graph energy is known to be linear for small densities (quadratic across the whole density spectrum). Among the uncorrelated band of metrics are the following: the eigenvector centrality metrics ($\langle e_C \rangle$,var(e_C)), modularity (μ), and the degree correlation (r). This is not surprising as ER graphs are supposed to have nodes of equal importance, not be modular, and have a zero degree correlation.

The *random modular ensemble* [Fig. 3(b)] is designed to have more links within modules than between modules with some probability [10]. Comparison with the ER heat map yields barely distinguishable differences. The degree correlation, rather than being zero everywhere and unrecorrelated, clusters with the rich club metrics. Given that there are modules with different node-neighbor degree profiles, this makes sense. Furthermore, the distance distribution mean correlates positively with the eigenvector centrality variance.

The *randomized hierarchies* graphs [Fig. 3(c)] have a lattice structure with shortcuts [11]. This heat map is quite distinct from the ER heat map. First, densities and distances are much more decoupled and not negatively correlated. Since these graphs have an underlying lattice structure, the distances in the graph are largely independent of density. Shortcuts increase density and decrease distances, but in this model this does not occur on average everywhere in the graph [11]. Second, correlation metrics are coupled, and some form a large cluster, namely the degree correlation (r), the s metric elasticity ($|s_{\max}\text{-}s_{\min}|$), the variance of the average degree (var(k)), and the average neighbor degree ($\langle k_n \rangle$). Because in this hierarchy nodes also have peer links to similar nodes, the rich club coefficient associates with the same cluster. The distance metrics correlate negatively with the degree correlations. The farther away two nodes are, the less likely it is they will have similar degrees and similar neighbors.

The *random preferential attachment ensemble* [Fig. 3(d)] is constructed by building random graphs from a power-law degree distribution [14]. The degree distribution exponent is chosen to be $\gamma = 2.5$ for the entire ensemble. There are two main distinctions that come with a skewed degree distribution. First, in random graphs, on average, the nodal degree variance correlates negatively with maximum s metric. That means that high variation in degrees results in more chances of high-to-low nodal degree edges, which does not maximize the s metric. For this model, there is no correlation between var(k) and $s_{\max}$. Due to the skewed degree distribution, the network is quite inelastic (hard to rewire), so while there is a large degree variance, the graph is close to its corresponding $s_{\max}$ graph in practice. The second distinction is that degree correlations and distance metrics are not as decoupled, and degree correlation itself (r) varies proportionally with distances. The shorter

the distances, the smaller the degree correlation, indicating higher degree variance, and more high-degree to low-degree connections.

The two tree ensembles have pronounced metric clusters, and wide areas of zero correlation. Because of the tree topology, metrics such as density, clustering coefficient and its variance, and number of loops and cliques become invariant.

The *preferential attachment tree* [Fig. 3(e)] has the regular density-distance groups and very decoupled degree-degree correlations. An exception is the s metric elasticity ($|s_{\max}\text{-}s_{\min}|$) which correlates positively with density-related metrics.

The *spatial distribution tree* heat map [Figs. 1(d) and 3(f)] does not have the regular density-distance clustering. Densities are decoupled, while degree correlations cluster strongly. Furthermore, the two elasticities, $|r_{\max}\text{-}r_{\min}|$ and $|s_{\max}\text{-}s_{\min}|$ correlate positively with each other, which does not occur for any other ensemble. This means that these graphs are both not very "scale-free" in the s metric sense, and also very elastic in terms of degree correlation. This is the opposite of the preferential attachment trees, which are inelastic (hard to rewire) and scale-free.

In summary, the random graph heat maps confirm that there are prevalent metric relationships, but they do not hold in all cases. Furthermore, even between similar topologies, varying model parameters can result in completely different correlation heat maps.

B. Airline networks and Wikipedia

For synthetic data, an ensemble of graphs is a set of graphs generated with the same rules, in independent trials. There is no real data that comes in this form. Systems that can be modeled as networks are seen as "... naturally occurring networks...intended to serve a single, coordinated purpose,...but which are built over long periods of time by many independent agents and authorities" [30]. Therefore, to talk about topology patterns, we have to look for instances of graphs that have grown to serve the same purpose under similar economic conditions or technological constraints.

