Percolation and the Effective Structure of Complex Networks

Antoine Allard[1,2] and Laurent Hébert-Dufresne[3,1]

[1]*Département de physique, de génie physique et d'optique, Université Laval, Québec, Québec, Canada G1V 0A6*

[2]*Centre interdisciplinaire de modélisation mathématique, Université Laval, Québec, Québec, Canada G1V 0A6*

[3]*Department of Computer Science and Vermont Complex Systems Center, University of Vermont, Burlington, Vermont 05405, USA*

(Received 9 July 2018; revised manuscript received 10 November 2018; published 5 February 2019)

Analytical approaches to model the structure of complex networks can be distinguished into two groups according to whether they consider an intensive (e.g., fixed degree sequence and random otherwise) or an extensive (e.g., adjacency matrix) description of the network structure. While extensive approaches—such as the state-of-the-art message passing approximation—typically yield more accurate predictions, intensive approaches provide crucial insights on the role played by any given structural property in the outcome of dynamical processes. Here we introduce an intensive description that yields almost identical predictions to the ones obtained with the message passing approximation using bond percolation as a benchmark. Our approach distinguishes nodes according to two simple statistics: their degree and their position in the core-periphery organization of the network. Our near-exact predictions highlight how accurately capturing the long-range correlations in network structures allows easy and effective compression of real complex network data.

DOI: 10.1103/PhysRevX.9.011023

Subject Areas: Complex Systems, Interdisciplinary Physics, Statistical Physics

I. INTRODUCTION

The structure of real complex networks lies somewhere between order and randomness [1–3], with the consequence that it cannot typically be fully characterized by a concise set of synthesizing observables. This *irreductibility* explains why most theoretical approaches to model complex networks are inspired by statistical physics in that they consider ensembles of networks constrained by the values of observables (e.g., density of links, degree-degree correlations, clustering coefficient, degree or motif distribution) and otherwise organized randomly. These approaches have three notable advantages. First, they usually yield analytical treatment. Second, they are *intensive* in network size, meaning that their complexity scales with the support of the observables (i.e., sublinearly with the numbers of nodes and links). Third, they provide null models, of which many have led to the identification of fundamental properties characterizing the structure of real complex networks [4,5].

Despite important leaps forward in recent years, these approaches still fail to capture enough information to systematically provide accurate quantitative predictions of most dynamical processes on real complex networks. The reason for this shortcoming is that the properties from which the ensembles are constructed are not constraining enough; the ensembles are "too large," such that the original real networks are exceptions, rather than typical instances, in the ensembles. As a result, the current state-of-the-art approach—the so-called message passing approximation (MPA) [6]—requires the whole structure to be specified as an input (i.e., the adjacency matrix, or a transformation thereof). This method is interesting because it is mathematically principled, meaning that it yields *exact* results on trees, and offers inexact, albeit generally good, predictions on networks containing loops (i.e., most real complex networks) [7].

However, by considering the whole structure of networks and thereby considering every link on equal footing, the accuracy of the MPA comes at a significant computational and conceptual cost. First, its time and space complexity are *extensive* in the number of links and therefore in the size of the network. Second, and most importantly, it does not provide any insight on the role played by any given structural property in the outcome of a dynamical process. With the MPA, getting good predictions comes at the expense of understanding what led to that outcome.

In this paper, we bridge the gap between intensive and extensive approaches to the mathematical modeling of

networks using the classic bond percolation problem as a benchmark. We introduce a random network ensemble that relies solely on an *intensive* description of the network structure that, nevertheless, yields predictions that are comparable to the ones from the MPA. This ensemble is based on the onion decomposition (OD), a refined k-core decomposition [8]. Critically, the OD can be translated into local connection rules allowing an exact mathematical treatment using probability generating functions (PGFs) in the limit of large network size. This approach leads to exact predictions on trees like the MPA and highlights the critical contribution of the OD to an accurate effective mathematical description of real complex networks. The significant step forward in accuracy provided by this new network description is confirmed by comparing its predictions with the ones of the classic configuration model, its degree-correlated variant and the MPA using a collection of 111 real network datasets. We included network data from several scientific domains, from the structure of social interactions among people, to food webs, power grids, road networks, and connectomes, making our study one of the most extensive comparisons of network models to date.

II. RESULTS AND DISCUSSIONS

To obtain useful analytical results, most models of complex networks must rely on some variation of the treelike approximation, which assumes that complex networks have essentially no loops beyond some local structure of interest [9,10]. This approximation allows an elegant mathematical treatment which typically works surprisingly well, although the vast majority of real complex networks are not treelike [11]. In the case of the MPA, the treelike approximation implies that a lot of information given to the model is thrown away due to loops being included in the input information (i.e., the adjacency matrix) only to be mathematically ignored.

To take advantage of the mathematical tools developed for treelike approximations, we propose to limit the information given to the model by compressing complex networks into an *effective tree*. While there are many approaches to network compression, most of them do not encode the network structure in a way that lends itself to mathematical treatment beyond the calculation of a few observables (e.g., degree distribution or clustering coefficients). The limitations of these approaches—which are mostly based on the concepts of motifs, local clustering, modules, or latent metric space—are discussed in Appendix E. Instead, we rely on a peeling process, which iteratively removes leaves (i.e., nodes located at the periphery) to unveil the effective tree of a network.

Taking this information into account, we then focus on predicting the outcome of bond percolation on complex networks: a canonical problem of network science analogous to many applied problems such as disease propagation or network resilience [12]. Given a network structure, this simple stochastic process consists in the occupation of each original link with probability p. We aim to predict the size of the largest connected component composed of occupied links S, as well as the percolation threshold p_c, above which that component corresponds to a macroscopic fraction of the network. The outcome of percolation depends on structural properties at all scales, thus making it a good benchmark for theoretical network models. Note that while real finite networks do not undergo phase transitions *per se*, their connectivity displays a behavior akin to a phase transition when varying p. We aim to predict this behavior using mathematical models in which phase transitions do take place.

A. Onion decomposition

The k-core decomposition is a well-known network metric that identifies a set of nested maximal subnetworks—the k-cores—in which each node shares at least k links with the other nodes [13,14]. A node belonging to the k-core but not to the $(k+1)$-core is said to be of *coreness* k and to be part of the k-shell. Nodes with a high coreness are generally seen as more central whereas nodes with low corenesses are seen as being part of the periphery of the network. The onion decomposition refines the k-core decomposition by assigning a layer l to each node to further indicate its position within its shell (e.g., in the middle of the layer or at its boundary). The OD therefore unveils the internal organization of each centrality shell and, unlike the original k-core decomposition, can be used to assess whether the structure of a core is more similar to a tree or to a lattice, among other things [8].

The OD of a given network structure is obtained via the following pruning process (see Fig. 1). First we remove every node with the smallest degree $k_{\min}$; the coreness of these

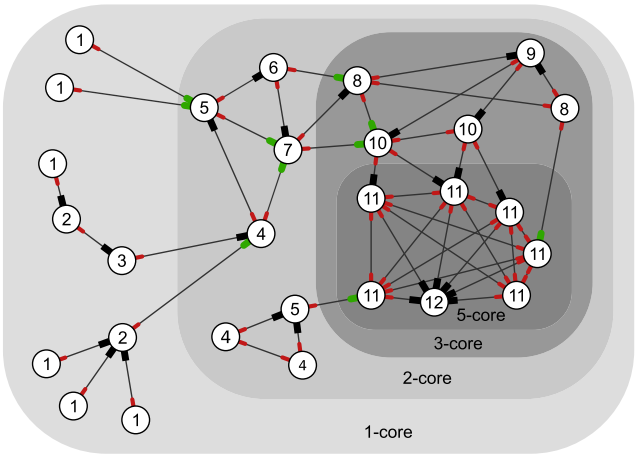


FIG. 1. Illustration of the onion decomposition (OD) of a simple network. The number of the layer to which each node belongs is indicated and the different k-cores are shown using increasingly darker background shades. The color of each stub according to the LCCM is also shown (different sizes and shapes are used for a clearer contrast between the three types of stubs).

nodes is equal to $k_{\min}$ and they are part of the first layer ($l = 1$). Removing these nodes may yield nodes whose remaining degree is now equal to or smaller than $k_{\min}$; these nodes must also be removed, have a coreness of $k_{\min}$ as well, but are part of the second layer ($l = 2$). If removing nodes of the second layer yields new nodes with a remaining degree equal to or lower than $k_{\min}$, they will be part of the third layer ($l = 3$), will have a coreness of $k_{\min}$, and will also be removed. This process is repeated until no new nodes with a remaining degree equal to or lower than $k_{\min}$ are left. We then update the value of $k_{\min}$ to reflect the lowest remaining degree and repeat this whole process until every node has been assigned a coreness and a layer. The layer number keeps increasing such that each layer corresponds to a unique coreness.

An efficient implementation of this procedure has a runtime complexity of $\mathcal{O}(L \log N)$, where L and N are, respectively, the number of links and nodes, which implies that the OD can be quickly obtained for virtually any real complex network [8]. Most importantly, nodes belonging to a same layer are *topologically similar* with regard to the mesoscale centrality organization of the network. Because the layer of a node is only weakly related to its degree (i.e., the coreness of a node provides a lower bound to its degree), the pair layer-degree can therefore be used to indicate how well a node is connected, but also to indicate its "topological position" in the network. It therefore allows us to discriminate central nodes from peripheral ones which, based on their degree alone, would have otherwise been deemed identical.

B. Effective random network ensemble: The layered and correlated configuration model (LCCM)

From the pruning process described above, it can be concluded that a node of coreness c belonging to the lth layer is in one of two scenarios. (1) It must have *exactly* c links to nodes in layers $l' \geq l$ if layer l is the first layer of the c shell (i.e., nodes in layer $l-1$ belong to the c' shell with $c' < c$). (2) Otherwise, if it is not in the first layer of its c shell, it must have *at least* $c+1$ links to nodes of layers $l' \geq l-1$ and *at most* c links to nodes of layers $l' \geq l$. The distinction between the two scenarios is that nodes not in the first layer of their shell require at least one link to the previous layer to anchor them to their own layer. Also, the common feature of these scenarios is that a node of coreness c needs at least c links with nodes of equal or greater coreness.

