How Small Are Building Blocks of Complex Networks

Almerima Jamakovic,[1] Priya Mahadevan,[2] Amin Vahdat,[3] Marián Boguñá,[4] and Dmitri Krioukov[5]

[1] *TNO Information and Communication Technology, Netherlands Organisation for Applied Scientific Research, P.O. Box 5050, 2600 GB Delft, The Netherlands*
[2] *HP Labs, 1501 Page Mill Rd, Palo Alto, California, 94304, USA*
[3] *Department of Computer Science and Engineering, University of California, San Diego (UCSD), 9500 Gilman Drive, La Jolla, California 92093, USA*
[4] *Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*
[5] *Cooperative Association for Internet Data Analysis (CAIDA), University of California, San Diego (UCSD), 9500 Gilman Drive, La Jolla, California 92093, USA*

Network motifs are small building blocks of complex networks, such as gene regulatory networks. The frequent appearance of a motif may be an indication of some network-specific utility for that motif, such as speeding up the response times of gene circuits. However, the precise nature of the connection between motifs and the global structure and function of networks remains unclear. Here we show that the global structure of some real networks is statistically determined by the distributions of local motifs of size at most 3, once we augment motifs to include node degree information. That is, remarkably, the global properties of these networks are fixed by the probability of the presence of links between node triples, once this probability accounts for the degree of the individual nodes. We consider a social web of trust, protein interactions, scientific collaborations, air transportation, the Internet, and a power grid. In all cases except the power grid, random networks that maintain the degree-enriched connectivity profiles for node triples in the original network reproduce all its local and global properties. This finding provides an alternative statistical explanation for motif significance. It also impacts research on network topology modeling and generation. Such models and generators are guaranteed to reproduce essential local and global network properties as soon as they reproduce their 3-node connectivity statistics.

I. INTRODUCTION

A promising direction in the studies of the structure and function of complex networks is to identify their building blocks, or motifs [1–3], which are small subgraphs in a real network. A great deal of research—in particular, research on gene regulatory networks—shows that specific motifs perform specific functions, such as speeding up response times of regulatory networks [4, 5]. However, motifs have also raised many questions [6–13], including continuing debates on whether and how motif statistical profiles are related to the *global* structure, function, and evolution of certain networks.

Our recent work [14] introduces dK-series, see Section II. The dK-series, with analogy to the Taylor or Fourier series, is the first systematic and complete basis for characterizing network structure. The dK-series is a generalization of known degree-based statistical characteristics of complex networks. The zero-th element of the dK-series, the $0K$-"distribution," is the average degree in a given network. The first element, the $1K$-distribution, is the network's degree distribution, or the number of nodes—subgraphs of size 1—of degree k. The second element, the $2K$-distribution, is the joint degree distribution, the number of subgraphs of size 2—links—involving nodes of degrees k_1 and k_2. For $d = 3$, the subgraphs are triangles and wedges, composed of nodes of degrees k_1, k_2, and k_3. Generalizing, the dK-distribution is the numbers of different subgraphs of size d involving nodes of degrees $k_1, k_2, \ldots, k_d$.

The dK-series is systematic and complete because it is *inclusive* and *converging*. Inclusiveness results from the fact that the $(d+1)K$-distribution contains the same information about the network as the dK-distribution, plus some additional information. That is, by increasing d, we provide increasingly more detail about the network structure. As d increases toward the network size, we fully specify the entire network structure, which explains the second *convergence* property of dK-series—it converges to the given network in the limit of large d.

Does this convergence happen only at d equal to the network size, or much sooner, at smaller d? In other words, how much *local* information, i.e., information about concentrations of degree-labeled subgraphs of what size, is needed to fully capture *global* network structure?

To answer these questions, we must compare a real network with typical random networks defined by its dK-distribution. If there is no difference between such dK-*random* networks and the real network, then the latter is fixed by its dK-distribution. To obtain a dK-random version of the real network, we dK-randomize it as illustrated in Fig. 1(a)—we randomly rewire (pairs of) links preserving the dK-distribution in the network, generalizing known network randomization techniques [17, 18] used to compute motif statistical significance. The result of this dK-randomization procedure are random networks that have the same dK-distribution as the original real network, but that are maximally random in all other respects.

Our question thus becomes what is the minimum value of d such that there is no difference between a real net-

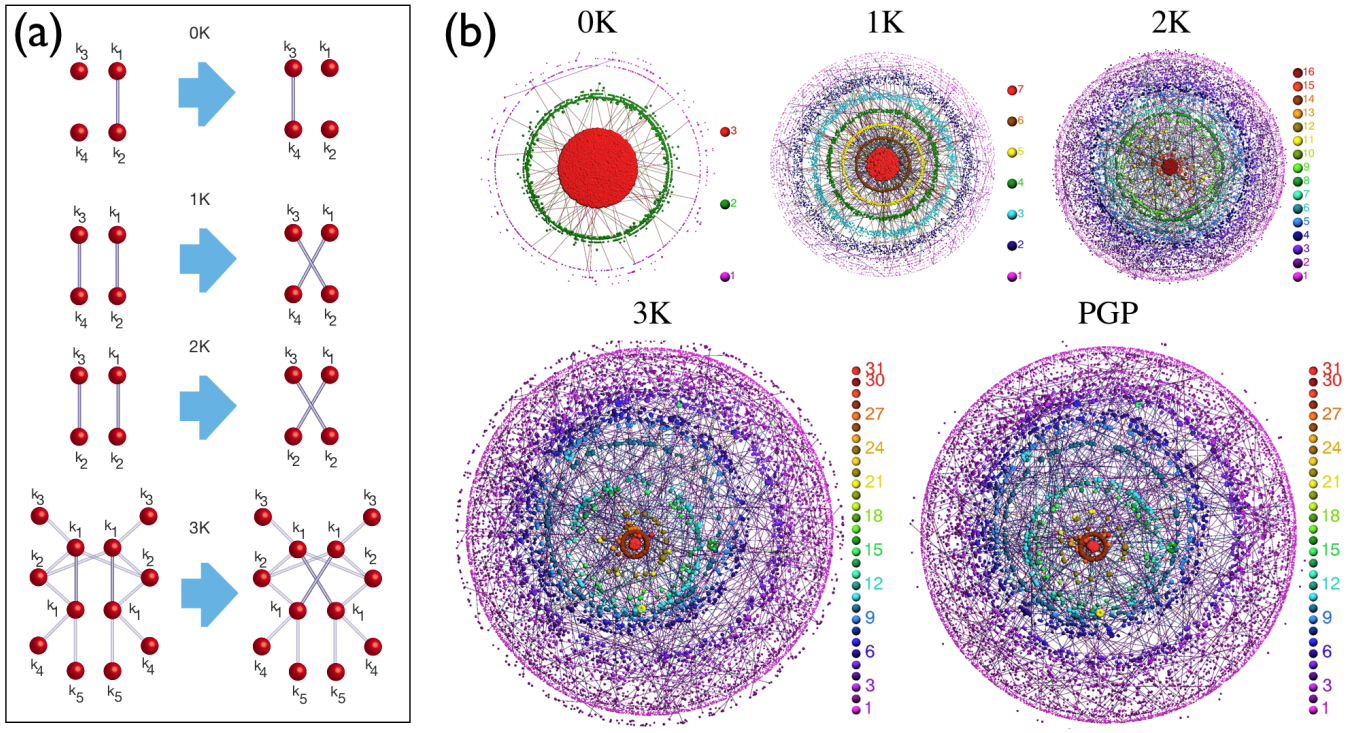


FIG. 1: **The dK-randomization null models for $d = 0, 1, 2, 3$. a) Illustration of dK-randomizing rewiring.** All nodes are labeled by their degrees, and a dK-rewiring preserves the graph's dK-distribution, and consequently its $d'K$-distributions for all $d' < d$, but randomizes the $d''K$-distributions for $d'' > d$. The $0K$-randomization involves rewiring of a link to any pair of disconnected nodes, which preserves the average degree only. The $1K$-randomization preserves the degree distribution, too, by rewiring a pair of links as shown. The $2K$-distribution preserves the joint degree distribution as well, because at least two nodes adjacent to the rewired links are of the same degree. The $3K$-randomization preserves the number of degree-labeled wedges and triangles. As d increases, the rewiring becomes increasingly more constrained since fewer links can be rewired without altering the dK-distribution. To dK-randomize a network, we randomly select a pair of links, and rewire them if they can be dK-rewired, or, if they cannot be rewired, select another random pair. This process is repeated for a sufficient number of successful rewirings, i.e., until all network properties stop changing, at which point we say that the graph has converged to its dK-randomization. **b) Visualization of the social web of trust (PGP network [15]) and its dK-randomizations.** We use the LaNet-vi tool [16] for visualization, which encodes the node coreness in color, see the right legends. The coreness is a measure of node centrality, i.e., how deeply in the network core the node lies [16]. Nodes with larger coreness are also placed closer to the circle centers. The quick convergence of the dK-randomizations to the original PGP network, and the similarity between it and its $3K$-randomization are remarkable.

work and its dK-randomizations? It seems at first that the answer to this question should strongly depend on the specific networks we consider.

