Quantifying the compressibility of complex networks

Christopher W. Lynn[a,b] and Danielle S. Bassett[c,d,e,f,g,h,1]

[a]Initiative for the Theoretical Sciences, Graduate Center, City University of New York, New York, NY 10016; [b]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544; [c]Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104; [d]Department of Physics and Astronomy, College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104; [e]Department of Electrical and Systems Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104; [f]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [g]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; and [h]Santa Fe Institute, Santa Fe, NM 87501

Edited by Albert-László Barabási, Northeastern University, Boston, MA, and accepted by Editorial Board Member Christopher Jarzynski June 17, 2021 (received for review November 11, 2020)

Many complex networks depend upon biological entities for their preservation. Such entities, from human cognition to evolution, must first encode and then replicate those networks under marked resource constraints. Networks that survive are those that are amenable to constrained encoding—or, in other words, are compressible. But how compressible is a network? And what features make one network more compressible than another? Here, we answer these questions by modeling networks as information sources before compressing them using rate-distortion theory. Each network yields a unique rate-distortion curve, which specifies the minimal amount of information that remains at a given scale of description. A natural definition then emerges for the compressibility of a network: the amount of information that can be removed via compression, averaged across all scales. Analyzing an array of real and model networks, we demonstrate that compressibility increases with two common network properties: transitivity (or clustering) and degree heterogeneity. These results indicate that hierarchical organization—which is characterized by modular structure and heterogeneous degrees—facilitates compression in complex networks. Generally, our framework sheds light on the interplay between a network's structure and its capacity to be compressed, enabling investigations into the role of compression in shaping real-world networks.

information theory | complex networks | rate distortion | compression

Complex networks are often encoded in biology and, thereby, utilized and replicated by biological systems. The brain encodes language (1), knowledge (2), music (3), social (4, 5), and transportation networks (6); the human mind uses these internal representations to engage in linguistic communication, build on existing understanding, sing a victorious melody, strengthen a valuable friendship, and walk the covered holloways (7). Similarly, biological networks among molecular and cellular components are encoded at various scales in genetic material (8–11), and evolution uses these encodings to propagate network topologies in a surviving species. From brains to genes, the biological materials that encode complex networks operate under marked constraints on time, energy, metabolism, and physical extent, among others. Such constraints determine which networks persist into the future—in particular, those whose topology can be efficiently encoded. These shared constraints raise a fundamental question: How does the structure of a network facilitate efficient encodings?

Encoding a network (indeed, encoding any piece of information) involves a natural trade-off between simplicity and accuracy. One could construct a simple representation that omits the fine-scale details of a network. Or one could build a representation that captures a network's intricate structure, but is complicated and unwieldy. An efficient encoding strikes an optimal balance between simplicity and accuracy; that is, it is a compression (12, 13). In fact, compression—a foundational branch of information theory—has provided key insights into optimal network representations, yielding principled algorithms for constructing coarse-grained maps of complex systems (14–16).

Building upon this progress, here, we investigate how the structure of complex networks facilitates compression. Intuitively, just as natural images are easier to compress than white noise due to their visual patterns and regularities, so, too, should networks with strong structural regularities be more compressible than random networks. But do homogeneous topologies, such as those found in lattice-like networks, make systems more compressible, or is compression facilitated by the hierarchical organization found in many real networks? To answer these questions, here, we develop a framework for quantifying the compressibility of complex networks. Applying our framework to several real and model networks, we identify specific network features that facilitate compression. Together, these results elucidate how a network's topology impacts its compressibility and suggest that many real-world networks may be shaped by the pressure to be compressed.

Rate-Distortion Theory of Network Clustering

In compression (13), one begins with an information source, a sequence of items that defines the object of interest. For networks, the details of information flow often vary from one

Significance

Real-world networks are complex, comprising vast webs of interconnected elements performing a diverse array of social and biological functions. Common among many networks, however, is the pressure to be efficiently compressed—either in the brain or in the genetic code. But just as files on a computer can be compressed to differing degrees, what makes one network more compressible than another? To answer this question, we adapt tools from information theory to quantify the compressibility of a network. Studying real-world and model networks, we find that hierarchical organization—with tight clustering and heterogeneous degrees—increases compressibility, enabling compressed representations across scales. Generally, our framework provides an information-theoretic method for investigating the interplay between network structure and compression.