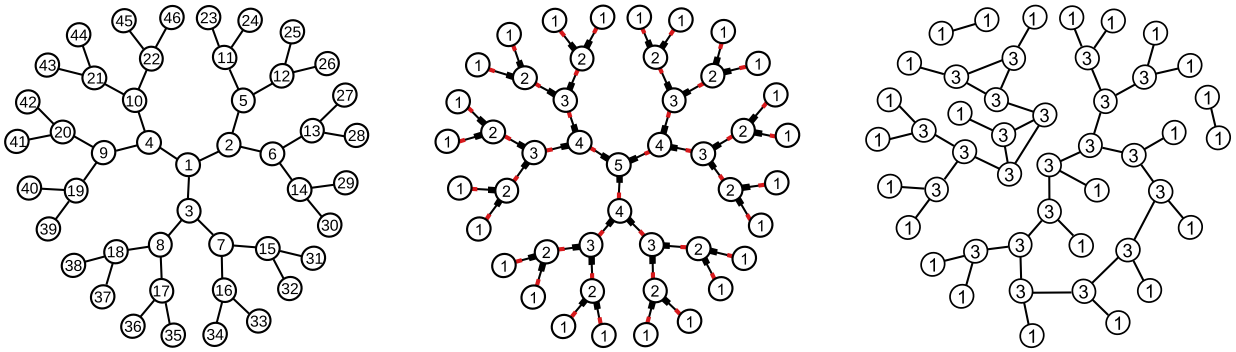
By rewiring the links of a given network using a degree-preserving procedure [15,16] while ensuring that the aforementioned rules are respected at all time, it is possible to explore the ensemble of all possible single networks with the same fixed layer-degree sequence [i.e., the sequence of every pair (l, k) in the original network]. Exactly preserving the layers—and thus the coreness of every node—is of critical significance since previous rewiring approaches could only approximately preserve the k-core decomposition [17].

Additionally, the pair layer-degree assigned to each node can be used to enforce two-point correlations [i.e., the (layer-degree)–(layer-degree) correlations], thus reducing the size of a random network ensemble. This correlated ensemble can be explored via a double link swap Markov chain method preserving both the layer-degree sequence and the number of links within and between every node class (i.e., nodes with the same layer-degree). One way to implement this method is by first choosing one link at random (e.g., joining nodes A and B) and then choosing another link at random (e.g., joining nodes C and D) among the links that are attached to at least one node whose layer-degree pair is the same at one of the two nodes connected by the first link (e.g., A and C have the same layer-degree) [18]. The two links are then swapped (e.g., A becomes connected to D and B to C) if no self-link or multilink would be created. Doing so ensures that both the degree sequence and the two-point correlations are preserved at all time. We call the layered and correlated configuration model (LCCM) the ensemble of maximally random networks with a given joint layer-degree sequence and (layer-degree)–(layer-degree) correlations.

Since it preserves both the degree sequence and the degree-degree correlations, the LCCM is a subset of two commonly used random network ensembles defined by the configuration model (CM) [19] and the correlated configuration model (CMM) [20]; the latter being known for its fair accuracy in many applications [11]. The LCCM, however, distinguishes itself from these models (and other variants) by enforcing a mesoscopic organization via the layers of the OD. This feature has the critical advantage of making the LCCM a mathematically principled approach in the sense that it exactly preserves the structure of a wide variety of trees (see Fig. 2). This is due to the fact that the network ensemble defined by the LCCM corresponds to the automorphisms of these trees. In other words, in such cases the only link swaps allowed by the connection rules of the LCCM yield a mere relabeling of the nodes without modifying the overall structure whatsoever. As we show below, this mesoscopic information accounts for a significant portion of the missing gap between the predictions of the intensive configuration models and the extensive, current state-of-the-art MPA.

C. Percolation on the LCCM

We adapt the approach of Ref. [10] to solve bond or site percolation on the LCCM in the limit of large network size. This approach requires the specification of (1) the classes of nodes, which here correspond to the distinct pairs layer-degree noted (l, k) and (2) the colors of stubs (i.e., half-links), which in the LCCM are identified based on the layer l' of the neighboring node. More precisely, from the connection rules stated in the previous section, the LCCM requires us to keep track of the number of links that each node in each layer l shares with nodes (i) in layers $l' \geq l$, (ii) in layer $l' = l-1$, and (iii) in layers $l' < l-1$. We identify the corresponding half-links as red, black, and



(a) Message passing approximation (b) Layered and correlated configuration model (c) Configuration model

FIG. 2. Compression of a perfect tree with different network models. (a) The message passing approximation assigns a unique ID to every node and preserves the full structure of the tree. (b) The layered and correlated configuration model assigns an ID to every node corresponding to its degree and its position in the core-periphery structure of the network. Degrees are not shown to lighten the presentation. Stubs are colored according to the layer to which they point: red if they point to more central layers and black if they point to the previous layer (different sizes are used for a clearer contrast between black and red stubs). There are no green stubs in this example. (c) The configuration model assigns an ID to every node according to its degree before randomly connecting them, thereby destroying the mesoscopic and macroscopic structure of the original network. The correlated configuration model fixes the number of links between different degree classes, and would therefore prohibit components formed by two nodes with degree 1, but would otherwise be very similar to the configuration model shown here.

green stubs, respectively. For instance, a link between nodes in layers 3 and 5 consists in a red stub stemming out of the node in layer 3 paired with a green stub belonging to the node in layer 5. Note that a link between two given layers can only consist in a unique pair of stub colors, and the only allowed combinations are red-red, red-black, and red-green.

From the link correlation matrix $\mathbf{L}$, whose entries specify the fraction of links within and between every class of nodes, we can derive the function (see Appendix B)

$$\varphi_{lk}(\boldsymbol{x}) = \sum_{k^r k^b k^g} P_{lk}(k^r, k^b, k^g)[x_{lk}^r]^{k^r}[x_{lk}^b]^{k^b}[x_{lk}^g]^{k^g} \quad (1)$$

generating the probability $P_{lk}(k^r, k^b, k^g)$ that a node in class (l, k) has k^r red stubs, k^b black stubs, and k^g green stubs, given the connection rules of the LCCM. From the same link correlation matrix, we can also derive the functions (see Appendix C)

$$\gamma_{lk}^{\alpha}(\boldsymbol{x}) = \sum_{l'k'} \sum_{\alpha' \in \{r,b,g\}} Q_{lk}^{\alpha}(l', k', \alpha') x_{l'k'}^{\alpha'}, \quad (2)$$

for every $\alpha \in \{r, b, g\}$, generating the probability $Q_{lk}^{\alpha}(l', k', \alpha')$ that a stub of color α stemming from a node of class (l, k) is attached to a stub of color α' belonging to a node in class (l', k'). Combining these two functions yields the PGF generating the distribution of the number of nodes of each class that are neighbors of a randomly chosen node of class (l, k):

$$g_{lk}(\boldsymbol{x}) = \varphi_{lk}(\boldsymbol{\gamma}(\boldsymbol{x})). \quad (3)$$

Note that this PGF also includes the colors of the stub through which these neighbors are connected to the node of class (l, k). Similarly, the number of such nodes that can be reached from a node of class (l, k) that has itself been reached by one of its stubs of color α is

$$f_{lk}^{\alpha}(\boldsymbol{x}) = \frac{1}{\langle k^{\alpha} \rangle_{lk}} \frac{\partial \varphi_{lk}(\boldsymbol{x}')}{\partial x_{lk}'^{\alpha}}\bigg|_{\boldsymbol{x}'=\boldsymbol{\gamma}(\boldsymbol{x})}, \quad (4)$$

where $\langle k^{\alpha} \rangle_{lk} = \{[\partial \varphi_{lk}(\mathbf{1})]/\partial x_{lk}^{\alpha}\}$ is the average number of stubs of color α nodes of class (l, k) have.

To compute the size of the extensive component, we assume that the networks in the ensemble are locally treelike, which occurs in the limit of large network size or when the detailed structure of matrix $\mathbf{L}$ permits only exact trees (i.e., when loops are structurally impossible). We define a_{lk}^{α} as the probability that attempting to reach a node in class (l, k) by one of its stubs of color α does not eventually lead to the extensive component. Noting p the probability that links are occupied, the probabilities $\{a_{lk}^{\alpha}\}$ are the solution of

$$a_{lk}^{\alpha} = 1 - p + p f_{lk}^{\alpha}(\boldsymbol{a}), \quad (5)$$

for all l, k, and α. This last expression encodes the simple self-consistent argument that attempting to reach the node will not lead to the extensive component if (1) the link is unoccupied, which occurs with probability $1 - p$, or if (2) the link is occupied, with probability p, but the attempts to reach the other neighbors of the node that has just been reached will all fail, which occurs with probability $f_{lk}^{\alpha}(\boldsymbol{a})$. Note that this argument relies on the assumption that the states of these neighbors are independent, which is true for a treelike structure. Having solved Eq. (5), the relative size of the extensive component S is then given by the probability that a randomly chosen node is found in S,

$$S = 1 - \sum_{lk} P(l,k) g_{lk}(\boldsymbol{a}), \tag{6}$$

where $P(l,k)$ is the fraction of nodes in class (l,k) which can be extracted from the link correlation matrix $\mathbf{L}$ (see Appendix A). Note that Eq. (5) remains valid in the case of site percolation—where nodes, instead of links, are occupied with probability p—since we assume the networks of the ensemble to be locally treelike. Equation (6) solely needs adjustment and becomes $S^{\text{site}} = pS$ to account for the fact that only a fraction p of nodes are occupied and can therefore be part of the extensive component [10,21]. Note also that the percolation threshold p_c is the value of p at which $\boldsymbol{a} = \mathbf{1}$ becomes an unstable solution of Eq. (5) (see Appendix D), which corresponds to the emergence of the extensive component.

D. Effective treelike structure

Because the LCCM is a special case of both the CM and CCM, the cardinality of the ensemble it generates should always be smaller than the cardinality of the ensembles generated by the CM and by the CCM. Consequently, if the mesoscale structural information provided by the layers l is of any significance, we expect the predictions of the LCCM to be the closest to the ones obtained with the MPA. Figure 3 confirms this observation using 111 real network datasets (see Appendix F for details on the datasets). In fact, our results demonstrate that identifying nodes using the layer in the OD alongside their degree does not merely improve the predictions, it drastically changes their nature, making them qualitatively very similar to the ones of the MPA when not strikingly quantitatively identical. Figure 4 further supports this conclusion with four representative datasets for which the gain in accuracy is the most manifest. Indeed, the LCCM reproduces the general shape of the curves, has the same number of inflection points, and always predicts a connected network when all links are occupied (i.e., S must be 1 at $p = 1$ since we considered the largest connected components of every datasets). Interestingly, only the LCCM and the MPA are able to capture the mesoscopic core-periphery and/or modular structures that were numerically shown to lead to smeared (or double) phase transitions [22] such as the one observed on the protein-protein interaction network.