We consider a variety of social, biological, transportation, communication, and technological networks, see Section III. Although the dK-series applies to directed and even annotated networks [19], here we report results for undirected networks. The dK-distributions for directed or annotated networks contain more information than for undirected networks. Therefore, dK-series converges faster in the former case [19]. Below we show the results for the well-studied social web of trust relationships extracted from Pretty Good Privacy (PGP) data [15]. The results for all other networks, except the power grid, are similar, cf. Section IV, where we also discuss possible reasons for why the power grid appears as an exception.

Fig. 1(b) visualizes the PGP network and its dK-randomizations. We observe that the dK-series converges at $d = 3$. While the $0K$-random network has little in common with the real network, the $1K$-random one is somewhat more similar, even more so for $2K$, and there is very little difference between the real PGP network and its $3K$-random counterpart.

To provide a more detailed and insightful comparison between the real network and its dK-randomizations, we compute a variety of metrics for each. Some popular metrics, such as degree distribution, average nearest neighbor connectivity, clustering, etc., are functions, sometimes peculiar, of dK-distributions, and therefore it is not surprising that they are properly captured by dK-series, as confirmed in Section IV A. We classify metrics that do not explicitly depend on dK-distributions as microscopic, mesoscopic, and macroscopic. We choose them to probe

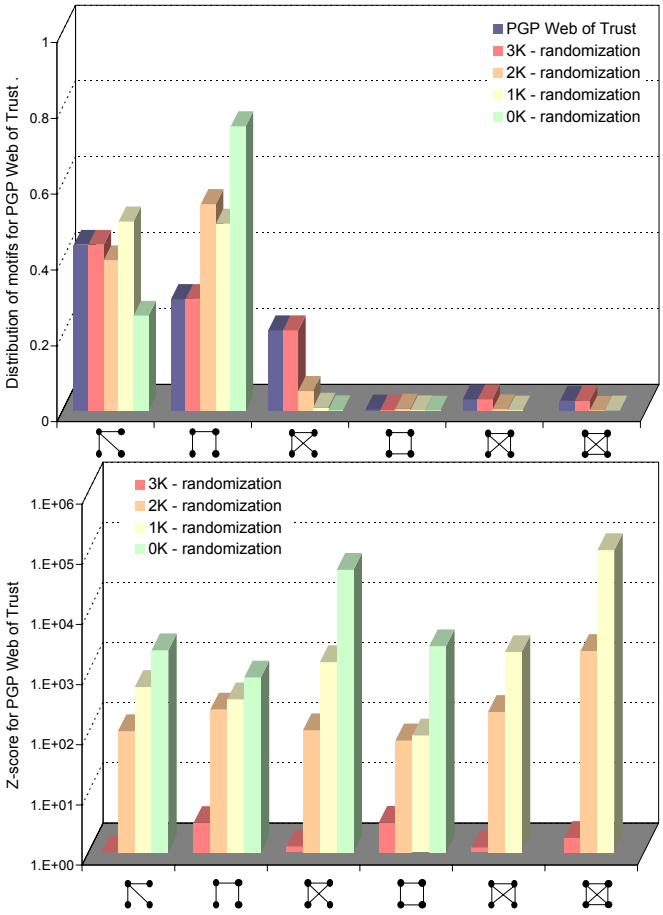


FIG. 2: **Microscopic scale: motifs.** There are six different graphs of size 4 shown on the x-axes. The **top** plot shows the distribution of the numbers of these subgraphs in the PGP network and its dK-randomizations, $d = 0, 1, 2, 3$. Each blue bar, for example, is the number of the corresponding subgraph occurrences in the PGP network divided by the total number of subgraphs of size 4 in it. For dK-randomizations, the values are averaged, for each d, over several realizations of the dK-randomized network. In the case of $0K$-randomization, the last two motifs did not occur in any randomized sample of the network. The **bottom** plot shows the Z-scores for the six subgraphs in the four dK-randomization null models. The Z-score [1] of a subgraph is a measure of its statistical significance in a real network, compared to a randomization null model. Specifically, the Z-score Z is the difference between the number N of the occurrences of a subgraph in the real network and the average number $\bar{N}$ of its occurrences in the corresponding randomized networks, divided by the standard deviation σ of its occurrences in the randomized networks, $Z = |N - \bar{N}|/\sigma$.

the network structure at the local, medium, and global scales.

The simplest *microscopic*, local-structure statistics, which are not fixed by the dK-distributions with $d \leqslant 3$, are the frequencies of motifs of size 4 without degree information. We compute these frequencies in the real network and its dK-randomizations, and show the results in

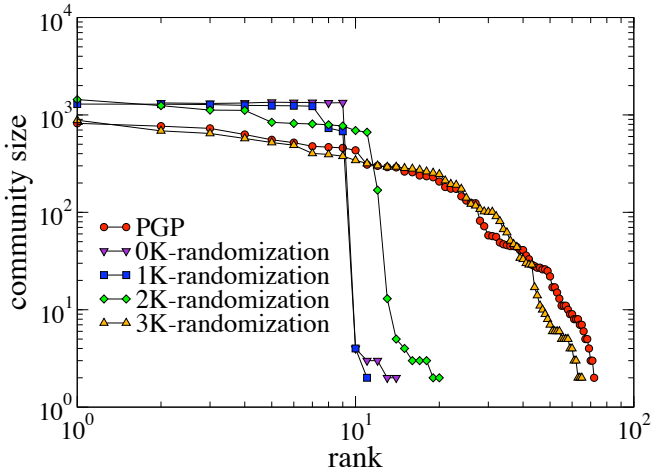


FIG. 3: **Mesoscopic scale: community structure.** We compute communities in the PGP network using the Extremal Optimization algorithm [20]. We then sort the found communities in the order of decreasing size. The size of a community is the number of nodes in it. The rank of a community is its position number in the size-ordered list. We then show the community size distribution by plotting the community sizes vs. their ranks.

Fig. 2. We find that the (relative) statistical significance of the motifs strongly depends on d. More importantly, no motif is statistically significant for $d = 3$.

At the *mesoscopic* scale, we consider the community structure of the PGP network. A community is a subgraph with many internal connections, and a relatively small number of connections external to the subgraph. Fig. 3 shows that the community structure is indeed a "mesoscopic" metric because the community sizes range from a few nodes to thousands of nodes for largest communities. Fig. 3 shows that the community size distributions in the PGP network and its $3K$-randomization are very similar.

At the *macroscopic* scale, we consider the two most popular and important statistics that depend on a network's global structure: the node betweenness centrality and the distribution of lengths of shortest paths in a network. Fig. 4 once again shows that $3K$ is sufficient to capture even global graph properties; the considered metrics are approximately the same for the PGP network and its $3K$-randomization.

We call a given real network *dK-random* if *all* its metrics, at all scales from local to global, are approximately the same as the corresponding metrics in its dK-randomizations. We see in Section IV that in agreement with the results of Vázquez *et al.* [12], almost all networks that we collected data for are $3K$-random at most (some networks are $2K$- or even $1K$-random). That is, surprisingly, *the global structure of these networks is captured entirely by the distribution of node triples and their degrees.*

It is an open question why many different real net-

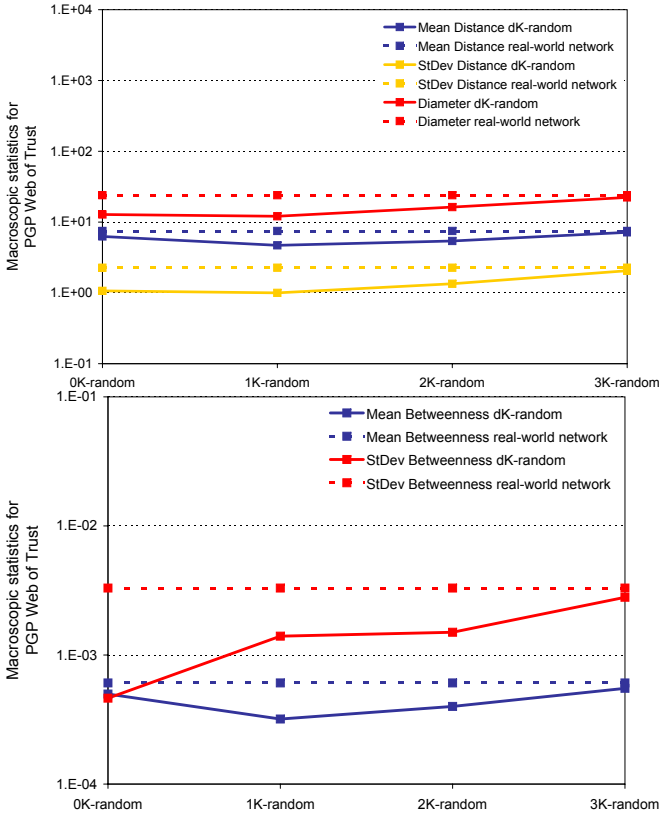


FIG. 4: **Macroscopic scale: the distance and betweenness distributions.** The **top** plot shows the metrics related to the hop length of shortest paths, or distances, between nodes in the PGP network and its dK-randomizations. These metrics are the average and maximum distance between nodes, the latter called the network *diameter*, and the standard deviation of the distance distribution. The **bottom** plot shows the average betweenness and the standard deviation of the betweenness distribution of nodes in the PGP network and its dK-randomizations. The betweenness of a node is a measure of its communication centrality [21]. It is equal to the number of shortest paths passing through the node, divided by the total number of shortest paths between the same source and destination, summed over all source-destinations pairs. In both plots the values for dK-randomizations are averaged, for each d, over several realizations of the dK-randomized network.