Our data contain an example of temporal ensembles: monthly airline networks [31]. Monthly instances are the same network with variations, such as seasonal patterns or growth. Even though these graphs are not independent, for this example they will approximate a statistical ensemble.

The first ensemble is the set of all US origin-destination pairs for 212 months from 1990 to 2007. An origin-destination pair is two airports connected by a flight. So this ensemble combines all airline flights in the U.S. monthly in 212 graphs [Fig. 4(a)]. Second, from the same data, we extract the legacy carriers only [Fig. 4(b)]. Six airlines, operating in 2007, are considered to have a "legacy" topology: American Airlines, Continental, Delta, Northwest, United, and US Airways. The third ensemble is of low-cost carriers, which are thought to have different operations and a distinct topology [Fig. 4(c)]. The low-cost carriers are Airtran, ATA, Frontier, Jetblue,

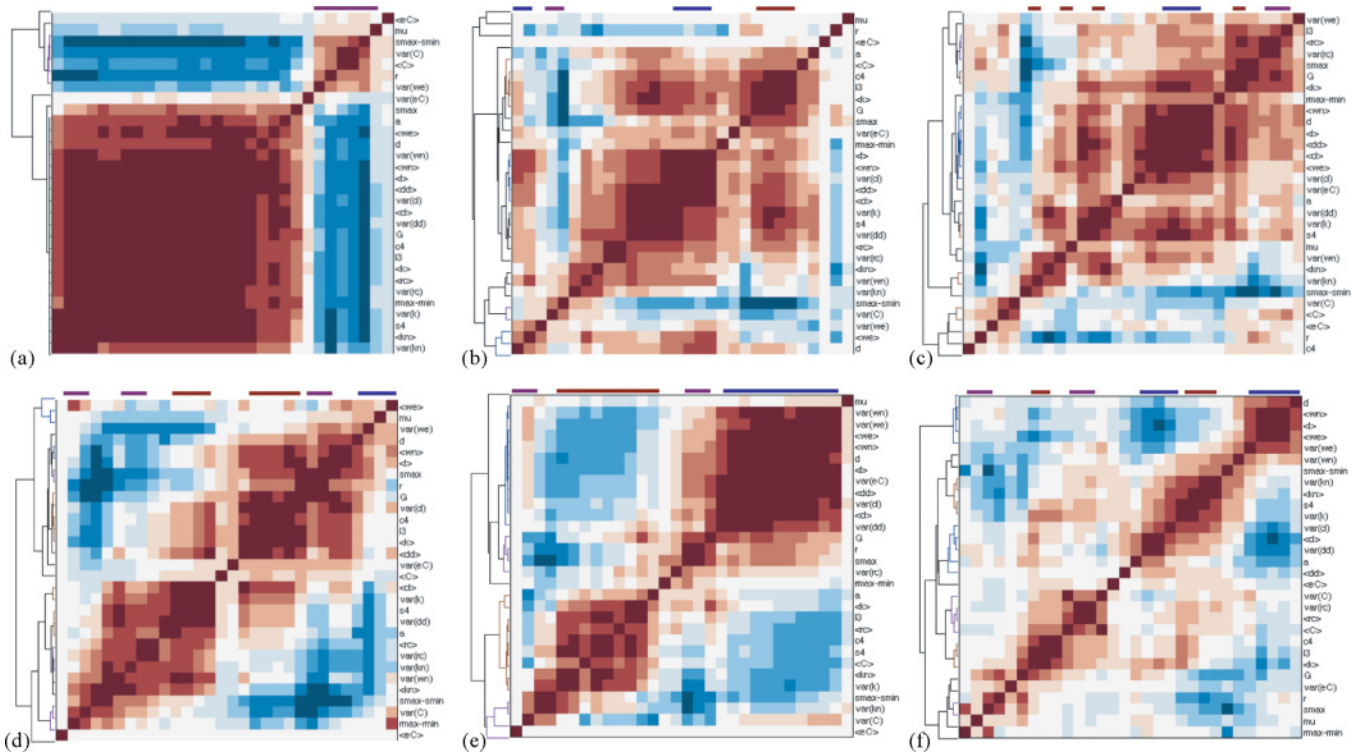

FIG. 4. (Color) Clustered heat maps for real ensembles. (a) All US airline data monthly from 1/1990 to 8/2007 (212 networks, 429–1105 nodes); (b) six US legacy carriers summer months only from 1990 to 2007 (432 networks, 64–202 nodes); (c) six US low-cost carriers summer months only from 1990 to 2007 (296 networks, 4–89 nodes); (d) Continental Airlines, all months from 1990 to 8/2007 (212 networks, 90–178 nodes); (e) Southwest Airlines, all months from 1990 to 8/2007 (212 networks, 31–79 nodes); (f) 82 language Wikipedia networks (1001–1065 nodes). For metrics nomenclature see Table I. Clusters are annotated with red for densities, blue for distances, and purple for degree-degree correlations.

Spirit, and USA3000. For the legacy and low-cost carriers only the summer month networks from 1990 to 2007 are considered, constructing ensembles of 432 and 296 graphs respectively. There are fewer low-cost months because many of these airlines are founded after 1990. Finally, to narrow the perspective even further, we select two individual airlines: Continental [Fig. 4(d)] and Southwest [Fig. 4(e)] and take 212 months of each airline history to make an ensemble. Continental is chosen as a generic example of a legacy carrier, while Southwest is chosen because it is an outlier in the industry, with known non-hub-spoke operations.

The last ensemble is an example of a nontemporal set of graphs: language Wikipedia networks [Fig. 4(f)].

Wikipedia can be represented as a network of hyperlinked pages. A node is an article on a given topic, consisting of the main article text only. A link to another page is a hyperlink within the main text. Therefore, this graph representation captures the article information content only. Wikipedia networks can be constructed for every language, because articles contain mostly references to pages in the same language. Every language Wikipedia evolves as pages get added and deleted. The set analyzed here is chosen at the snapshot in time when each language has between 1000 and 1100 pages.