Perhaps most importantly, the LCCM approximates to high accuracy the percolation threshold predicted by the MPA, as seen in Fig. 3(a), with a relative error of less than 1.5% for 75% of the 111 network datasets considered. Additionally, Fig. 3(b) shows the expected error on the size of the extensive component averaged over the entire range of occupation probability p. When using the LCCM to compress the network structure, we find that the error, relative to the MPA, is at most of the order of 10^{-3} for 75% of the datasets considered; an improvement of at least 1 order of magnitude from existing approaches. Altogether, these results indicate that categorizing nodes with the classes (l,k) captures critical features of the local and mesoscopic treelike organization of many real complex networks, thus offering an intensive effective description of their structure.

It is important to mention that the MPA's predictions were considered here not as the ground truth of bond percolation but rather as the best analytical predictions available. While we acknowledge that there exist approximations of the MPA taking into account clustering that work quite well for site percolation [27], the accuracy of these approximations does not translate well to bond percolation due to the over-counting redundant paths. Similar approximations exist for bond percolation but based on nonunique decompositions of networks into triangles [28]. More importantly, the

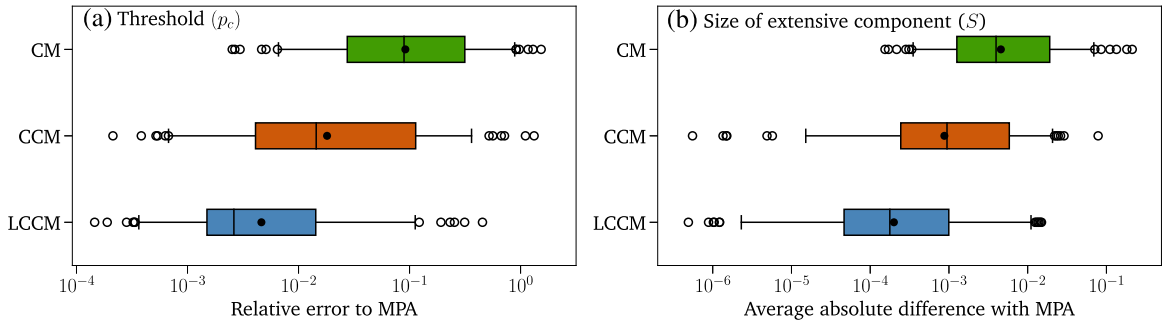


FIG. 3. Predictions of the intensive models (CM, CCM, and LCCM) compared to the predictions of the extensive MPA for a collection of 111 real network datasets from different scientific domains, including, but not limited to, social networks (where nodes are often people and links are interactions), biological networks (e.g., protein-protein interaction networks), connectomes (e.g., axonal connections between brain regions), food webs, infrastructure networks (e.g., power grids), and transportation networks (e.g., flights between airports). Further details are provided in Appendix F. The whiskers cover the range between the fifth and the 95th percentiles, the black dots indicate the mean, and the outliers' data points are shown with a circle. Each box indicates the first, second, and third quartiles, as usual. (a) Relative error of the percolation threshold defined as $|p_c^{\text{model}} - p_c^{\text{MPA}}|/p_c^{\text{MPA}}$. The calculation of p_c^{LCCM} is detailed in Appendix D. (b) Area of the region bounded by the curves S^{model} and S^{MPA} computed as $\int_0^1 |S^{\text{model}} - S^{\text{MPA}}| dp$. References [6,19,20] provide the methods to compute p_c^{model} and S^{model} for the CM, the CCM, and the MPA.

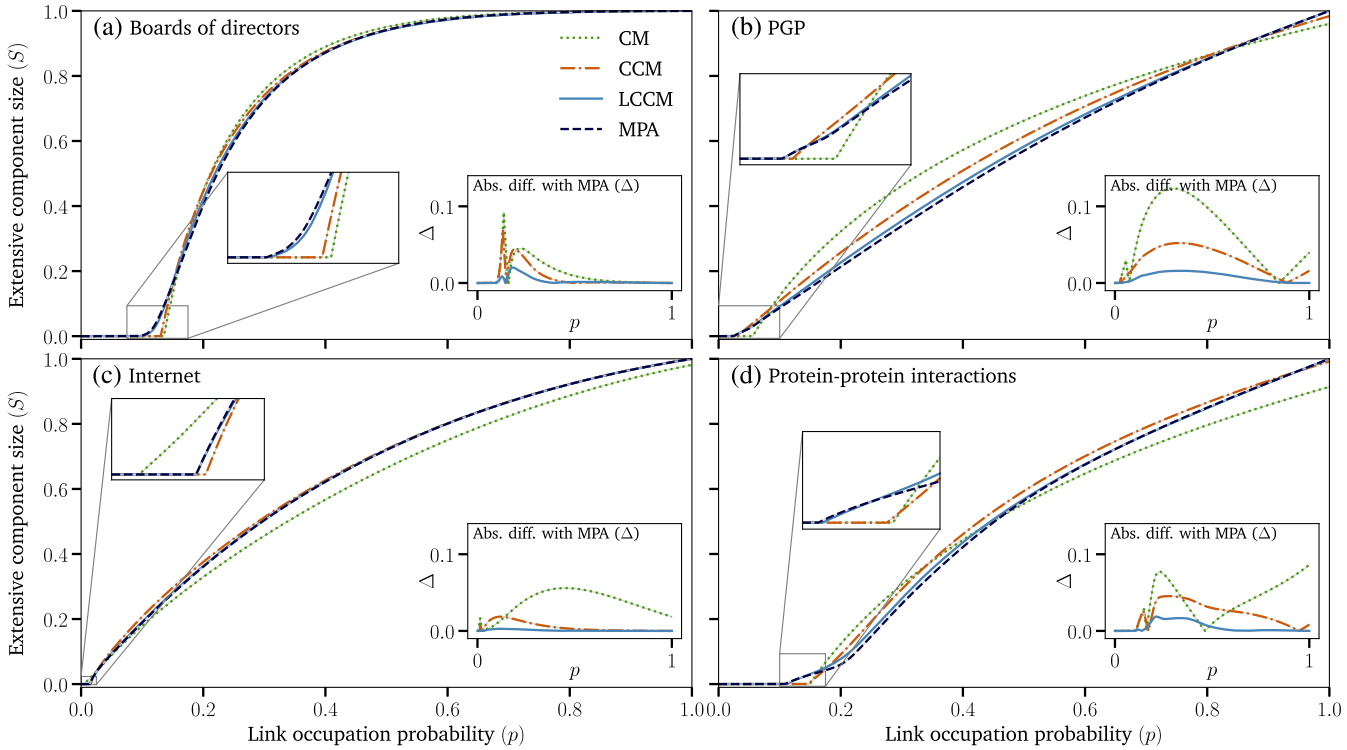


FIG. 4. Relative size of the extensive component predicted by the LCCM with the CM, the CCM, and the MPA for four representative real network datasets. (a) One-mode projection of a bipartite network of Norwegian boards of directors [23]. (b) PGP web of trust [24]. (c) A subset of the Internet at the autonomous level [25]. (d) Protein-protein interaction network of *Homo sapiens* [26]. The insets show the absolute value of the difference Δ between the MPA and the CM, the CCM and the LCCM as a function of p, as well as an enlargement of the region around the percolation threshold. The largest connected component was used for all datasets.

"original" formulation of the MPA and the LCCM both rely on the same simplifying assumptions of infinite network size and treelike local structure, which allowed us to compare the predictions of both approaches on the same footing. Doing so allows us to clearly demonstrate that the LCCM offers a way to substantially compress the information used indiscriminately by the MPA while still providing predictions with a similar accuracy. That being said, approximations introduced in the MPA to account for clustering could also potentially be used within the LCCM.

III. SUMMARY AND FUTURE WORK

We introduced a random network ensemble that relies solely on an *intensive* description of the network structure that yields predictions for percolation that are either essentially quantitatively identical—or at least strikingly qualitatively similar—to the ones obtained with the state-of-the-art MPA. This ensemble assigns two structural features to each node—its degree k (local) and its position l in the onion decomposition of the network (mesoscale)—and creates links according to simple, local connection rules that exactly preserve these two features. This ensemble lends itself to exact analytical calculations using probability generating functions in the limit of large network size. This model is mathematically principled, meaning that it leads to exact predictions on trees, like the MPA, but unlike other intensive approaches, such as the classic configuration model and existing variants. The accuracy of the predictions of the LCCM shows that the OD easily captures important features of the mesoscale structural organization of many real complex networks, which is at the root of the accuracy of the current state-of-the-art message passing approximation. Given that the vast majority of network models rely on local, pairwise connection rules, the significant gain in accuracy demonstrated here for percolation is therefore readily accessible to other modeling approaches. We consequently strongly argue that this information should be included in the future generations of models of complex networks.

A. Recurrent state dynamics

One salient advantage of effective ensembles of random networks is that they can be leveraged to describe complex dynamical processes on networks, whereas generalizing the MPA to complex recurrent state dynamics is nontrivial [29,30]. With that in mind, the exact mathematical description of the LCCM can easily be included in a system of ordinary differential equations (ODEs) describing, e.g., the susceptible-infectious-susceptible (SIS) model of disease spread. Instead of directly following all types of nodes

generated by Eq. (1), we can average all nodes with the same layer and degree and simply follow the fractions $I_{l,k}(t)$ that are infectious at time t. The model specifies that these infectious nodes recover at rate α and infect susceptible neighbors at rate β, such that we can write:

$$\begin{aligned}\frac{d}{dt}I_{l,k}(t) = &-\alpha I_{l,k}(t) + \beta[P(l,k) - I_{l,k}(t)] \\ &\times \sum_{k^r,k^g,k^b} P_{l,k}(k^r,k^g,k^b) \\ &\times \left[k^r \frac{\sum_{l'\geq l,k'} L_{lk,l'k'}(1+\delta_{ll'}\delta_{kk'})I_{lk,l'k'}}{\sum_{l'\geq l,k'} L_{lk,l'k'}(1+\delta_{ll'}\delta_{kk'})}\right. \\ &+ k^g \frac{\sum_{l'<l-1,k'} L_{lk,l'k'} I_{lk,l'k'}}{\sum_{l'<l-1,k'} L_{lk,l'k'}} \\ &\left.+ k^b \frac{\sum_{k'} L_{lk,(l-1)k'} I_{lk,(l-1)k'}}{\sum_{k'} L_{lk,(l-1)k'}}\right].\end{aligned}$$

This system of ODEs is equivalent to a mean-field metapopulation model where we follow the average state of a population of nodes with a given structural role (given pair l, k) coupled with the other populations according to the LCCM structure. The dynamical rules of the SIS model simply specify the sign and rates of the transition events, and we could as easily write similar equations for other dynamical processes. This approach relies on an averaging approximation but overlayed on an exact description of the LCCM structure.