works are $3K$-random. A trivial answer would be that $d = 3$ is just "constraining enough." There may only be a few possible rewirings preserving the $3K$-distribution. But why exactly is $d = 3$ sufficient for real networks? There are many classes of synthetic graphs, such as lattices, for which no d substantially smaller than the graph size is "constraining enough." Perhaps the answer can be obtained by studying the hidden metric spaces underlying real networks [22]. The distances in such spaces abstract intrinsic similarities between nodes. If these spaces are metric—and there is empirical evidence that they are indeed such [23]—then the triangle inequality naturally yields and explains network clustering, which the $3K$-distribution captures by definition.

Whatever the actual explanation, our results have diverse implications. First, our dK-randomization basis makes it clear that there is no preferred null model for network randomization. To tell how statistically important a given motif is, it is necessary to compare its frequency in the real network with the same frequency in a network randomization, a null model. But one can dK-randomize any network for any d. Therefore choosing any specific value of d, or more generally, any specific null model to compute motif significance requires some non-trivial justification.

The second implication concerns the difference between motifs and dK-series. This difference is small but crucial. Motifs are subgraphs whose nodes can have any degree in the original network, while dK-series preserves the information about these degrees. This difference is crucial because a motif-based series cannot be inclusive. Node degrees are necessary to make the series inclusive and thus systematic, see Section V.

Our finding that many networks are $3K$-random can assist our understanding of how functions of an evolving network shape its structure. Indeed, one can potentially simplify such explanations to how the observed $3K$-distribution has emerged in the network. As soon as one explains the emergence of the $3K$-distribution, all other network structural properties follow.

Finally, our work very practically impacts the design of network topology models and generators. For simulation experiments, hypothesis testing, etc., network researches in many sciences, including biology [9, 24–26] and computer science [27–30], must model real networks in laboratory settings, and generate random graphs that reproduce important properties of the real network. Our results show that it is sufficient to generate $3K$-random graphs for such purposes. But even if these graphs do not capture some important property not previously considered, the dK-series will remain applicable given its convergence property and a sufficient increase in d.

We conclude this introduction with a reference to [19] for a detailed discussion of various graph generation techniques based on dK-series and extensions to generate random graphs with rich semantic, structural, or functional annotations of nodes and links.

II. THE dK-SERIES ILLUSTRATED

In Fig. 5(a) we illustrate dK-series for a graph of size 4. The $4K$-distribution is the graph itself. The $3K$-distribution consists of its three subgraphs of size 3: one triangle connecting nodes of degrees 2, 2, and 3, and two wedges connecting nodes of degrees 2, 3, and 1. The $2K$-distribution is the joint degree distribution in the graph. It specifies the number of links (subgraphs of size 2) connecting nodes of different degrees: one link connects nodes of degrees 2 and 2, two links connect nodes of degrees 2 and 3, and one link connects nodes of degree

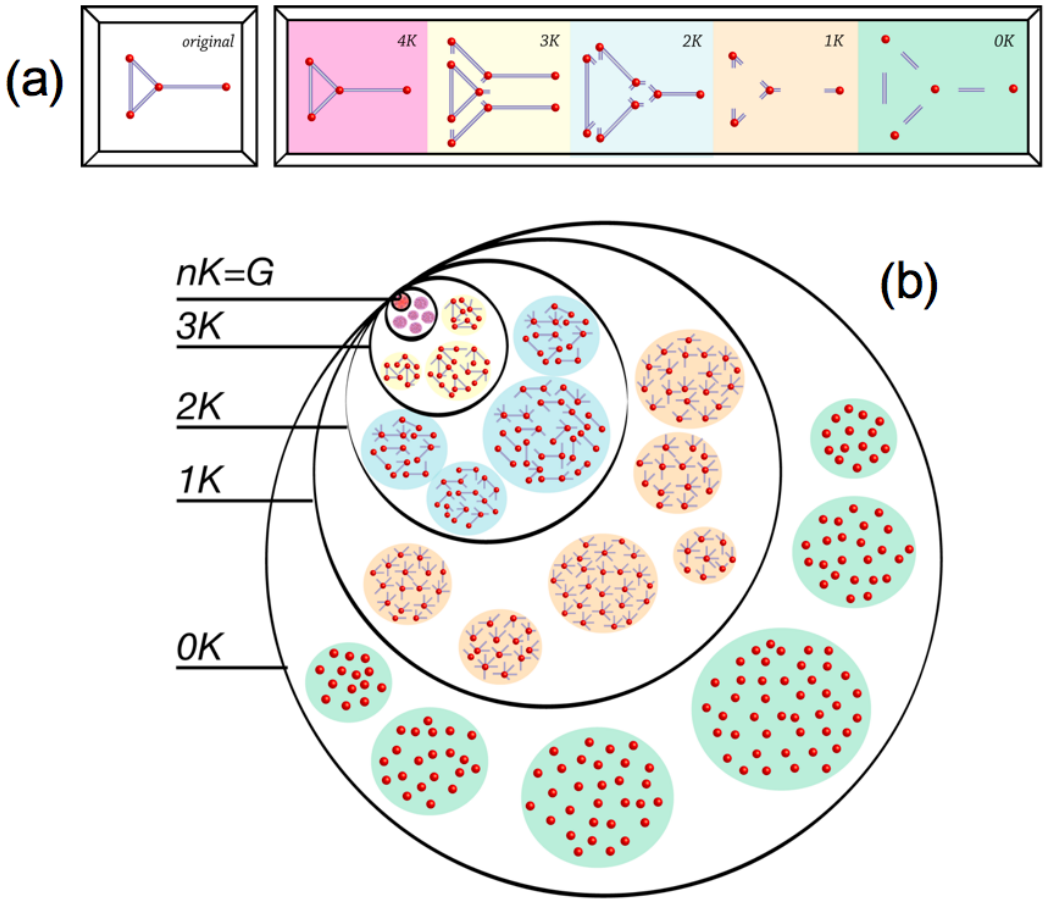


FIG. 5: The dK-series illustrated: **a)** dK-distributions for a graph of size 4; **b)** convergence and inclusiveness of dK-series.

3 and 1. The $1K$-distribution is the degree distribution in the graph. It lists the number of nodes (subgraphs of size 1) of different degree: one node of degree 1, two nodes of degree 2, and one node of degree 3. The $0K$-distribution is just the average degree in the graph, which is 2.

Fig. 5(b) illustrates the inclusiveness and convergence of dK-series by showing the hierarchy of dK-graphs, which are graphs that have the same dK-distribution as some graph G of size n. The black circles schematically shows the sets of dK-graphs.

The set of $0K$-graphs is largest: the number of different graphs that have the same average degree as G is enormous. These graphs may have a structure drastically different from G's. The set of $1K$-graphs is a subset of $0K$-graphs, because each graph with the same degree distribution as in G has also the same average degree as G, but not *vice versa*. As a consequence, typical ("maximally random") $1K$-graphs tend to be more similar to G than $0K$-graphs. The set of $2K$-graphs is a subset of $1K$-graphs, also containing G.

As d increases, the circles become smaller because the number of different dK-graphs decreases. Since all the dK-graph sets contain G, the circles "zoom-in" on it, and while their number decreases, dK-graphs become increasingly more similar to G. In the $d = n$ limit, the set of nK-graphs consists of only one element, G itself.

III. THE REAL NETWORKS CONSIDERED

We collected data for a number of real networks. We wanted the set of considered networks to be representative, in the sense that it should contain networks of different nature, coming from different domains, thus showing the universality of our dK-basis. The considered networks include social, biological, transportation, and technological networks. Specifically, we report results for:

- The social web of trust relationships among people. The trust relationships are inferred using the data from the Pretty Good Privacy (PGP) encryption algorithm [15]. We extract the strongly connected component from this network. The nodes are people, and there is a link between two people if they trust each other.