For the airlines, combining all data or slicing it in legacy and low cost, or looking at individual players, yields different results. The entire US airline network [Fig. 4(a)] grows from about 450 reported airports in 1990 to over 1000 in 2007. Looking at the entire period yields a high-correlation pattern, in which only a few degree-degree correlation metrics (r, $|s_{\max}\text{-}s_{\min}|$) stand out. One explanation could be that the combined networks of all airline flights do not form a statistical ensemble. Deeper analysis shows that the graphs in this ensemble are of two types. The airline network prior to 2002 is twice as small (400–500 airports), and is denser, compared to the map after 2002 (1000 airports). Interestingly, 2002 is a transition year with intermediate size and density. This could be due to regulation changes after 2002 or to the Bureau of Transportation Statistics [31] reporting their data differently after 2002. The heat maps of the two separate time periods are similar to the legacy and low-cost data slices.

This is an additional argument that pulling data together, even from the same domain, cannot be done without careful investigation.

The legacy and low-cost airline heat maps are not too different from each other [Figs. 4(b) and 4(c)]. Some regular patterns are present, but with overall more positive correlations merging otherwise distinct clusters. The same set of degree-degree correlation metrics as in [Fig. 4(a)] stand out as independent or negatively correlated with the rest.

Single-airline ensembles are most similar to the random graph data sets. Both Continental [Fig. 4(d)] and Southwest [Fig. 4(e)] have pronounced clusters, but they do not share the same metrics. Southwest shows the typical density versus distance pattern, whereas Continental has a large degree correlations cluster, and more degree correlation metrics coupling with other metrics, similar to the randomized hierarchies map [Fig. 3(c)].

Wikipedia [Fig. 4(f)] heat maps are most decoupled among the real data sets, and closest to trees, compared to the random ensembles. Their early networks are thought to have the topology of randomized hierarchies [32], but perhaps if very sparse, they resemble trees. This argument rests on the assumption that the graphs grow under similar conditions and can be considered an ensemble.

To summarize, different classes of random graphs exhibit different metric correlations and have distinct heat maps. Erdös-Rényi graphs show two major correlation clusters of metrics: densities and distances. Random modular graphs are not very different from ER graphs. Preferential attachment graphs and spatial distribution trees differ in elasticity and scale-freeness. Airline heat maps have much more correlation and show mixing of positive correlation between the two general clusters of the random ensembles. Wikipedia networks have the most independent metrics from the real data examples and are closest to trees.