B. Network comparison

The LCCM is fundamentally a compression method for network structure. While we demonstrated its potential in the context of percolation, it is also applicable to other problems where an extensive description of network structure is prohibitive. For instance, one classic solution to the task of comparing two networks is to calculate their graph edit distance (GED). The GED is defined along sequences of elementary graph operations, such as the insertion or deletion of a node or a link, that turns a network $\mathcal{G}_1$ into another network $\mathcal{G}_2$. Each edit operation e_i is associated with a given cost $c(e_i)$ based on the operation it performs. The distance $D(\mathcal{G}_1,\mathcal{G}_2)$ between networks $\mathcal{G}_1$ and $\mathcal{G}_2$ is then defined as the minimal cost for a sequence $\mathcal{P}(\mathcal{G}_1,\mathcal{G}_2)$ of edits that transforms $\mathcal{G}_1$ into $\mathcal{G}_2$: $D(\mathcal{G}_1,\mathcal{G}_2) = \min_{(e_1,\ldots,e_k)\in\mathcal{P}(\mathcal{G}_1,\mathcal{G}_2)} \sum_{i=1}^{k} c(e_i)$.

Algorithms for calculating these edit distances are well studied but unfortunately tend to scale exponentially with network size and are therefore usually tested on networks with only a handful of nodes [31]. This impractical scaling means that GED is virtually impossible to apply to the real, large, complex networks considered here. To solve this problem, a new avenue of research recently introduced a more statistical perspective to network comparison [32,33]. For example, Bagrow and Bollt introduced a principled approach where, instead of comparing networks, one compares graph portraits: the matrix $B_{k,\ell}$ containing the number of nodes with k neighbors at a shortest distance of ℓ [33]. A similar approach could be used on the LCCM and its joint layer-degree distribution, such that LCCM ensembles are compared instead of individual networks.

The LCCM has three main advantages over other statistical description of networks. First, the underlying network compression algorithm, the onion decomposition, is fast. Second, the compressed network can be mapped to connection rules yielding a link swapping algorithm to explore the ensemble. Third, the ensemble is described by an exact analytical formalism. The second and third advantages are particularly powerful since they allow one to interpolate between two networks, both computationally and analytically. When doing network comparison, it would therefore be possible to not only quantify the distance between two networks, but also generate intermediate networks between them. This could be leveraged, for example, to better explore the space of all possible robust designs for infrastructure networks.

ACKNOWLEDGMENTS

A. A. acknowledges financial support from the project Sentinelle Nord of the Canada First Research Excellence Fund, from "la Caixa" Foundation, and from the Spanish "Juan de la Cierva-incorporación" program (IJCI-2016-30193). L. H.-D. acknowledges support from the National Science Foundation Grant No. DMS-1829826. Both authors thank the Santa Fe Institute, the Institute for Disease Modeling, and the Vermont Complex Systems Center, where this work was completed.

APPENDIX A: LINK CORRELATION MATRIX

We define the symmetrical link correlation matrix $\mathbf{L}$ whose elements $L_{lk,l'k'}$ correspond to the fraction of links between nodes of class (l,k) and (l',k'). It has the following properties,

$$\frac{1}{2}\sum_{lk}\sum_{l'k'}(1+\delta_{ll'}\delta_{kk'})L_{lk,l'k'} = 1, \tag{A1}$$

since each type of link appears twice in the matrix except for the links connecting nodes of the same class (i.e., diagonal elements), and

$$\frac{1}{2}\sum_{l'k'}(1+\delta_{ll'}\delta_{kk'})L_{lk,l'k'} = \frac{kP(l,k)}{\langle k\rangle}, \tag{A2}$$

where $P(l,k)$ is the fraction of nodes belonging to the class (l,k) and $\langle k\rangle = \sum_{lk} kP(l,k)$ is the average degree.

APPENDIX B: DISTRIBUTION OF THE NUMBER AND OF THE COLOR OF STUBS

The connection rules of the LCCM indicate that a node of degree k in layer l and coreness c_l have at most c_l red stubs. Since red stubs are defined as half-links toward nodes in layers $l' \geq l$, they represent a fraction

$$\frac{1}{2}\sum_{l'\geq l}\sum_{k'}(1+\delta_{ll'}\delta_{kk'})L_{lk,l'k'} \tag{B1}$$

of all stubs in the network ensemble, where $\delta_{ll'}\delta_{kk'}$ accounts for the fact that a link connecting two nodes of class (l,k) contributes to two red stubs. This last quantity would be equal to

$$\frac{c_l P(l,k)}{\langle k\rangle} \tag{B2}$$

if every of these nodes had exactly c_l red stubs. Consequently, since the LCCM only dictates bounds on the number of each color, the probability that a node of degree k in layer l has exactly k^r red stubs is simply

$$\binom{c_l}{k^r}[p^r_{lk}]^{k^r}[1-p^r_{lk}]^{c_l-k^r}, \tag{B3}$$

where

$$p^r_{lk}=\frac{\sum_{l'\geq l}\sum_{k'}(1+\delta_{ll'}\delta_{kk'})L_{lk,l'k'}}{2c_l P(l,k)/\langle k\rangle}. \tag{B4}$$

Note that whenever layer l is the first layer of its core—when $c_l > c_{l-1}$—Eq. (B4) reduces to $p^r_{lk}=1$, meaning that each node has exactly c_l red stubs, as prescribed by the connection rules of the LCCM.

Similarly, the fraction of half-links shared with nodes in layers $l' < l-1$ (i.e., green stubs) is

$$\frac{1}{2}\sum_{l'<l-1}\sum_{k'}L_{lk,l'k'}. \tag{B5}$$

The maximal value of this quantity, however, varies in function of l. If the layer is the first layer of its shell (i.e., if $c_l > c_{l-1}$), then each node has c_l red stubs and up to $k-c_l$ green stubs according to the connections rules. If $c_l = c_{l-1}$, nodes that have exactly c_l red stubs can have up to $k-c_l-1$ green stubs since they must have at least one black stub, and can have up to $k-c_l$ otherwise. The maximal value of Eq. (B5) can therefore be summarized as

$$\frac{(k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}})P(l,k)}{\langle k\rangle}, \tag{B6}$$

such that the probability that a node of degree k in layer l has exactly k^g green stubs is

$$\binom{k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}{k^g}\left[\frac{k-c_l}{k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}p^g_{lk}\right]^{k^g}\left[1-\frac{k-c_l}{k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}p^g_{lk}\right]^{k-c_l-k^g-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}, \tag{B7}$$

with

$$p^g_{lk}=\frac{\sum_{l'<l-1}\sum_{k'}L_{lk,l'k'}}{2(k-c_l)P(l,k)/\langle k\rangle}. \tag{B8}$$

Combining Eqs. (B3) and (B7) yields the probability that a node in layer l and of degree k has k^r, k^g, and k^b red, green, and black stubs, respectively:

$$\begin{aligned}P_{lk}(k^r,k^g,k^b)=&\,\delta_{k,k^r+k^b+k^g}\binom{c_l}{k^r}[p^r_{lk}]^{k^r}[1-p^r_{lk}]^{c_l-k^r}\\&\times\binom{k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}{k^g}\left[\frac{k-c_l}{k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}p^g_{lk}\right]^{k^g}\left[1-\frac{k-c_l}{k-c_l-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}p^g_{lk}\right]^{k-c_l-k^g-\delta_{k^r,c_l}\delta_{c_l,c_{l-1}}}.\end{aligned} \tag{B9}$$

Finally, after some elementary algebra, it can be shown that the generating function $\varphi_{lk}(\boldsymbol{x})$ associated with this distribution is

$$\begin{aligned}\varphi_{lk}(\boldsymbol{x})=&\sum_{k^r,k^b,k^g}P(k^r,k^g,k^b|l,k)[x^r_{lk}]^{k^r}[x^g_{lk}]^{k^g}[x^b_{lk}]^{k^b}\\=&\,\delta_{c_l,c_{l-1}}x^b_{lk}[p^r_{lk}x^r_{lk}]^{c_l}\left[\left(1-\frac{k-c_l}{k-c_l-1}p^g_{lk}\right)x^b_{lk}+\frac{k-c_l}{k-c_l-1}p^g_{lk}x^g_{lk}\right]^{k-c_l-1}-\delta_{c_l,c_{l-1}}[p^r_{lk}x^r_{lk}]^{c_l}[(1-p^g_{lk})x^b_{lk}+p^g_{lk}x^g_{lk}]^{k-c_l}\\&+[(1-p^r_{lk})x^b_{lk}+p^r_{lk}x^r_{lk}]^{c_l}[(1-p^g_{lk})x^b_{lk}+p^g_{lk}x^g_{lk}]^{k-c_l}.\end{aligned} \tag{B10}$$

APPENDIX C: TRANSITION PROBABILITIES

With the distribution of the number of stubs of each color that nodes have being provided by Eq. (B10), the only missing quantities are the transition probabilities: the probability $Q^{\alpha}_{lk}(l',k',\alpha')$ that a stub of color α stemming from a node of class (l,k) leads to a stub of color α' attached to a node of class (l',k'). Once more, this information can be extracted from the link correlation matrix $\mathbf{L}$.