- The social network of scientific collaborations extracted from the `arXiv` condensed-matter

TABLE I: The considered networks and their abbreviations.

Network	Abbreviation
PGP Web of Trust [15]	PGP
Scientific collaboration network [31]	Collab.
Protein interaction network [32]	Protein
US air transportation network [33]	Air
Internet at the level of ASs [34]	Internet
Power grid in the western US [35]	Power

database [31]. The nodes are authors, and there is a link between two authors if they co-authored a paper.

- The biological network of protein interactions in the yeast *Saccharomyces cerevisiae* collected from the database of interacting proteins [32]. The nodes are proteins, and there is a link between two proteins if they interact.

- The US air transportation network [33]. The nodes are airports, and there is a link between two airports if there is a direct flight between them.

- The topology of the Internet at the level of Autonomous Systems (ASs) [34]. The nodes are ASs, i.e., organizations owing parts of the Internet infrastructure, and there is a link between two ASs if they are physically connected.

- The electrical power grid in the western US [35]. The nodes are generators, transformers, or substations, two of which are linked if there is a high-voltage transmission line between them.

Table I lists these networks and their abbreviations used in the subsequent figures and tables.

IV. TOPOLOGIES OF REAL NETWORKS AND THEIR dK-RANDOMIZATIONS

In this section we compare the real networks to their dK-randomizations across a number of topological metrics.

A. Metrics defined by dK-distributions

We first consider the most basic metrics, which are defined by the appropriate dK-distributions. Therefore it is not surprising that dK-random graphs with appropriate d have the values of these metrics equal exactly to those in the real networks. Nevertheless, we report these results for consistency and illustration purposes.

1. $1K$: degree distribution

Fig. 6 shows the distributions $P(k)$ of node degrees k:

$$P(k) = \frac{N(k)}{N}, \tag{1}$$

where $N(k)$ is the number of nodes of degree k in the network, and N is the total number of nodes in it, so that $P(k)$ is normalized, $\sum_k P(k) = 1$ (we do not consider nodes of degree $k = 0$). The $1K$-distribution fully defines the $0K$-distribution, i.e., the average degree $\bar{k}$ in the network, by

$$\bar{k} = \sum_k kP(k), \tag{2}$$

but not *vice versa*.

We observe in Fig. 6 that while $0K$-randomizations are off, the $1K$-random graphs reproduce the degree distributions in the real networks exactly, which is by dentition: the $1K$-distribution is the degree distribution, and $1K$-randomization does not alter it. The dK-randomizations with $d > 1$ do not alter the $1K$-distribution either, therefore they also match the degree distributions in the real networks exactly (not shown).

2. $2K$: average neighbor degree

Fig. 7 shows the average degree $\bar{k}_{nn}(k)$ of neighbors of nodes of degree k. This function is a commonly used projection of the joint degree distribution (JDD) $P(k, k')$, i.e., the $2K$-distribution. The JDD is defined as

$$P(k, k') = \mu(k, k')\frac{N(k, k')}{2M}, \tag{3}$$

where $N(k, k') = N(k', k)$ is the number of links between nodes of degrees k and k' in the network, M is the total number of links in it, and

$$\mu(k, k') = \begin{cases} 2 & \text{if } k = k', \\ 1 & \text{otherwise,} \end{cases} \tag{4}$$

so that $P(k, k')$ is normalized, $\sum_{k,k'} P(k, k') = 1$. The $2K$-distribution fully defines the $1K$-distribution by

$$P(k) = \frac{\bar{k}}{k}\sum_{k'} P(k, k'), \tag{5}$$

but not *vice versa*. The average neighbor degree $\bar{k}_{nn}(k)$ is a projection of the $2K$-distribution $P(k, k')$ via

$$\bar{k}_{nn}(k) = \frac{\bar{k}}{kP(k)}\sum_{k'} k'P(k, k') = \frac{\sum_{k'} k'P(k, k')}{\sum_{k'} P(k, k')}. \tag{6}$$

We observe in Fig. 7 that while $0K$-randomizations are way off, the $1K$-randomization are much closer to the

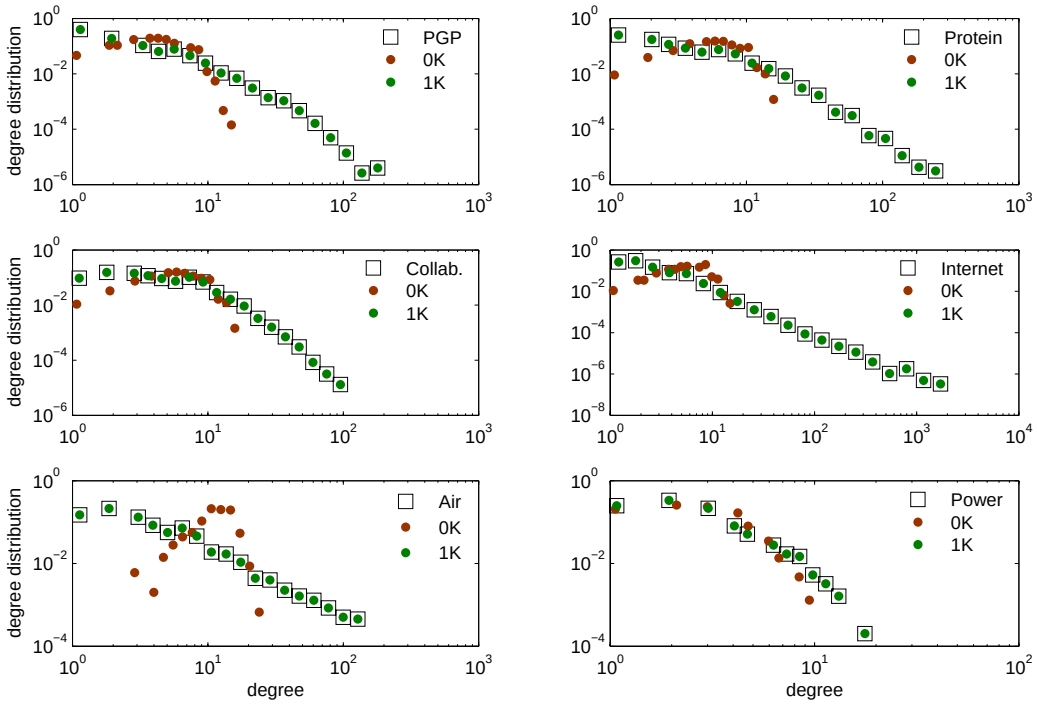


FIG. 6: The degree distribution in the real networks and their dK-randomizations.

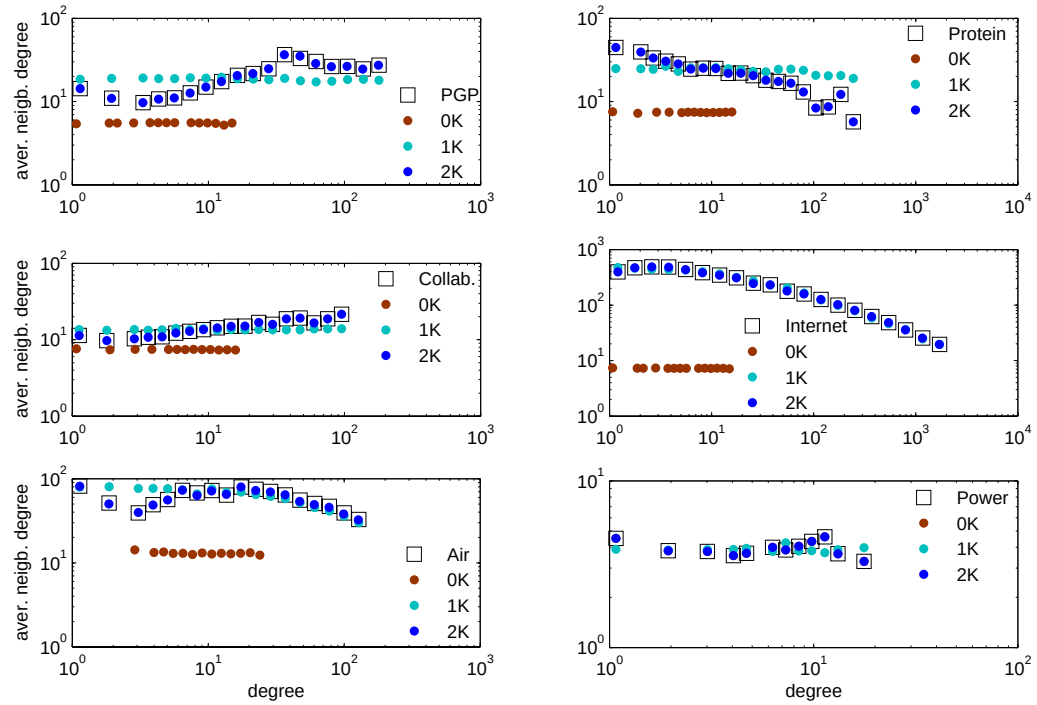


FIG. 7: The average neighbour degree in the real networks and their dK-randomizations.