There are three main takeaways: (i) Generalization of metric correlations over arbitrary data sets is not possible because the patterns are ensemble specific and sometimes sample specific, (ii) some graphs might not form an ensemble even though they are constructed under similar conditions and rules, and (iii) the difference between two topologies can manifest itself in only one or a small set of measures.

To make the first point stronger, we will add that constructing the comprehensive heat map by merging all data sets into a single ensemble provides no meaningful result. The map does not resemble the Erdös-Rényi ensemble or any of the others.

V. TOPOLOGY METRICS DISCUSSION

The heat-map representation in Sec. IV emphasizes ensemble differences, and that generalization of metric relationships can be hard to justify for real data. Detecting differences, however, also provides useful information. The following discussion concentrates on the metrics that vary the most across ensembles.

Table II summarizes all data sets, with their size, variation in number of nodes, and average degree. Based on the heat maps, the most uncorrelated metrics are listed, along with the highest variance metrics (top 5–6). There is significant similarity in the last two columns across ensembles. The highest variance metrics tend to be all degree-degree correlations. These are the degree correlation (r), its elasticity (r_e), the maximum s metric ($s_{\max}$), the s metric elasticity ($|s_{\max}\text{-}s_{\min}|$), and the rich club distribution moments ($\langle r_c \rangle$, var(r_c)). Algebraic connectivity (a) and modularity (μ) also have high variance, but the variation is within ensembles and is not helpful for classification. Eigencentrality mean and variance ($\langle e_C \rangle$,var(e_C)) are the ubiquitous uncorrelated metrics. Given that these are used as a relative node importance, it is interesting that they do not correlate with the other centrality measures in this list. Tree ensembles have the highest number of uncorrelated metrics, because of the lack of loops, cliques, the constant average degree, and zero clustering coefficient.

To compare ensembles in fewer dimensions, we project the data onto its first three principal components and also onto some of the highest variance dimensions. Figure 5 shows the projections. The marginal histograms indicate the density of points.

Figure 5 shows not only the relative position of ensembles, but also that some sets of graphs do not behave as an

TABLE II. Ensemble statistics with uncorrelated and highest variance metrics.

Ensemble	Ensemble size	Number of nodes [mean,std]	Average degree [mean,std]	Uncorrelated metrics	Highest variance metrics
ER classic	100	[1000, 0]	[6.9, 0.11]	$\langle e_C\rangle$, var(e_C), r	a, r, μ, $\lvert s_{\max}\text{-}s_{\min}\rvert$, $\lvert r_{\max}\text{-}r_{\min}\rvert$, $s_{\max}$
ER modular	100	[1000, 0]	[6, 0.11]	$\langle e_C\rangle$	a, μ, r, $\lvert r_{\max}\text{-}r_{\min}\rvert$, $\lvert s_{\max}\text{-}s_{\min}\rvert$, $s_{\max}$
Randomized hierarchies	100	[1000, 0]	[2.50, 0.01]	$\langle d_d\rangle$, μ, $\langle e_C\rangle$	μ, r, $\lvert r_{\max}\text{-}r_{\min}\rvert$, $\lvert s_{\max}\text{-}s_{\min}\rvert$, var($r_c$), a
General preferential attachment	100	[1002, 2.2]	[9.6, 0.22]	c_4, $\langle e_C\rangle$, var(e_C)	a, r, $\lvert s_{\max}\text{-}s_{\min}\rvert$, var($r_c$), $\lvert r_{\max}\text{-}r_{\min}\rvert$, $s_{\max}$
Preferential attachment trees	100	[1000,0]	[1.99, 0]	c_4, l_3, $\langle e_C\rangle$, $\langle k\rangle$,μ, var(C), var(e_C), $\langle C\rangle$	$\lvert r_{\max}\text{-}r_{\min}\rvert$, r, $\lvert s_{\max}\text{-}s_{\min}\rvert$, $s_{\max}$, d, var(r_c)
Spatial distribution trees	100	[1000, 0]	[1.99, 0]	var(C), $\langle e_C\rangle$, μ, $\langle k\rangle$, $\langle e_C\rangle$, c_4, l_3, $\langle C\rangle$, $\langle d_d\rangle$	r, $\lvert r_{\max}\text{-}r_{\min}\rvert$, d, $\lvert s_{\max}\text{-}s_{\min}\rvert$, $s_{\max}$, $\langle l\rangle$
All US airlines	212	[660, 226]	[12.5, 0.75]	var(e_C)	μ, a, r, $\lvert r_{\max}\text{-}r_{\min}\rvert$, var($C$), $\langle C\rangle$, $\lvert s_{\max}\text{-}s_{\min}\rvert$, $s_{\max}$
US legacy airlines	432	[146, 26]	[6.2, 1.1]	μ, $\langle e_C\rangle$	μ, a, $\langle k_n\rangle$, r, $\langle C\rangle$, $\lvert r_{\max}\text{-}r_{\min}\rvert$, $s_{\max}$
US low-cost airlines	296	[33, 19.7]	[3.1, 0.84]	c_4	a, μ, $s_{\max}$, $\langle k_n\rangle$, $\lvert s_{\max}\text{-}s_{\min}\rvert$, r, $\langle C\rangle$
Southwest Airlines	212	[53, 12.6]	[11, 2.7]	$\langle C\rangle$, $\langle e_C\rangle$	a, μ, $\lvert s_{\max}\text{-}s_{\min}\rvert$, r, $s_{\max}$, $\langle r_c\rangle$
Continental Airlines	212	[143, 19]	[4.7, 0.7]	μ, $\langle e_C\rangle$	μ, a, $\langle k_n\rangle$, r, $s_{\max}$, $\lvert r_{\max}\text{-}r_{\min}\rvert$, $\langle r_c\rangle$
Wikipedia languages	82	[1015, 16.6]	[4.8, 2.0]	$\langle e_C\rangle$, var(e_C), μ	μ, $s_{\max}$, $\lvert r_{\max}\text{-}r_{\min}\rvert$, r, $\langle C\rangle$, $\langle k_n\rangle$, $\lvert s_{\max}\text{-}s_{\min}\rvert$