Let us recall that black stubs stemming from nodes of class (l,k) can lead only to red stubs attached to nodes in the previous layer (i.e., $l'=l-1$), which can be summarized by

$$Q^{b}_{lk}(l',k',\alpha')=\frac{\delta_{\alpha',r}\delta_{l',l-1}L_{l'k',lk}}{\sum_{l''}\sum_{k''}\delta_{l'',l-1}L_{l''k'',lk}}, \tag{C1}$$

where the denominator is proportional to the fraction of all stubs that are black and that are stemming from nodes of class (l,k). Similarly, since green stubs can only lead to red stubs attached to nodes in layer $l'<l-1$, we have

$$Q^{g}_{lk}(l',k',\alpha')=\begin{cases}\frac{\delta_{\alpha',r}L_{l'k',lk}}{\sum_{l''<l-1}\sum_{k''}L_{l''k'',lk}} & \text{if } l'<l-1\\ 0 & \text{otherwise.}\end{cases} \tag{C2}$$

Because red stubs can lead to all three colors of stubs, we first consider the case where a red stub leads to a black stub (i.e., to a node in layer $l'=l+1$), which corresponds to

$$Q^{r}_{lk}(l',k',b)=\frac{\delta_{l',l+1}L_{lk,l'k'}}{\sum_{l''\geq l}\sum_{k''}(1+\delta_{ll''}\delta_{kk''})L_{lk,l''k''}}, \tag{C3}$$

where the denominator is proportional to the fraction of all stubs that corresponds to red stubs stemming from nodes of class (l,k). In the case of red stubs leading to red stubs—i.e., links between nodes in the same layer—we need to double the contribution of $L_{lk,lk}$ since each link between nodes of the same class contributes to two red stubs, which yields

$$Q^{r}_{lk}(l',k',r)=\frac{\delta_{ll'}(1+\delta_{kk''})L_{lk,l''k''}}{\sum_{l''\geq l}\sum_{k''}(1+\delta_{ll''}\delta_{kk''})L_{lk,l''k''}}. \tag{C4}$$

The case of red stubs leading to green stubs is similar to Eq. (C2) and is straightforward to obtain:

$$Q^{r}_{lk}(l',k',g)=\begin{cases}\frac{L_{lk,l'k'}}{\sum_{l''\geq l}\sum_{k''}(1+\delta_{ll''}\delta_{kk''})L_{lk,l''k''}} & \text{if } l'>l+1\\ 0 & \text{otherwise.}\end{cases} \tag{C5}$$

Finally, by inserting Eqs. (C1)–(C5) in Eq. (2), we obtain

$$\gamma^{r}_{lk}(\boldsymbol{x})=\frac{\sum_{l'\geq l}\sum_{k'}L_{lk,l'k'}[\delta_{ll'}(1+\delta_{kk'})x^{r}_{l'k'}+\delta_{ll'-1}x^{b}_{l'k'}+(1-\delta_{ll'})(1-\delta_{ll'-1})x^{g}_{l'k'}]}{\sum_{l''\geq l}\sum_{k''}(1+\delta_{ll''}\delta_{kk''})L_{lk,l''k''}}, \tag{C6a}$$

$$\gamma^{b}_{lk}(\boldsymbol{x})=\frac{\sum_{k'}L_{l-1k',lk}x^{r}_{l-1k'}}{\sum_{k''}L_{l-1k'',lk}}, \tag{C6b}$$

$$\gamma^{g}_{lk}(\boldsymbol{x})=\frac{\sum_{l'<l-1}\sum_{k'}L_{l'k',lk}x^{r}_{l'r}}{\sum_{l''<l-1}\sum_{k''}L_{l''k'',lk}}. \tag{C6c}$$

APPENDIX D: PERCOLATION THRESHOLD

The value of the percolation threshold p_c can be computed analytically by a linear stability analysis of the solution $\boldsymbol{a}=\mathbf{1}$ of Eq. (5). Substituing $a^{\alpha}_{lk}=1-\varepsilon^{\alpha}_{lk}$, where $\varepsilon^{\alpha}_{lk}\ll 1$, yields

$$\varepsilon^{\alpha}_{lk}=p\sum_{l'k'\alpha'}\left.\frac{\partial f^{\alpha}_{lk}(\boldsymbol{x})}{\partial x^{\alpha'}_{l'k'}}\right|_{\boldsymbol{x}=\mathbf{1}}\varepsilon^{\alpha'}_{l'k'}, \tag{D1}$$

when limiting the expansion of $f^{\alpha}_{lk}(\mathbf{1}-\boldsymbol{\varepsilon})$ to the first order. The last equation can be rewritten as an eigenvalue problem,

$$\boldsymbol{\varepsilon}=p\mathbf{M}\boldsymbol{\varepsilon}, \tag{D2}$$

thus indicating that the fixed point $\boldsymbol{a}=\mathbf{1}$ loses its stability—i.e., the extensive component emerges—when the largest eigenvalue of $p\mathbf{M}$ exceeds 1. The percolation threshold p_c therefore equals the reciprocal of the largest eigenvalue of $\mathbf{M}$, which, by virtue of the Perron-Frobenius theorem, is real and positive.

The elements of $\mathbf{M}$ can be written as

$$\frac{\partial f^{\alpha}_{lk}(\mathbf{1})}{\partial x^{\alpha'}_{l'k'}}=\frac{1}{\langle k^{\alpha}\rangle_{lk}}\sum_{\alpha''}\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x^{\alpha}_{lk}\partial x^{\alpha''}_{lk}}\frac{\partial\gamma^{\alpha''}_{lk}(\mathbf{1})}{\partial x^{\alpha'}_{l'k'}}, \tag{D3}$$

where the derivatives are calculated directly from Eq. (B10) and (C6a)–(C6c). While the derivatives of $\gamma_{lk}^{\alpha}\left(\boldsymbol{x}\right)$ are straightforward, the derivatives of $\varphi_{lk}(\boldsymbol{x})$ require special care with respect to the value of $k-c_l$. To facilitate the numerical implementation of the formalism, we provide the explicit expression of the derivatives of $\varphi_{lk}(\boldsymbol{x})$:

$$\langle k^r\rangle_{lk}=\frac{\partial\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^r}=c_l p_{lk}^r, \tag{D4a}$$

$$\langle k^g\rangle_{lk}=\frac{\partial\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^g}=\begin{cases}(k-c_l)p_{lk}^g-\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)p_{lk}^g & \text{if } k-c_l\le 1\\ (k-c_l)p_{lk}^g & \text{otherwise,}\end{cases} \tag{D4b}$$

$$\langle k^b\rangle_{lk}=\frac{\partial\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^b}=\begin{cases}c_l(1-p_{lk}^g)+(k-c_l)(1-p_{lk}^g)+\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)p_{lk}^g & \text{if } k-c_l\le 1\\ c_l(1-p_{lk}^r)+(k-c_l)(1-p_{lk}^g) & \text{otherwise,}\end{cases} \tag{D4c}$$

$$\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^{r2}}=c_l(c_l-1)[p_{lk}^r]^2, \tag{D4d}$$

$$\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^r\partial x_{lk}^g}=\begin{cases}c_l(k-c_l)p_{lk}^r p_{lk}^g-\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}c_l(k-c_l)p_{lk}^g & \text{if } k-c_l\le 1\\ c_l(k-c_l)p_{lk}^r p_{lk}^g & \text{otherwise.}\end{cases} \tag{D4e}$$

$$\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^r\partial x_{lk}^b}=\begin{cases}c_l(c_l-1)p_{lk}^r(1-p_{lk}^r)+c_l(k-c_l)p_{lk}^r(1-p_{lk}^g)+\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}c_l(k-c_l)p_{lk}^g & \text{if } k-c_l\le 1\\ c_l(c_l-1)p_{lk}^r(1-p_{lk}^r)+c_l(k-c_l)p_{lk}^r(1-p_{lk}^g) & \text{otherwise.}\end{cases} \tag{D4f}$$

$$\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^{g2}}=\begin{cases}(k-c_l)(k-c_l-1)[p_{lk}^g]^2-\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)(k-c_l-1)[p_{lk}^g]^2 & \text{if } k-c_l\le 2\\ (k-c_l)(k-c_l-1)[p_{lk}^g]^2-\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)(k-c_l-1)[p_{lk}^g]^2+\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)^2[p_{lk}^g]^2\frac{k-c_l-2}{k-c_l-1} & \text{otherwise,}\end{cases} \tag{D4g}$$

$$\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^g\partial x_{lk}^b}=\begin{cases}c_l(k-c_l)p_{lk}^g(1-p_{lk}^r) & \text{if } k-c_l\le 1\\ c_l(k-c_l)p_{lk}^g(1-p_{lk}^r)+(k-c_l)(k-c_l-1)p_{lk}^g(1-p_{lk}^g)+\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)(k-c_l-1)[p_{lk}^g]^2 & \text{if } k-c_l=2\\ c_l(k-c_l)p_{lk}^g(1-p_{lk}^r)+(k-c_l)(k-c_l-1)p_{lk}^g(1-p_{lk}^g)+\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)(k-c_l-1)[p_{lk}^g]^2 & \\ -\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)^2[p_{lk}^g]^2\frac{k-c_l-2}{k-c_l-1} & \text{otherwise,}\end{cases} \tag{D4h}$$

$$\frac{\partial^2\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^{b2}}=\begin{cases}c_l(c_l-1)(1-p_{lk}^r)^2+2c_l(k-c_l)(1-p_{lk}^r)(1-p_{lk}^g) & \text{if } k-c_l\le 1\\ c_l(c_l-1)(1-p_{lk}^r)^2+2c_l(k-c_l)(1-p_{lk}^r)(1-p_{lk}^g)+(k-c_l)(k-c_l-1)(1-p_{lk}^g)^2 & \\ -\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)(k-c_l-1)[p_{lk}^g]^2 & \text{if } k-c_l=2\\ c_l(c_l-1)(1-p_{lk}^r)^2+2c_l(k-c_l)(1-p_{lk}^r)(1-p_{lk}^g)+(k-c_l)(k-c_l-1)(1-p_{lk}^g)^2 & \\ -\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)(k-c_l-1)[p_{lk}^g]^2+\delta_{c_l,c_{l-1}}[p_{lk}^r]^{c_l}(k-c_l)^2[p_{lk}^g]^2\frac{k-c_l-2}{k-c_l-1} & \text{otherwise.}\end{cases} \tag{D4i}$$

Let us recall that $c_l\neq c_{l-1}$ and $p_{lk}^r=1$ whenever $k=c_l$, since these nodes are in the first layer of their core by definition, and that we set $c_1\neq c_0$ to simplify the notation. Finally, note that Eqs. (D4) satisfy

$$\sum_{\alpha}\frac{\partial\varphi_{lk}(\mathbf{1})}{\partial x_{lk}^{\alpha}}=\sum_{\alpha}\langle k^{\alpha}\rangle_{lk}=k \tag{D5}$$

and

$$\sum_{\alpha,\alpha'} \frac{\partial^2 \varphi_{lk}(\mathbf{1})}{\partial x_{lk}^{\alpha} \partial x_{lk}^{\alpha'}} = \sum_{\alpha,\alpha'} \langle k^{\alpha}(k^{\alpha'} - \delta_{\alpha\alpha'}) \rangle_{lk} = k(k-1) \quad \text{(D6)}$$

for $\alpha, \alpha' \in \{r, g, b\}$ and regardless of the value of $k - c_l$, as expected.