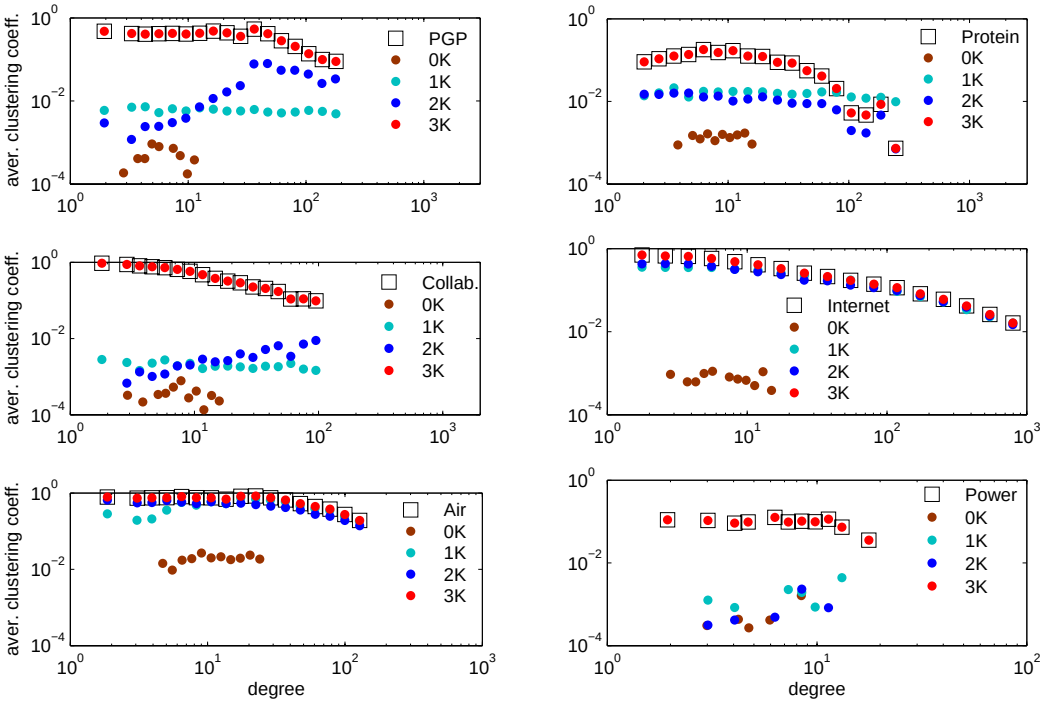


FIG. 8: The degree-dependent clustering in the real networks and their dK-randomizations.

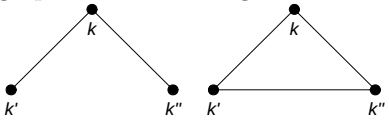
real networks, whereas the $2K$-randomizations have exactly the same average neighbor degrees as the real networks, which is again by definition: $2K$-randomization does not change $P(k,k')$. In the Internet case, even $1K$-randomization does not noticeably affect $\bar{k}_{nn}(k)$. The dK-randomizations with $d > 2$ do not alter $P(k,k')$ and consequently $\bar{k}_{nn}(k)$ at all, therefore they reproduce the latter exactly as well for all the networks (not shown).

3. $3K$: clustering

Fig. 8 shows degree-dependent clustering $\bar{c}(k)$. Clustering of node i is the number of triangles $\triangle_i$ it forms, or equivalently the number of links among its neighbors, divided by the maximum such number, which is $k(k-1)/2$, where k is i's degree, $\deg(i) = k$. Averaging over all nodes of degree k, the degree-dependent clustering is

$$\bar{c}(k) = \frac{2\triangle(k)}{k(k-1)N(k)}, \text{ where } \triangle(k) = \sum_{i:\,\deg(i)=k} \triangle_i. \quad (7)$$

The degree-dependent clustering is a commonly used projection of the $3K$-distribution [38]. The $3K$-distribution is actually two distributions characterizing the concentrations of the two non-isomorphic degree-labeled subgraphs of size 3, wedges and triangles:



Let $N_\wedge(k',k,k'') = N_\wedge(k'',k,k')$ be the number wedges involving nodes of degrees k, k', and k'', where k is the central node degree, and let $N_\triangle(k,k',k'')$ be the number of triangles consisting of nodes of degrees k, k', and k'', where $N_\triangle(k,k',k'')$ is assumed to be symmetric with respect to all permutations of its arguments. Then the two components of the $3K$-distribution are

$$P_\wedge(k',k,k'') = \mu(k',k'')\frac{N_\wedge(k',k,k'')}{2W}, \quad (8)$$

$$P_\triangle(k,k',k'') = \nu(k,k',k'')\frac{N_\triangle(k,k',k'')}{6T}, \quad (9)$$

where T and W are the total numbers of triangles and wedges in the network, and

$$\nu(k,k',k'') = \begin{cases} 6 & \text{if } k = k' = k'', \\ 1 & \text{if } k \neq k' \neq k'', \\ 2 & \text{otherwise,} \end{cases} \quad (10)$$

so that both $P_\wedge(k',k,k'')$ and $P_\triangle(k,k',k'')$ are normalized, $\sum_{k,k',k''} P_\wedge(k',k,k'') = \sum_{k,k',k''} P_\triangle(k,k',k'') = 1$. The $3K$-distribution defines the $2K$-distribution (but not *vice versa*), by

$$P(k,k') = \frac{1}{k+k'-2}\sum_{k''}\left\{\frac{6T}{M}P_\triangle(k,k',k'') + \frac{W}{M}\left[P_\wedge(k',k,k'') + P_\wedge(k,k',k'')\right]\right\}. \quad (11)$$

The normalization of $2K$- and $3K$-distributions implies the following identity between the numbers of triangles, wedges, edges, nodes, and the second moment of the degree distribution $\bar{k^2} = \sum_k k^2 P(k)$:

$$2\frac{3T + W + M}{N} = \bar{k^2}. \quad (12)$$

The degree-dependent clustering coefficient $\bar{c}(k)$ is the following projection of the $3K$-distribution

$$\bar{c}(k) = \frac{6T}{N}\frac{\sum_{k',k''} P_\triangle(k,k',k'')}{k(k-1)P(k)}. \quad (13)$$

We observe in Fig. 8 that clustering in the real networks and their dK-randomizations with $d = 3$ is exactly the same, which is again by definition. For $d < 3$, clustering differs drastically in many cases, except for the air transportation network and especially the Internet. Therefore we can say that the Internet is very close to being $1K$-random, i.e., fully defined by its degree distribution, as far as the dK-based metrics are concerned. Neither $3K$-, $2K$-, nor even $1K$-randomization alter its dK-based (projection) metrics noticeably.

B. Motifs and their Z-scores

There are six non-isomorphic motifs of size 4, shown as the x-axes in Figs. 9,10. For each network and for each $d = 0, 1, 2, 3$, we obtain several dK-randomized samples of the network, and then for each motif we compute its distribution (normalized to the total number of subgraphs of size 4) in the real network, and its average distribution in the dK-randomized samples of the network. The results are in Fig. 9. Fig. 10 reports the corresponding Z-scores. In certain cases, often for $0K$-randomizations, some motifs do not occur at all in any randomized samples, which explains the absence of some bars in the figures.

The key observation is that when the randomization null model is $3K$, the distributions of all motifs in the randomizations of all the networks except the power grid, are close to those in the real networks. The corresponding Z-scores are either low or zero. In other words, all motifs are statistically non-significant.