ensemble in the projection space. Projection onto the first three principal component axes shows the random ensembles clustering tightly and away from the airlines and Wikipedia. The real data sets show much more variation. Legacy, some of the low-cost airlines, and Continental Airlines overlap in this space. Southwest is an outlier in all projections, while the low-cost airlines do not cluster at all. This is the example of a set of graphs that do not form a good ensemble. Wikipedia networks cluster on their own. So does the all-airline data set, but in the two partitions prior to and after 2002, as explained in Sec. IV A.

Figure 5 also shows the two most interesting projections onto high variance metrics. The $s_{\max}$ versus degree correlation (r) plot shows the relative "scale-freeness" versus assortativity for all ensembles. Random ensembles such as ER, randomized hierarchies, and general scale-free graphs have the highest degree correlation and s metric (ER graphs centered at $r = 0$, as expected). Their nonskewed degree distributions explain the higher r and the higher $s_{\max}$, because they have many edges connecting nodes of similar degree.

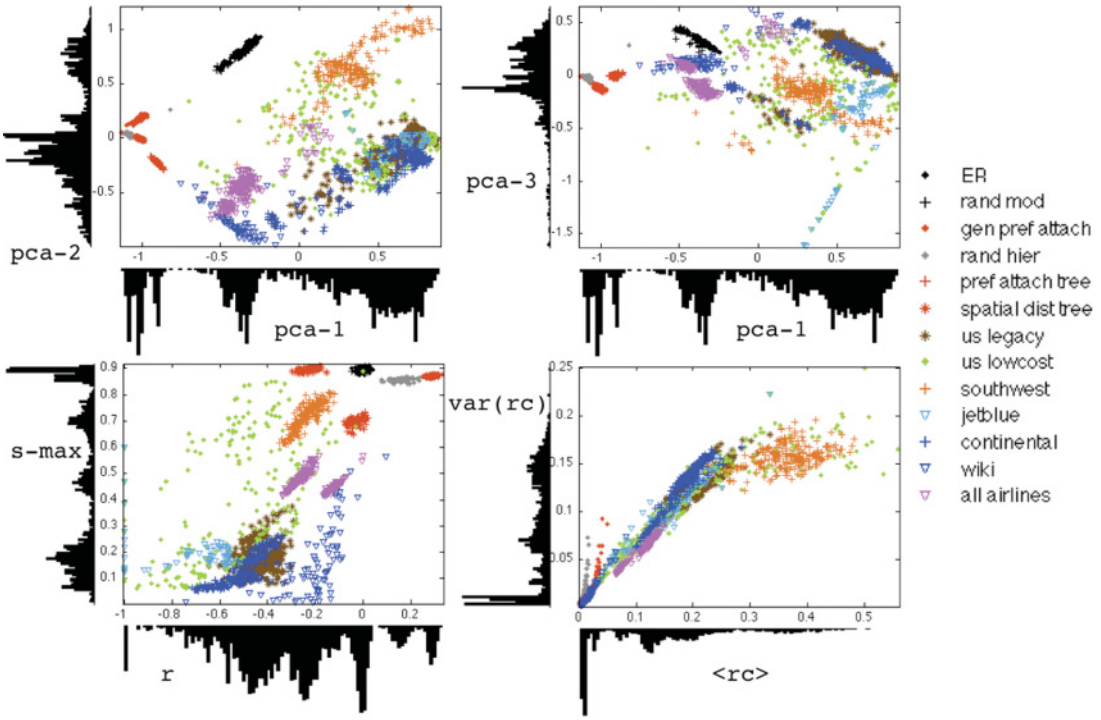

FIG. 5. (Color) All ensembles projected onto their first three principal components and onto some of the highest-variance topology descriptors: degree correlation (r), s metric ($s_{\max}$), mean rich club metric ($\langle r_c\rangle$), and rich club variance (var(r_c)).

All other data sets have relatively more skewed degree distributions, are disassortatively connected (low r), and are less scale-free in the sense defined by Li *et al.* [13].

As in the PCA projections, airline data sets occupy the same part of r-$s_{\max}$ space, except for Southwest Airlines, the all-airline ensemble and the wide spread of low-cost carriers. Airlines have a relatively low degree correlation (-0.5), because of their hub-spoke structure. The Southwest ensemble as an outlier is much closer to the random ensembles in r-$s_{\max}$ space. Previous research has also suggested that the Southwest network viewed as a simple graph resembles a random graph [32]. The combined ensemble of all airlines forms two clusters, which is consistent for most of the 30 metrics presented here. The two clusters correspond to months before and after 2002.