APPENDIX E: OTHER APPROACHES TO NETWORK COMPRESSION

Let us recall that we aimed to find a compression of networks that (1) provides an intensive description of network structure, (2) allows an exact mathematical description, and (3) preserves the accuracy of the message passing approximation. In this Appendix, we briefly discuss some other promising network compression approaches, but to which one or more of the following shortcomings led us to discard them from our contribution. First, some descriptions of network structure were simply not designed to capture structures relevant to percolation or other dynamical processes. Second, some allow simple mathematical description, but move the difficulty of the problem to inferring the structure on which the description should apply which, in some cases, is still an active area of research. Third, some simply do not have known mathematical descriptions.

1. Stochastic block models (SBMs)

In its most general version, the stochastic block model (SBM) is described as N nodes divided into B blocks with a given number of edges e_{rs} between blocks r and s. More relevant to dynamical processes and percolation is the degree-corrected version that also maintains the degree distribution of nodes. The problem of inferring the block structure within the SBM is not a simple one, but it allows principled solutions in terms of minimum description length or Bayesian model selection [34]. Once inferred, the framework of multitype probability generating functions used in the main text can be used to mathematically describe the structure around nodes with a given degree and block membership.

Unfortunately, the goal of the SBM is to describe large-scale block structure in networks and not necessarily the microscale or local connections at which our approach operates. In fact, the treelike networks for which our mathematical framework is exact do not contain significant blocks encoding their structure. For example, the optimal description of the finite Cayley tree shown in Fig. 2 in terms of a degree-corrected SBM reduces the tree to a bipartite network with one block containing the even generations of the tree and the other block containing the odd generations. The ensemble of networks defined by this block structure is obviously very similar to that of the configuration model or correlated configuration model with small exceptions. The degree-corrected SBM will therefore not preserve the connected tree structure, simply because it is not the objective of the SBM.

2. Motif decompositions (MD)

Motif decomposition (MD) is related philosophically to the SBM, but focuses on smaller (microscopic) structures [9,35]. The objective of MD is to find common small motifs with a fixed structure of connections among each other. Nodes are then described by their number of memberships to different motifs; the complex treelike composition of motifs can be described mathematically by solving their internal structure independently [9,10].

The larger problem with MD is that there is currently no principled approach to infer the set of relevant motifs in real networks to accurately predict the outcome of percolation. For example, the tree of Fig. 2 could be exactly described by keeping the root node as a motif of size one, and then by distinguishing all wedges (motifs of size 3) by their motif degree. This would result in three unique motifs and an exact description. However, this decomposition breaks down if two or more generations to the Cayley tree are added, which would then require motifs that span more than two generations of the tree. This issue illustrates the larger problem: There is no systematic way to exactly decompose these trees in motifs, and even less so for complex networks.

3. *dk* series

The dk series is one way to provide a systematic MD [36]. The dk series hierarchically extends the concept of motif degree to motifs of increasing sizes as a nested series. Per definition, the zeroth element in the dk series, given by $d = 0$ and described by the $0k$ distribution, is simply the average degree of the network. The next element, the $1k$ distribution, looks at motifs of size 1 (nodes) and their connections through motif degree to other motifs of size 1, hence corresponding to the standard degree distribution. The second element, the $2k$ distribution, follows motifs of size 2 (edges), which corresponds to the joint degree distribution defining the number of edges connecting nodes of degree k and k'.

For $d = 3$, the two possible motifs are triangles and wedges, distinguished by the degree k, k', and k'' of the nodes involved. This series extends to motifs of larger size for higher values of d, and the motif-degree correlations become exponentially more involved (and become extensive in network size once we start distinguishing every unique motif of a given size d because of the unique sequence of d degrees involved in the motif).

While a mathematical description exists for $d \in \{0, 1, 2\}$, there is currently no mathematical description for the ensemble of all networks preserving all features captured by the dk series with $d \geq 3$. Given how constraining the dk series becomes for higher d [36], extending exact mathematical descriptions to these highly constrained

ensembles is a hard but potentially important problem to tackle.

4. Geometric embeddings

The models considered above all consist of pairwise interactions (i.e., either two nodes connected by a link or a node connected to a motif), thus explaining our difficulty to model general forms of clustering realistically, which are three-body interactions. To circumvent this limitation, an interesting approach assumes that complex networks are embedded in a latent metric space where the distance between each pair of nodes controls their probability of being connected. From a mathematical point of view, the interactions are still pairwise, but the triangle inequality of the underlying metric space indirectly allows the reproduction of the clustering patterns observed in real complex networks [37–39].

Critically, most applications require the inference of the positions of the nodes of real complex networks into a latent hyperbolic space [40–43], an optimization problem that scales badly with the number of nodes. Many heuristics, some inspired by machine-learning methods, have been proposed over the years [38,41,44–47], and Refs. [40–43] provide compelling evidence of their success. However, the range of complex networks that these methods can map into hyperbolic space is still too restricted for this methodology to become the state-of-the-art approach to model the structure of complex systems. Moreover, although these maps provide invaluable information to understand the overall organization of these systems, the inferred positions of the nodes are not precise enough to generate representative surrogates of the original network.

While the next generation of embedding algorithms will surely overcome these limitations, the remaining challenge will consist in incorporating the inferred positions of the nodes in hyperbolic space into a mathematical formalism to predict the outcome of various dynamical processes such as percolation. To the best of our knowledge, such a formalism has yet to be developed.

APPENDIX F: NETWORK DATASETS

Table I provides the domain, source, brief description, and reference (when available) for the 111 networks used in this study.

TABLE I. Description of the network datasets.

Network name	Domain	Source	Description or name within source[a]	Reference
AdoHealth	Social	[48]	Adolescent health (1994)	[49]
Arabidopsis	Biological	[48]	Arabidopsis interactome (2011)	[50]
ArtExhibit	Social	[48]	Art exhibit dynamic contacts (2011)	[51]
arXiv	Social	[48]	Scientific collaborations in physics (1995–2005)	[52]
Brightkite	Social	[48]	Location-based social networks	[53]
Cargoships	Transport	associated web page[b]	Shipping journeys between major commercial ports	[54]
CargoshipsBB	Transport	from the authors	Shipping journeys between major commercial ports	[55]
CatBrain2013	Connectome	[48]	Cat brain (2013)	[56,57]
CatCortexThalamus1999a	Connectome	[48]	Cat cerebral hemisphere	[58]
CatCortexThalamus1999b	Connectome	[48]	Cat cerebral hemisphere	[58]
CElegans	Connectome	[48]	*C. elegans* neurons (1986)	[59,60]
CElegansGenetic	Biological	[48]	Multiplex protein interactions (2015)	[61]
CoastalFoodWeb[c]	Food Web	[48]	Coastal food webs with metazoan parasites	[62]
Corporate	Economic	[48]	Corporate ownership (2002)	[63]
Counties	Geographical	public dataset[d]	Adjacent counties in the U.S.	
Digg	Information	[48]	Digg reply network (2008)	[64]
Drosophila13	Connectome	from the original paper	Synapses in the optic medulla of Drosophila	[65]
Drosophila13_full	Connectome	[48]	Fly medulla (2013)	[65]
DrosophilaGenetic	Biological	[48]	Multiplex protein interactions (2015)	[61]
Ecoli	Biological	from the authors	Shared participation of metabolites in *E. coli*	[43,66]
EColi2	Biological	from the authors	Shared participation of metabolites in *E. coli*	[43,66]
Email	Social	[48]	Email network (EU research inst.)	[67]
Enron	Social	[48]	Email network (Enron corpus)	[68]
EuropeGridKit	Infrastructure	public dataset[e]	Transmission network	
EuropeSciGrid	Infrastructure	public dataset[f]	Transmission network	
FloridaBayDry	Food Web	[48]	Florida cypress wetlands food web (1998)	
FloridaBayWet	Food Web	[48]	Florida cypress wetlands food web (1998)	
FoodWeb	Food Web	[48]	Little Rock Lake food web (1991)	[69]
GermanRoads	Infrastructure	[48]	German highway system (2002)	[70]

(Table continued)

TABLE I. *(Continued)*

Network name	Domain	Source	Description or name within source[a]	Reference
GermanySciGrid	Infrastructure	public dataset[f]	Transmission network	
Gnutella	Technological	[48]	Gnutella p2p networks (2002)	[71]
HollywoodFilmMusic	Social	[48]	Hollywood film music	[72]
HomoGenetic	Biological	[48]	Multiplex genetic interactions (2014)	[61]
Human08	Connectome	from the authors	Axonal connections between brain regions	[73,74]
Human12a	Connectome	from the authors	Axonal connections between brain regions	[73,74]
Human13l	Connectome	[48]	Axonal connections between brain regions	[75]
Human13m	Connectome	[48]	Axonal connections between brain regions	[75]
HumanHerpes4Genetic	Biological	associated web page[g]	Genetic-protein interactions Epstein-Barr virus	[61]
HumanHIV1	Biological	associated web page[g]	Genetic-protein interactions human HIV type 1	[61]
InternetAS	Technological	[48]	Route Views AS graphs (1997–1998)	
InternetCaida	Technological	[48]	CAIDA AS graphs (2004–2007)	
InternetOregon	Technological	[48]	Route Views (extended) AS graphs (2001)	
Internet	Technological	[48]	Internet AS graph (2006)	[6]
Jazz	Social	[48]	Jazz collaboration network	[76]
Macaque93	Connectome	[48]	Macaque cortical connectivity (Young)	[77]
Macaque95	Connectome	associated web page[h]	Cortical connectivity	[78]
MangroveEstuaryDry	Food Web	[48]	South Florida ecosystems (2005)	
MangroveEstuaryWet	Food Web	[48]	South Florida ecosystems (2005)	
MetabolicTrade	Biological	unpublished dataset	Metabolite exchanges of synthetic microbial populations	
Mouse14	Connectome	from the original paper	Axonal connections between brain regions	[79]
MouseRetina2013	Connectome	[48]	Mouse retina (2013)	[80]
MouseRetina2013_full	Connectome	[48]	Mouse retina (2013)	[80]
MusGenetic	Biological	[48]	Multiplex protein interactions (2015)	[61]
NAmericaGridKit	Infrastructure	public dataset[i]	Transmission network	
NetScientists	Social	[48]	Scientific collaborations in network science (2006)	[81]
NorwegianBoard	Social	[48]	Norwegian Boards of Directors (2002–2011)	[23]
PGP	Information	[48]	PGP web of trust (2004)	[24]
PGP2009	Information	[48]	PGP web of trust (2009)	[82]
PINHSapiens	Biological	associated web page[j]	Protein-protein interactions in *H. sapiens*	[26]
PlantPollinator	Food Web	[48]	Clements and Long plant-pollinator web	
PlantPollinator2	Food Web	[48]	McCullen plant-pollinator web	
PlantPollinator3	Food Web	[48]	Robertson plant-pollinator web	
Plasmodium	Biological	[48]	Multiplex protein interactions (2015)	[61]
PolBlogs	Information	[48]	Political blogs network (2004)	[83]
PolBlogs2	Information	[48]	Political blogs network (2004)	[83]
PolBooks	Information	[48]	Political books network (2004)	
PolishGrid	Infrastructure	associated web page[k]	Subset of the power grid of Poland	[84]
PowerGrid	Infrastructure	[48]	Western U.S. Power Grid	[1]
ProteinCore	Biological	associated web page[l]	Protein-protein interactions in *S. cerevisiae*	[85]
RattusGenetic	Biological	[48]	Multiplex protein interactions (2015)	[61]
RattusNorvegicus1	Connectome	[48]	Rat brain (2011–2013)	[86]
RattusNorvegicus2	Connectome	[48]	Rat brain (2011–2013)	[86]
RattusNorvegicus3	Connectome	[48]	Rat brain (2011–2013)	[86]
RhesusBrain2012	Connectome	[48]	Rhesus brain (2012)	[87]
SacchCere	Biological	[48]	Multiplex protein interactions (2015)	[61]
Slashdot	Social	[48]	Slashdot Zoo friend-foe network (2009)	[88]
Square	Synthetic	⋯	50×50 square lattice	
UCIrvine	Social	[48]	Facebook100	[89]
USAirports500	Transport	[48]	U.S. airport network (top 500; 2002)	[90]
USAirports	Transport	[48]	U.S. airport networks (2010)	
UScommodities[m]	Economic	from the original paper	Trade between industrial sectors in the U.S. in 2007	[91]
UScommute[m]	Transport	from the original paper	Daily flow of commuters between counties in the US 2000	[91]