C. Distance and betweenness distributions

Fig. 11 shows the distance distribution in the real networks and in their dK-randomizations. The distance distribution is the distribution of hop-lengths of shortest paths between nodes in a network. Formally, if $N(h)$ is the number of node pairs located at hop distance h from each other, then the distance distribution $\delta(h)$ is

$$\delta(h) = \frac{2N(h)}{N(N-1)}, \quad (14)$$

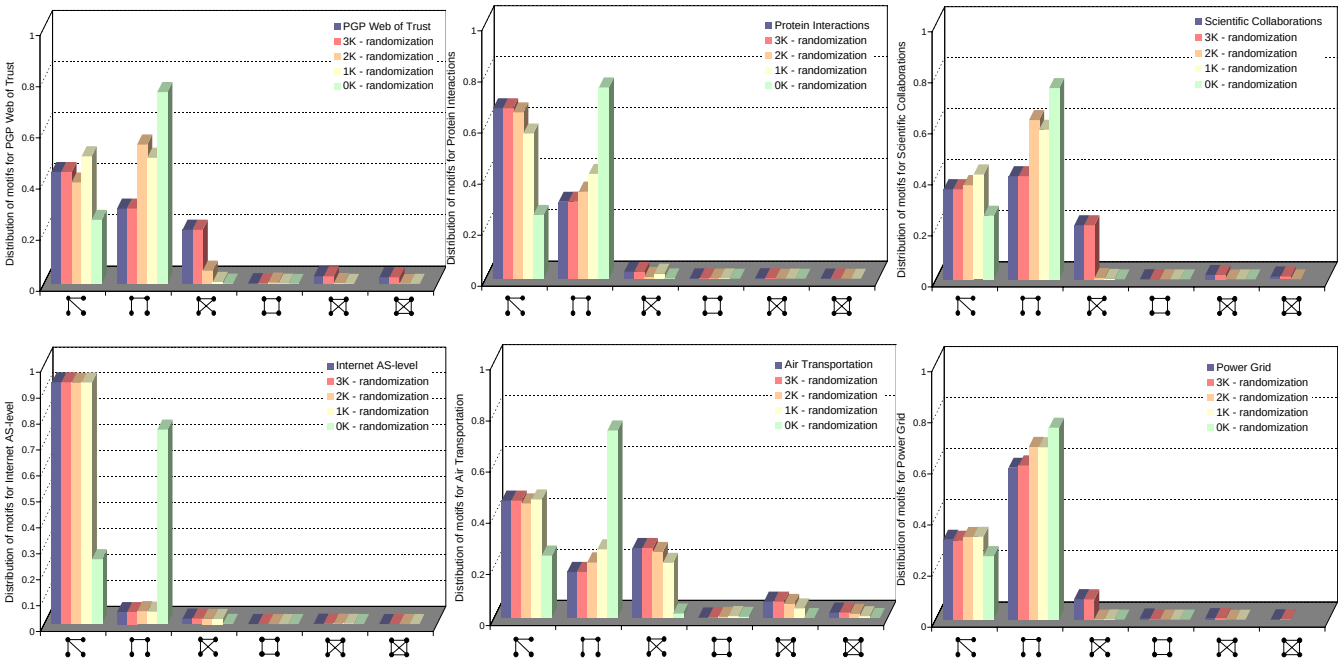


FIG. 9: The motif distributions in the real networks and their dK-randomizations.

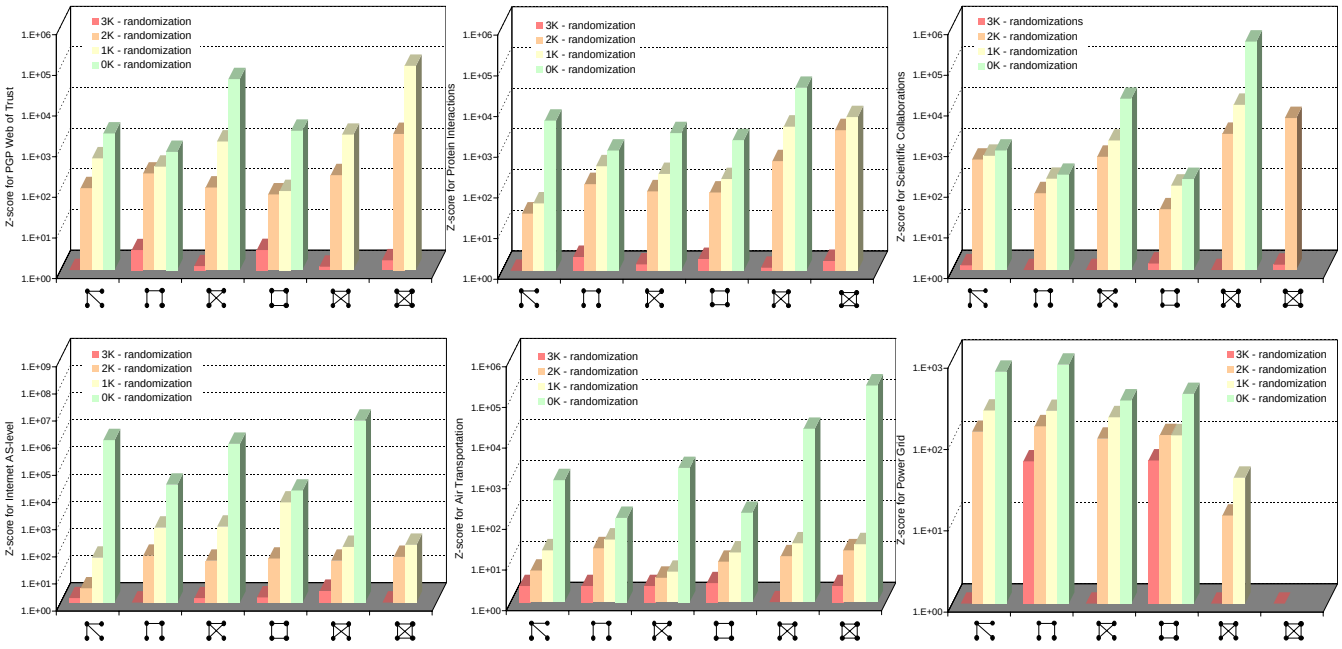


FIG. 10: The motif Z-scores in the real networks and their dK-randomizations.

where $N(N-1)/2$ is the total number of nodes pairs in the network.

To provide a clearer view of how close the distance distributions in dK-randomizations are to the real networks, we show in Fig. 12 some scalar summary statistics of the distance distribution as functions of d. These summary statistics are the average distance

$$\bar{h} = \sum_h h\delta(h), \tag{15}$$

and the standard deviation of the distance distribution $\delta(h)$. In addition we show in Fig. 12 the network diameter, i.e., the maximum hop-wise distance between nodes in the network, which is an extremal statistics of the dis-

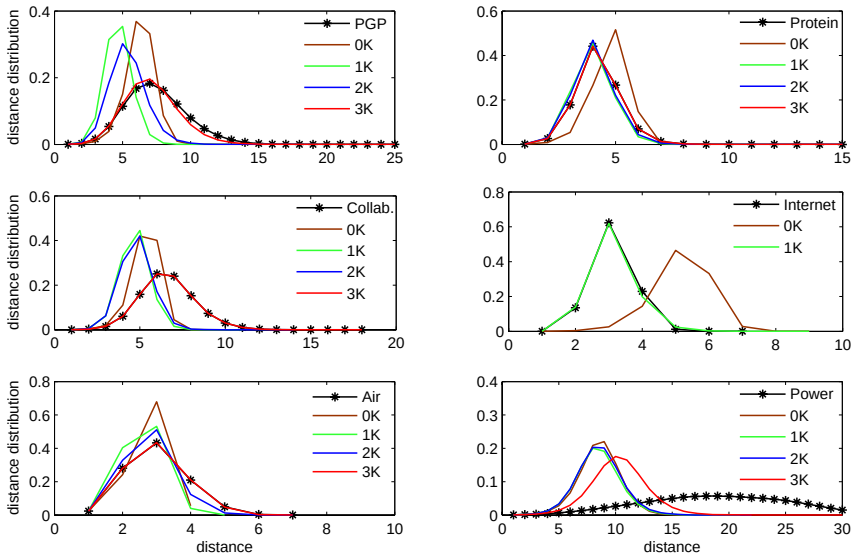


FIG. 11: The distance distribution in the real networks and their dK-randomizations.

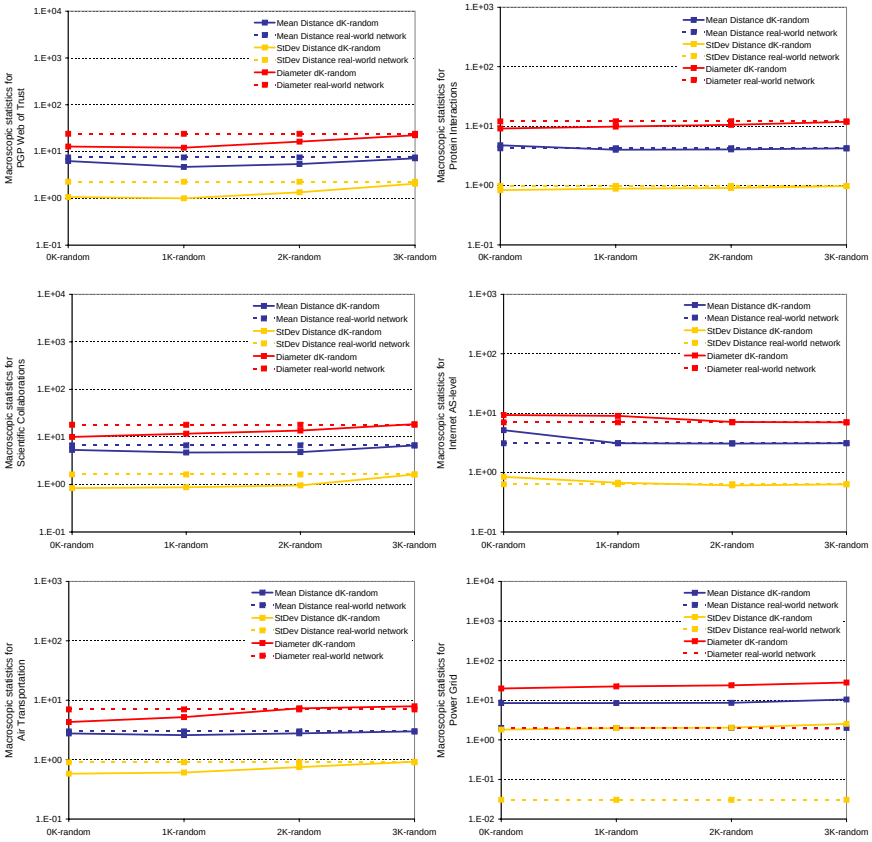


FIG. 12: The average distance, the standard deviation of the distance distribution, and the network diameter as functions of d for dK-randomisations of the real networks. The corresponding values for the real networks are shown by dashed lines.