Another interesting relationship is the consistent positive correlation of rich club mean and rich club variance (Fig. 5). All ensembles form a line, with a distinction between random networks and airlines. Note that the rich club distribution is the set of rich club coefficients for all possible threshold degrees. For networks with uniform degree distribution, where the average degree will never be too high, the rich club distribution will have many zeros, and the mean will be low. These are the ER graphs, and randomized hierarchies rising superlinearly on the plot. It is natural that if the mean is higher, the variation is higher, because we know that the rich club sequence is decreasing. The airlines have higher average rich club coefficient and higher rich club variance. As Table II shows they are much denser networks than the random ensembles, hence there will be more links in every "rich club" at threshold degree. This probably explains the higher variance as well. It also indicates that airlines have rich club cores in their networks, especially Southwest Airlines which swings far right on this plot.

In summary, projections of all data onto lower dimensions either via principal component analysis or onto the highest variance metrics, confirms that the ensembles we studied can be clustered separately, and should be analyzed separately. The differences in their projections also confirm previously known patterns, and reveal new facts, such as the different structure of the entire US airline map before and after 2002. In addition, some proposed ensembles should probably not be analyzed as one set of data, such as the low-cost airlines.

VI. CONCLUSION

This paper presents an overview of topology metrics and their relationships for various ensembles of synthetic and real data. The metrics span degree distribution moments, degree-degree correlations, graph distance metrics, and spectral properties. The ensembles are derived from classical random graph models such as the Erdös-Rényi model and general preferential attachment. Real data sets include airline networks and Wikipedia graphs. For every ensemble we compute the set of metrics for every graph instance and compute their pairwise correlations. The correlation matrix is represented as a heat map, which is used to discuss general patterns in metric relationships and key differences between ensembles.

Finally, the ensembles are projected onto the entire data's principal component axes and onto the highest variance metrics. The projections are used to discuss ensemble clustering and proximity.