(Table continued)

TABLE I. *(Continued)*

Network name	Domain	Source	Description or name within source[a]	Reference
WAirports	Transport	[48]	Openflights airport network (2010)	
WikipediaNorms	Information	[48]	Wikipedia norms (2015)	[92]
WorldAirports	Transport	[48]	Openflights airport network (2016)	
WordAssoc	Information	[48]	USF word associations	[93]
WTW2013	Economic	from the authors	Trade exchanges between countries in 2013	[42]
ZebraFinch17	Connectome	from the original paper	Synapses in the basal-ganglia (Area X) of the Zebra Finch	[94]
ZebraFinch17b	Connectome	from the authors	Synapses in the basal-ganglia (Area X) of the Zebra Finch	[94]

[a]A brief description of the network dataset or its name can be found in Ref. [48].
[b]Ref. [95].
[c]21 unique edge lists that can be downloaded from Ref. [96].
[d]Ref. [97].
[e]Ref. [98].
[f]Ref. [99].
[g]Ref. [100].
[h]Ref. [101].
[i]Ref. [98].
[j]Ref. [102].
[k]Ref. [103].
[l]Ref. [104].
[m]The unweighted backbone considered in Ref. [55] was also used.

[1] D. J. Watts and S. H. Strogatz, *Collective Dynamics of 'Small-World' Networks*, Nature (London) **393**, 440 (1998).
[2] N. Goldenfeld and L. P. Kadanoff, *Simple Lessons from Complexity*, Science **284**, 87 (1999).
[3] S. H. Strogatz, *Complex Systems: Romanesque Networks*, Nature (London) **433**, 365 (2005).
[4] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010), p. 720.
[5] A.-L. Barabási, *Network Science* (Cambridge University Press, Cambridge, England, 2016), p. 474.
[6] B. Karrer, M. E. J. Newman, and L. Zdeborová, *Percolation on Sparse Networks*, Phys. Rev. Lett. **113**, 208702 (2014).
[7] F. Radicchi, *Percolation in Real Interdependent Networks*, Nat. Phys. **11**, 597 (2015).
[8] L. Hébert-Dufresne, J. A. Grochow, and A. Allard, *Multi-Scale Structure and Topological Anomaly Detection via a New Network Statistic: The Onion Decomposition*, Sci. Rep. **6**, 31708 (2016).
[9] B. Karrer and M. E. J. Newman, *Random Graphs Containing Arbitrary Distributions of Subgraphs*, Phys. Rev. E **82**, 066118 (2010).
[10] A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé, *General and Exact Approach to Percolation on Random Graphs*, Phys. Rev. E **92**, 062807 (2015).
[11] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, *The Unreasonable Effectiveness of Tree-Based Theory for Networks with Clustering*, Phys. Rev. E **83**, 036112 (2011).
[12] V. Latora, V. Nicosia, and G. Russo, *Complex Networks: Principles, Methods and Applications* (Cambridge University Press, Cambridge, England, 2017), p. 594.
[13] S. B. Seidman, *Network Structure and Minimum Degree*, Soc. Networks **5**, 269 (1983).
[14] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, *k-Core Organization of Complex Networks*, Phys. Rev. Lett. **96**, 040601 (2006).
[15] A. C. C. Coolen, A. De Martino, and A. Annibale, *Constrained Markovian Dynamics of Random Graphs*, J. Stat. Phys. **136**, 1035 (2009).
[16] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, *Configuring Random Graph Models with Fixed Degree Sequences*, SIAM Rev. **60**, 315 (2018).
[17] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé, *Percolation on Random Networks with Arbitrary k-Core Structure*, Phys. Rev. E **88**, 062820 (2013).
[18] P. Colomer-de-Simón, M. Á. Serrano, M. G. Beiró, J. I. Alvarez-Hamelin, and M. Boguñá, *Deciphering the Global Organization of Clustering in Real Complex Networks.*, Sci. Rep. **3**, 2517 (2013).
[19] M. E. J. Newman, *Spread of Epidemic Disease on Networks*, Phys. Rev. E **66**, 016128 (2002).
[20] A. Vázquez and Y. Moreno, *Resilience to Damage of Graphs with Degree Correlations*, Phys. Rev. E **67**, 015101 (2003).
[21] F. Radicchi and C. Castellano, *Breaking of the Site-Bond Percolation Universality in Networks*, Nat. Commun. **6**, 10196 (2015).
[22] P. Colomer-de-Simón and M. Boguñá, *Double Percolation Phase Transition in Clustered Complex Networks*, Phys. Rev. X **4**, 041020 (2014).
[23] C. Seierstad and T. Opsahl, *For the Few Not the Many? The Effects of Affirmative Action on Presence, Prominence, and Social Capital of Women Directors in Norway*, Scand. J. Manag. **27**, 44 (2011).

[24] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, *Models of Social Networks Based on Social Distance Attachment*, Phys. Rev. E **70**, 056122 (2004).
[25] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graphs Over Time: Densification Laws, Shrinking Diameters and Possible Explanations*, in *Proceedings of the Elevnth ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)* (ACM, New York, NY, USA, 2005), p. 177.
[26] C. Song, S. Havlin, and H. A. Makse, *Self-Similarity of Complex Networks*, Nature (London) **433**, 392 (2005).
[27] F. Radicchi and C. Castellano, *Beyond the Locally Treelike Approximation for Percolation on Real Networks*, Phys. Rev. E **93**, 030302 (2016).
[28] P. Zhang, *Spectral Estimation of the Percolation Transition in Clustered Networks*, Phys. Rev. E **96**, 042303 (2017).
[29] M. Shrestha, S. V. Scarpino, and C. Moore, *Message-Passing Approach for Recurrent-State Epidemic Models on Networks*, Phys. Rev. E **92**, 022821 (2015).
[30] C. Castellano and R. Pastor-Satorras, *Relevance of Backtracking Paths in Epidemic Spreading on Networks*, Phys. Rev. E **98**, 052313 (2018).
[31] M. Neuhaus, K. Riesen, and H. Bunke, *Fast Suboptimal Algorithms for the Computation of Graph Edit Distance*, in *Structural, Syntactic, and Statistical Pattern Recognition. SSPR/SPR 2006*, edited by D. Y. Yeung, J. T. Kwok, A. Fred, F. Roli, D. de Ridder Lecture Notes in Computer Science, Vol. 4109 (Springer, New York, 2006).
[32] T. A. Schieber, L. Carpi, A. Díaz-Guilera, P. M. Pardalos, C. Masoller, and M. G. Ravetti, *Quantification of Network Structural Dissimilarities*, Nat. Commun. **8**, 13928 (2017).
[33] J. P. Bagrow and E. M. Bollt, *An Information-Theoretic, All-Scales Approach to Comparing Networks*, arXiv: 1804.03665.
[34] T. P. Peixoto, *Efficient Monte Carlo and Greedy Heuristic for the Inference of Stochastic Block Models*, Phys. Rev. E **89**, 012804 (2014).
[35] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network Motifs: Simple Building Blocks of Complex Networks*, Science **298**, 824 (2002).
[36] C. Orsini, M. M. Dankulov, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguná, G. Caldarelli *et al.*, *Quantifying Randomness in Real Networks*, Nat. Commun. **6**, 8627 (2015).
[37] D. Krioukov, F. Papadopoulos, A. Vahdat, and M. Boguñá, *Curvature and Temperature of Complex Networks*, Phys. Rev. E **80**, 035101 (2009).
[38] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov, *Popularity versus Similarity in Growing Networks*, Nature (London) **489**, 537 (2012).
[39] M. Á. Serrano, D. Krioukov, and M. Boguñá, *Self-Similarity of Complex Networks and Hidden Metric Spaces*, Phys. Rev. Lett. **100**, 078701 (2008).
[40] G. Alanis-Lobato, P. Mier, and M. Andrade-Navarro, *The Latent Geometry of the Human Protein Interaction Network*, Bioinformatics **34**, 2826 (2018).
[41] M. Boguñá, F. Papadopoulos, and D. Krioukov, *Sustaining the Internet with Hyperbolic Mapping.*, Nat. Commun. **1**, 62 (2010).
[42] G. García-Pérez, M. Boguñá, A. Allard, and M. Á. Serrano, *The Hidden Hyperbolic Geometry of International Trade: World Trade Atlas 1870-2013*, Sci. Rep. **6**, 33441 (2016).
[43] M. Á. Serrano, M. Boguñá, and F. Sagués, *Uncovering the Hidden Geometry Behind Metabolic Metworks*, Mol. Biosyst. **8**, 843 (2012).
[44] G. Alanis-Lobato, P. Mier, and M. A. Andrade-Navarro, *Efficient Embedding of Complex Networks to Hyperbolic Space via Their Laplacian*, Sci. Rep. **6**, 30108 (2016).
[45] T. Blasius, T. Friedrich, A. Krohmer, and S. Laue, *Efficient Embedding of Scale-Free Graphs in the Hyperbolic Plane*, IEEE/ACM Trans. Netw. **26**, 920 (2018).
[46] A. Muscoloni, J. M. Thomas, S. Ciucci, G. Bianconi, and C. V. Cannistraci, *Machine Learning Meets Complex Networks via Coalescent Embedding in the Hyperbolic Space*, Nat. Commun. **8**, 1615 (2017).
[47] F. Papadopoulos, R. Aldecoa, and D. Krioukov, *Network Geometry Inference Using Common Neighbors*, Phys. Rev. E **92**, 022807 (2015).
[48] https://icon.colorado.edu.
[49] J. Moody, *Peer Influence Groups: Identifying Dense Clusters in Large Networks*, Soc. Networks **23**, 261 (2001).
[50] *Arabidopsis Interactome Mapping Consortium, Evidence for Network Evolution in an Arabidopsis Interactome Map*, Science **333**, 601 (2011).
[51] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, *What's in a Crowd? Analysis of Face-to-Face Behavioral Networks*, J. Theor. Biol. **271**, 166 (2011).
[52] M. E. J. Newman, *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[53] E. Cho, S. A. Myers, and J. Leskovec, *Friendship and Mobility: User Movement in Location-Based Social Networks*, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)* (ACM New York, NY, USA, 2011), p. 1082.
[54] P. Kaluza, A. Kolzsch, M. T. Gastner, and B. Blasius, *The Complex Network of Global Cargo Ship Movements*, J. R. Soc. Interface **7**, 1093 (2010).
[55] A. Allard, M. Á. Serrano, G. García-Pérez, and M. Boguñá, *The Geometric Nature of Weights in Real Complex Networks*, Nat. Commun. **8**, 14103 (2017).
[56] J. W. Scannell, C. Blakemore, and M. P. Young, *Analysis of Connectivity in the Cat Cerebral Cortex*, J. Neurosci. **15**, 1463 (1995).
[57] M. A. de Reus and M. P. van den Heuvel, *Rich Club Organization and Intermodule Communication in the Cat Connectome*, J. Neurosci. **33**, 12929 (2013).
[58] J. W. Scannell, *The Connectional Organization of the Cortico-Thalamic System of the Cat*, Cereb. Cortex **9**, 277 (1999).
[59] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, *Structural Properties of the Caenorhab-*