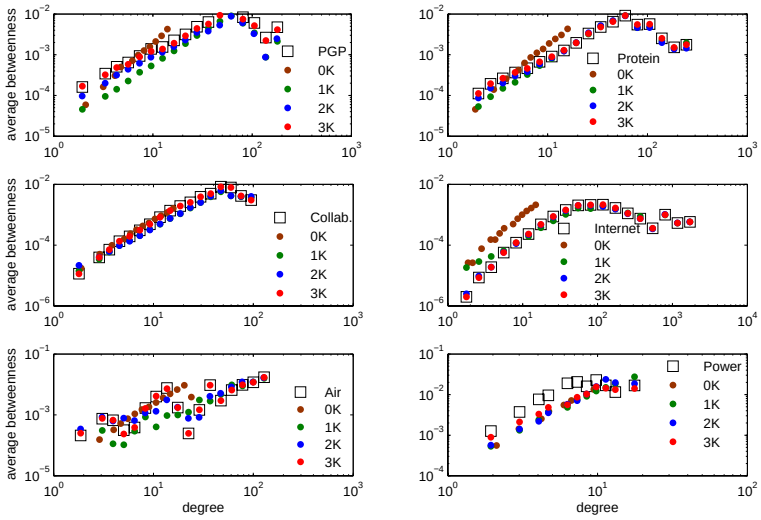


FIG. 13: The average betweenness of nodes of a given degree in the real networks and their dK-randomizations.

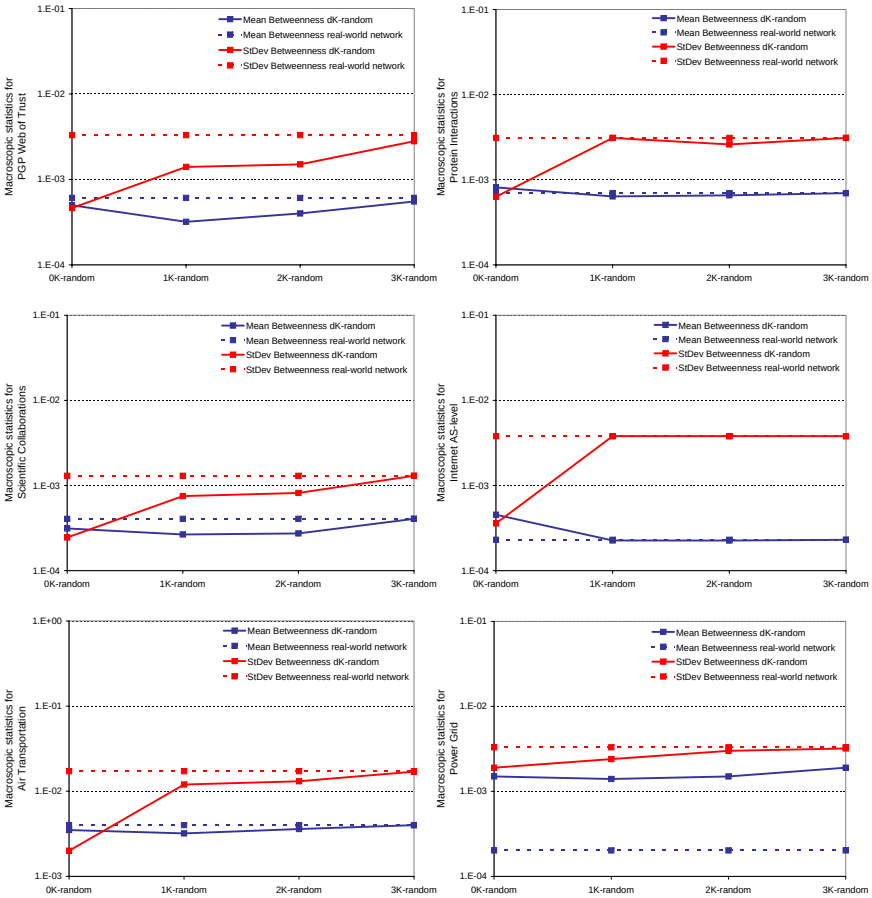


FIG. 14: The average betweenness and the standard deviation of the betweenness distribution as functions of d for dK-randomisations of the real networks. The corresponding values for the real networks are shown by dashed lines.

TABLE II: The scalar topological metrics of the real networks and the minimum value of d such that the network's dK-randomizations approximately preserve all the metrics.

Metrics	PGP	Collab.	Protein	Air	Internet	Power
$\bar{k}$	4.6	6.4	6.4	11.9	6.3	4.7
r	0.238	0.157	-0.137	-0.268	-0.236	-0.273
$\bar{c}$	0.27	0.65	0.09	0.62	0.46	0.68
$\bar{h}$	7.5	6.6	4.2	3.0	3.1	2.0
$\bar{b}$	$6 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$4 \cdot 10^{-3}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$
dK	$3K$	$3K$	$3K$	$2K$	$1K$	?

tance distribution.

Fig. 13 shows degree-dependent betweenness centrality $\bar{b}(k)$ in the real networks and their dK-randomizations. Betweenness $b(i)$ of node i is a measure of how "important" i is in terms of the number of shortest paths passing through it. Formally, if $\sigma_{st}(i)$ is the number of shortest paths between nodes $s \neq i$ and $t \neq i$ that pass through i, and σ_{st} is the total number of shortest paths between the two nodes $s \neq t$, then betweenness of i is

$$b(i) = \sum_{s,t} \frac{\sigma_{s,t}(i)}{\sigma_{s,t}}. \tag{16}$$

Averaging over all nodes of degree k, degree-dependent betweenness $\bar{b}(k)$ is

$$\bar{b}(k) = \sum_{i:\, \deg(i)=k} \frac{b(i)}{N(k)}. \tag{17}$$

We also compute the betweenness distribution, and show its average and standard deviation in Fig. 14.

We observer similar trends with respect to both distance and betweenness metrics. The power grid cannot be approximated even by its $3K$-randomization. The Internet lies at the other extreme: even $1K$-randomization does not disturb its global metrics too much. The air transportation network appears to come next, as its $2K$-randomizations resemble it closely. But all the networks other than the power grid are very similar to their $3K$-randomizations.

D. Scalar topological metrics and dK-randomness of real networks

To conclude this section we show in Table II the most important scalar topological metrics for the real networks. These metrics are coarse summary statistics of the more informative and detailed metrics that we have considered in this section. Specifically, these coarse summaries are:

- $\bar{k}$ is the average degree in the network, Eq. (2), which is both the $0K$-distribution and a summary statistics of the $1K$-distribution in the dK-series terminology;
- r is the assortativity coefficient,

$$r = \frac{\langle k \rangle^2 \sum_{kk'} kk' P(k,k') - \langle k^2 \rangle^2}{\langle k^3 \rangle \langle k \rangle - \langle k^2 \rangle^2} \tag{18}$$

  which is nothing but the Pearson correlation coefficient of the $2K$-distribution $P(k,k')$;
- $\bar{c}$ is the average clustering

$$\bar{c} = \sum_k \bar{c}(k) P(k), \tag{19}$$

  which is a coarse summary statistics of the $3K$-distribution;
- $\bar{h}$ is the average distance, Eq. (15), which is unrelated to dK-distributions;
- $\bar{b}$ is the average betweenness,

$$\bar{b} = \sum_k \bar{b}(k) P(k), \tag{20}$$

  unrelated to dK-distributions as well.

In Table II we also show the minimum value of d such the dK-randomization null model approximately reproduces the real network with respect to all the metrics above.