This work is a critical analysis of studying graph topology via multiple metric analysis. It provides examples of why various data sets should not be combined without careful examination to study the same patterns. This includes examples of real data that has the same type and origin but should not be analyzed as one ensemble. We show that among the metrics we study the degree-degree correlations show the most variation and potential for classifying different types of graphs. In particular, the degree correlation (r), the $s_{\max}$ and their elasticities, as well as the rich club distribution mean and variance ($\langle r_c \rangle$,var(r_c)) are among the top high variance metrics for all ensembles.

The ideas presented in this paper leave many open questions. One important question is how to analyze the topology of a single graph from this multiple-metric view. The short answer is that it is not possible. Correlations cannot be computed from one data point. A single data point can be plotted in lower dimensions against synthetic data ensembles as in Fig. 5. Another way is to generate neighborhood ensembles, analogous to bootstrapping, for example, from the same degree distribution. The most precise way to construct graphs with the same topology would be to construct random graphs not just with the same degree sequence, but with the same distribution of subgraphs up to a certain size [4]. We have experimented in this direction, but leave this as future work and out of the scope of this paper.

Finally, we think that ensemble-based statistics is a key idea in studying topology not just for random graphs. It can be enriched further by developing new ways to create neighborhood ensembles, extending the set of descriptors and attempting analysis with different correlation and statistical approaches. The heat-map representation is convenient but it does collapse some of the information in the distributions being studied. Extending the computation to larger graphs and a wider set of descriptors will provide a higher resolution view of graph topology.

[1] D. Alderson, Oper. Res. **56**, 1047 (2008).
[2] D. Alderson and L. Li, Phys. Rev. E **75**, 046102 (2007).
[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[4] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, Systematic Topology Analysis and Generation Using Degree Correlations, SIGCOMM06, September 11–15, 2006, Pisa, Italy.
[5] D. Braha and Y. Bar-Yam, Phys. Rev. E **69**, 016113 (2004).
[6] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Nature Phys. **3**, 63 (2007).
[7] P. Erdös and A. Rényi, Publ. Math. Debrecen **6**, 290 (1959).
[8] M. E. J. Newman, Proc. Natl. Acad. Sci. **103**, 8577 (2006).

[9] C. Baldwin and K. Clark, Harvard Business Review **75**, 84 (1997).
[10] A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[11] P. S. Dodds, D. J. Watts, and C. F. Sabel, Proc. Natl. Acad. Sci. USA **100**, 12516 (2003).
[12] M. Gastner and M. E. J. Newman, Eur. Phys. J. B **49**, 247 (2006).
[13] L. Li, D. Alderson, J. Doyle, and W. Willinger, in *Internet Mathematics*, Vol. 2, No. 4 (Taylor & Francis, New York, 2005), pp. 431–523.
[14] A. Clauset, Phys. Rev. E **72**, 026132 (2005).
[15] J. M. Carlson and J. Doyle, Phys. Rev. E **60**, 1412 (1999).
[16] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[17] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Networks to the Internet and WWW* (Oxford University Press, Oxford, 2003).
[18] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, 1st ed. (Cambridge University Press, Cambridge, 2008).
[19] S. Roy and V. Filkov, Phys. Rev. E **80**, 040902 (2009).
[20] V. Filkov, Z. M. Saul, S. Roy, D. Souza, R. M. Devanbu, and P. T. Modeling, Europhys. Lett. **86**, 28003 (2009).
[21] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[22] D. J. Watts and S. Strogatz, Nature (London) **393**, 440 (1998).
[23] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[24] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani, Nature Phys. **2**, (2006).
[25] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[26] I. Gutman, Ber. Math. Statist. Sekt. Forschungszentrum Graz. **103**, 1 (1978).
[27] K. Sinha and O. de Weck, Spectral and Topological Features of Real-World Product Structures, 11th International Design Structure Matrix Conference, DSM'09, 12-13 October, 2009, Greenville, South Carolina.
[28] M. Fiedler, Czechoslovak Mathematical Journal **23**, 298 (1973).
[29] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, New York, 2005.
[30] M. Newman, A.-L. Barabási, and D. J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).
[31] Bureau of Transportation Statistics, http://www.bts.gov.
[32] G. Bounova, Ph.D. thesis, Massachusetts Institute of Technology, 2009.