ditis elegans Neuronal Network, PLoS Comput. Biol. **7**, e1001066 (2011).

[60] Y.-Y. Ahn, H. Jeong, and B. J. Kim, *Wiring Cost in the Organization of a Biological Neuronal Network*, Physica (Amsterdam) **367A**, 531 (2006).

[61] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora, *Structural Reducibility of Multilayer Networks*, Nat. Commun. **6**, 6864 (2015).

[62] J. A. Dunne, K. D. Lafferty, A. P. Dobson, R. F. Hechinger, A. M. Kuris, N. D. Martinez, J. P. McLaughlin, K. N. Mouritsen, R. Poulin, K. Reise, D. B. Stouffer, D. W. Thieltges, R. J. Williams, and C. D. Zander, *Parasites Affect Food Web Structure Primarily through Increased Diversity and Complexity*, PLoS Biol. **11**, e1001579 (2013).

[63] K. Norlen, G. Lucas, M. Gebbie, and J. Chuang, *EVA: Extraction, Visualization and Analysis of the Telecommunications and Media Ownership Network*, in *Proceedings of International Telecommunications Society 14th Biennial Conference (ITS2002), Seoul, Korea, 2002* (Elsevier, Amsterdam, The Netherlands, 2002), pp. 27–129.

[64] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, *Social Synchrony: Predicting Mimicry of User Actions in Online Social Media*, in *Proceedings of the 2009 International Conference on Computational Science and Engineering* (IEEE, Piscataway, NJ, 2009), pp. 151–158.

[65] S.-Y. Takemura *et al.*, *A Visual Motion Detection Circuit Suggested by Drosophila Connectomics*, Nature (London) **500**, 175 (2013).

[66] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson, *A Comprehensive Genome-Scale Reconstruction of Escherichia coli Metabolism–2011*, Mol. Syst. Biol. **7**, 535 (2011).

[67] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graph Evolution: Densification and Shrinking Diameters*, ACM Trans. Knowl. Discovery Data **1**, 2 (2007).

[68] B. Klimt and Y. Yang, *The Enron Corpus: A New Dataset for Email Classification Research*, in *European Conference on Machine Learning*, (Springer, Berlin, Heidelberg, 2004), pp. 217–226.

[69] N. D. Martinez, *Artifacts or Attributes? Effects of Resolution on the Little Rock Lake Food Web*, Ecol. Monogr. **61**, 367 (1991).

[70] M. Kaiser and C. C. Hilgetag, *Spatial Growth of Real-World Networks*, Phys. Rev. E **69**, 036103 (2004).

[71] R. Matei, A. Iamnitchi, and P. Foster, *Mapping the Gnutella Network*, IEEE Internet Comput. **6**, 50 (2002).

[72] R. R. Faulkner, *Music on Demand: Composers and Careers in the Hollywood Film Industry.*, Administrative science quarterly **34**, 318 (1989).

[73] A. Avena-Koenigsberger, J. Goni, R. F. Betzel, M. P. van den Heuvel, A. Griffa, P. Hagmann, J.-P. Thiran, and O. Sporns, *Using Pareto Optimality to Explore the Topology and Dynamics of the Human Connectome*, Philos. Trans. R. Soc. B Biol. Sci. **369**, 20130530 (2014).

[74] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, *Mapping the Structural Core of Human Cerebral Cortex*, PLoS Biol. **6**, e159 (2008).

[75] W. R. G. Roncal, Z. H. Koterba, D. Mhembere, D. M. Kleissas, J. T. Vogelstein, R. Burns, A. R. Bowles, D. K. Donavos, S. Ryman, R. E. Jung, L. Wu, V. Calhoun, and R. J. Vogelstein, *MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics*, in *Proceedings of the IEEE Global Conference on Signal and Information* (IEEE, Piscataway, NJ, 2013), pp. 313–316.

[76] P. M. Gleiser and L. Danon, *Community Structure in Jazz*, Adv. Complex Syst. **06**, 565 (2003).

[77] M. P. Young, *The Organization of Neural Systems in the Primate Cerebral Cortex*, Proc. R. Soc. B **252**, 13 (1993).

[78] M. Kaiser and C. C. Hilgetag, *Nonoptimal Component Placement, but Short Processing Paths, due to Long-Distance Projections in Neural Systems*, PLoS Comput. Biol. **2**, e95 (2006).

[79] S. W. Oh *et al.*, *A Mesoscale Connectome of the Mouse Brain*, Nature (London) **508**, 207 (2014).

[80] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, *Connectomic Reconstruction of the Inner Plexiform Layer in the Mouse Retina*, Nature (London) **500**, 168 (2013).

[81] M. E. J. Newman, *Finding Community Structure in Networks Using the Eigenvectors of Matrices*, Phys. Rev. E **74**, 036104 (2006).

[82] O. Richters and T. P. Peixoto, *Trust Transitivity in Social Networks*, PLoS One **6**, e18384 (2011).

[83] L. A. Adamic and N. Glance, *The Political Blogosphere and the 2004 U.S. Election: Divided They Blog*, in *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD '05)* (ACM, New York, NY, USA, 2005), pp. 36–43.

[84] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, *MATPOWER: Steady-State Operations, Systems Research and Education*, IEEE Transactions on Power Systems;IEEE Transactions on Power Electronics **26**, 12 (2011).

[85] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society*, Nature (London) **435**, 814 (2005).

[86] M. Bota and L. W. Swanson, *Online Workbenches for Neural Network Connections*, J. Comp. Neurol. **500**, 807 (2007).

[87] L. Harriger, M. P. van den Heuvel, and O. Sporns, *Rich Club Organization of Macaque Cerebral Cortex and Its Role in Network Communication*, PLoS One **7**, e46497 (2012).

[88] J. Kunegis, A. Lommatzsch, and C. Bauckhage, *The Slashdot Zoo: Mining a Social Network with Negative Edges*, in *Proceedings of the 18th International Conference on the World Wide Web (WWW '09)* (ACM, New York, NY, USA, 2009), p. 741.

[89] A. L. Traud, P. J. Mucha, and M. A. Porter, *Social Structure of Facebook Networks*, Physica A (Amsterdam) **391A**, 4165 (2012).

[90] V. Colizza, R. Pastor-Satorras, and A. Vespignani, *Reaction-Diffusion Processes and Metapopulation Models in Heterogeneous Networks*, Nat. Phys. **3**, 276 (2007).

[91] D. Grady, C. Thiemann, and D. Brockmann, *Robust Classification of Salient Links in Complex Networks*, Nat. Commun. **3**, 864 (2012).
[92] B. Heaberlin and S. DeDeo, *The Evolution of Wikipedia's Norm Network*, Futur. Internet **8**, 14 (2016).
[93] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, *The University of South Florida Free Association, Rhyme, and Word Fragment Norms*, Behav. Res. Methods Instrum. Comput. **36**, 402 (2004).
[94] S. N. Dorkenwald, P. J. Schubert, M. F. Killinger, G. Urban, S. Mikula, F. Svara, and J. Kornfeld, *Automated Synaptic Connectivity Inference for Volume Electron Microscopy*, Nat. Methods **14**, 435 (2017).
[95] http://www.icbm.de/mathematische-modellierung/forschung/marine-bioinvasion/cargo-ship-movement/.
[96] https://datadryad.org//resource/doi:10.5061/dryad.b8r5c.
[97] https://www.census.gov/geo/reference/county-adjacency.html.
[98] https://zenodo.org/record/47317.
[99] https://www.power.scigrid.de/pages/downloads.html.
[100] https://comunelab.fbk.eu/data.php.
[101] https://www.dynamic-connectome.org.
[102] http://dip.doe-mbi.ucla.edu/dip/Main.cgi.
[103] http://www.pserc.cornell.edu/matpower/.
[104] http://www.cfinder.org/.