The observation that the power grid cannot be approximated even by its $3K$-randomization is instructive. It shows that there are networks for which there is no sufficiently small d capable of preserving the network structure upon dK-randomizing. In case of the power grid, the explanation why this network is not even $3K$-random may be related to the fact that it is carefully designed and fully controlled by human engineers. Informally, we can think of it as rather "non-random," designed, and thus bearing a number of constraints that the dK-distributions with low d cannot capture. Informally, the higher d required to approximately preserve the network structure upon dK-randomization, the less "random" the network is. The commonly referred explanation that the power grid is an "outlier" because it is spatially embedded, may be less relevant here because two other networks that we have considered (the Internet and air transportation) are also spatially embedded.

What is different between the power grid and the other considered networks is that the latter are self-evolving. They may be engineered to a certain degree, such as the Internet, but their global structure and evolution are not fully controlled by any single human or organization. In the Internet case, for example, the global network topology is a cumulative effect of independent decisions made by tens of thousands of separate organizations, roughly corresponding to Autonomous Systems, i.e., nodes of the Internet graph.

In that sense, self-evolving complex networks are "more random." However, why the level of their "randomness" is at $d \leqslant 3$ remains an open question.

TABLE III: dK-series vs. d-series

d	dK-statistics	d-statistics
0	$\bar{k}$	-
1	$N(k)$	N
2	$N(k,k')$	M
3	$N_\wedge(k,k',k'')$	W
	$N_\triangle(k,k',k'')$	T

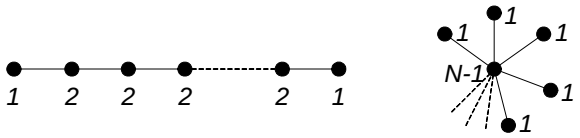
V. MOTIF-BASED SERIES VS. dK-SERIES

In this section we compare dK-series with the series based on motifs, and show that the latter cannot form a systematic basis for topology analysis.

The difference between dK-series and motif-series, which we can call d-series, is that the former is the series of distributions of d-sized subgraphs labeled with node degrees in a given network, while the d-series is the distributions of such subgraphs in which this degree information is ignored. This difference explains the mnemonic names for these two series: 'd' in 'dK' refers to the subgraph size, while 'K' signifies that they are labeled by node degrees—'K' is a standard notation for node degrees.

This difference between the dK-series and d-series is crucial. The dK-series are inclusive, in the sense that the $(d+1)K$-distribution contains the full information about the dK-distribution, plus some additional information, which is not true for d-series.

To see this, let us consider the first few elements of both series in Table III. In Section IV A we show explicitly how the $(d+1)K$-distributions define the dK-distribution for $d = 0, 1, 2$. The key observation is that the d-series does not have this property. The 0'th element of d-series is undefined. For $d = 1$ we have the number of subgraphs of size 1, which is just N, the number of nodes in the network. For $d = 2$, the corresponding statistics is M, the number of links, subgraphs of size 2. Clearly, M and N are independent statistics, and the former does not define the latter. For $d = 3$, the statistics are W and T, the total number of wedges and triangles, subgraphs of size 3, in the network. These do not define the previous element M either. Indeed, consider the following two networks of size N—the chain and the star:



There are no triangles in either network, $T = 0$. In the chain network, the number of wedges is $W = N - 2$, and in the star $W = (N-1)(N-2)/2$. We see that even though W ($d = 3$) scales completely differently with N in the two networks, the number of edges $M = N - 1$ ($d = 2$) is the same.

In summary, d-series is not inclusive. For each d, the corresponding element of the series reflects a differen kind of statistical information about the network topology, unrelated or only loosely related to the information conveyed by the preceding elements. At the same time, similar to dK-series, the d-series is also converging since at $d = N$ it specifies the whole network topology. However, this convergence is much slower that in the dK-series case. In the two networks considered above, for example, neither $W = N - 2$, $T = 0$ nor $W = (N-1)(N-2)/2$, $T = 0$, fix the network topology as there are many non-isomorphic graphs with the same (W, T) counts, whereas the $3K$-distributions $N_\wedge(1,2,2) = 2$, $N_\wedge(2,2,2) = N-4$ and $N_\wedge(1,N-1,1) = (N-1)(N-2)/2$ define the chain and star topologies exactly.

The node degrees thus provide necessary information about subgraph locations in the original network, which improves convergence, and makes the dK-series basis inclusive and systematic.

Acknowledgments

We thank Alex Arenas and Alessandro Vespignani for useful comments and discussions, and Connie Lyu and Bradley Huffaker for their help with Figs. 1,5. This work was supported in part by DGES grant FIS2007-66485-C02-02, by NSF CNS-0434996 and CNS-0722070, by DHS N66001-08-C-2029, and by Cisco Systems.

[1] R. Milo, S. Shen-Orr, S. Itzkovic, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[2] U. Alon, Nat Rev Genet **8**, 450 (2007).
[3] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall/CRC, Boca Raton, 2006).
[4] N. Rosenfeld, M. Elowitz, and U. Alon, J Mol Biol **323**, 785 (2002).
[5] S. Mangana, S. Itzkovitz, A. Zaslaver, and U. Alon, J Mol Biol **356**, 1073 (2006).
[6] J. Knabe, C. Nehaniva, and M. Schilstra, Biosystems **94**, 68 (2008).
[7] P. Ingram, M. Stumpf, and J. Stark, BMC Genomics **7**, 108 (2006).
[8] O. Cordero and P. Hogeweg, Mol Biol Evol **23**, 1931 (2006).
[9] P. Kuo, W. Banzhaf, and A. Leier, Biosystems **85**, 177 (2006).
[10] A. Mazurie, S. Bottani, and M. Vergassola, Genome Biol **6**, R35 (2005).
[11] S. Sakata, Y. Komatsu, and T. Yamamori, Neurosci Res **51**, 309 (2005).
[12] A. Vázquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabási, Proc Natl Acad Sci USA **101**,

17940 (2004).
[13] Y. Artzy-Randrup, S. Fleishman, N. Ben-Tal, and L. Stone, Science **305**, 1107 (2004).
[14] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, Comput Commun Rev **36**, 135 (2006).
[15] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys Rev E **70**, 056122 (2004).
[16] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, in *Advances in Neural Information Processing Systems 18*, edited by Y. Weiss, B. Schölkopf, and J. Platt (MIT Press, Boston, 2006), pp. 41–50.
[17] S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[18] S. Maslov, K. Sneppen, and U. Alon, *Handbook of Graphs and Networks* (Wiley-VCH, Berlin, 2003), chap. Correlations Profiles and Motifs in Complex Networks.
[19] X. Dimitropoulos, D. Krioukov, G. Riley, and A. Vahdat, ACM Transactions on Modeling and Computer Simulation (to appear) (2009), `arXiv:0708.3879`.
[20] J. Duch and A. Arenas, Phys Rev E **72**, 027104 (2005).
[21] L. Freeman, Sociometry **40**, 35 (1977).
[22] M. Boguñá, D. Krioukov, and kc claffy, Nature Physics **5**, 74 (2009).
[23] M. Á. Serrano, D. Krioukov, and M. Boguñá, Phys Rev Lett **100**, 078701 (2008).
[24] T. V. den Bulcke, K. V. Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. D. Moor, and K. Marchal, BMC Bioinformatics **7** (2006).
[25] J. Knabe, C. Nehaniva, and M. Schilstra, Artif Life **14**, 135148 (2008).
[26] S. Roy, M. Werner-Washburne, and T. Lane, Bioinformatics **24**, 13181320 (2008).
[27] B. M. Waxman, IEEE J Sel Area Comm **6**, 1617 (1988).
[28] E. Zegura, K. Calvert, and S. Bhattacharjee, in *Proc INFOCOM* (1996), vol. 2, pp. 594–602.
[29] A. Medina, A. Lakhina, I. Matta, and J. Byers, in *Proc MASCOTS* (2001), pp. 346–353.
[30] J. Winick and S. Jamin, Technical Report UM-CSE-TR-456-02, University of Michigan (2002).
[31] M. E. J. Newman, Proc Natl Acad Sci USA **98**, 404 (2001).
[32] V. Colizza, A. Flammini, M. Á. Serrano, and A. Vespignani, Nat Phys **2**, 110 (2006).
[33] V. Colizza, R. Pastor-Satorras, and A. Vespignani, Nat Phys **3**, 276 (2007).
[34] P. Mahadevan, D. Krioukov, M. Fomenkov, B. Huffaker, X. Dimitropoulos, kc claffy, and A. Vahdat, Comput Commun Rev **36**, 17 (2006).
[35] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[36] M. Á. Serrano and M. Boguñá, Phys Rev E **74**, 056114 (2006).
[37] M. A. Serrano and M. Boguñá, Phys Rev E **74**, 056115 (2006).
[38] See [36, 37] for an alternative formalism involving three point correlations