Author contributions: C.W.L. and D.S.B. designed research; C.W.L. performed research; C.W.L. contributed new reagents/analytic tools; C.W.L. analyzed data; and C.W.L. and D.S.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.-L.B. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: dsb@seas.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2023473118/-/DCSupplemental.

Published August 4, 2021.

context to another. Therefore, a logical choice for the information source is a random walk, which contains all of the details about the structure of a network and nothing more (14).

One then seeks to reduce the amount of information in the sequence, which can be accomplished in two complementary, yet distinct, ways. In *lossless* compression, one removes statistical redundancy in the sequence while maintaining an exact description of the network. This approach has provided important information-theoretic perspectives on the problem of community detection, wherein one constructs a coarse-grained representation at a specific scale of description (14). By contrast, here, we seek to quantify the compressibility of a network itself, without selecting a desired scale. To do so, we employ rate-distortion theory, the foundation of *lossy* compression. In lossy compression, rather than removing statistical redundancy in the sequence, one instead removes redundant features of the network directly. As we will see, directly coarse-graining the network will enable tractable strategies for compressing networks across all scales and, in doing so, will allow us to develop an intuitive definition for network compressibility.

Compressing Random Walks. To see how compression unfolds in practice, consider the network in Fig. 1*A*. A random walk on the network defines a sequence of nodes $\boldsymbol{x} = (x_1, x_2, \ldots)$, with each node transitioning to one of its four neighbors uniformly at random. The rate at which this sequence generates information is given by the entropy $H(\boldsymbol{x})$, which (because there are four possible nodes at each step) equals 2 bits (see *Materials and Methods* for a definition of $H(\boldsymbol{x})$). To reduce the amount of information in the sequence, we can construct a coarse-grained representation by clustering nodes together (14–16). This clustering yields a new sequence $\boldsymbol{y} = (y_1, y_2, \ldots)$, where y_t is the cluster containing node x_t (Fig. 1*B*), which communicates information at a rate equal to the mutual information $I(\boldsymbol{x},\boldsymbol{y}) = H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x})$ (12, 13, 15, 16). If the clusters are chosen deterministically, as is common (4, 14, 17), then the conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$ vanishes, and the information rate simplifies to the entropy of the clustered sequence, $I(\boldsymbol{x},\boldsymbol{y}) = H(\boldsymbol{y})$.

Consider, for example, a trivial clustering in which each node belongs to its own cluster (Fig. 1 *B*, *Top*). In this case, we maintain a complete description of the network, but we have not reduced the information rate, since $I(\boldsymbol{x},\boldsymbol{y}) = H(\boldsymbol{x}) = 2$ bits. By contrast, consider the opposite setting in which all nodes belong to the same large cluster (Fig. 1 *B*, *Bottom*). Now, we have reduced the information rate to zero ($I(\boldsymbol{x},\boldsymbol{y}) = 0$ bits), but all details about the network structure have been lost. Between these two extremes lies a range of clusterings (such as that in Fig. 1 *B*, *Middle*), each inducing its own information rate and yielding a unique distortion of the network structure.

Scale as a Measure of Distortion. Building representations that strike an optimal balance between minimizing information rate while also minimizing distortion is precisely the purview of rate-distortion theory (12, 13). As in any rate-distortion problem, one must choose a specific definition for the distortion of the object of interest. When clustering a network, a natural choice for the distortion presents itself: the scale of description. Specifically, for a network with N nodes and a clustering with n clusters, we define the scale to be $S = 1 - \frac{n-1}{N}$. For example, if $n = N$, then we have an exact fine-grained description of the network at a scale $S = 1/N$ (Fig. 1 *B*, *Top*), whereas if $n = 1$, then one cluster encloses the entire network and $S = 1$ (Fig. 1 *B*, *Bottom*).

At each scale S (equivalently, for each number of clusters n), we seek to identify the clustering that minimizes the information rate $I(\boldsymbol{x},\boldsymbol{y})$. This optimal information rate, denoted $R(S)$, defines a unique rate-distortion curve for each network (Fig. 1*C*). If a network is easier to compress, then at each scale S, one should be able to find a clustering that is more efficient, thereby reducing the information rate R (Fig. 1*C*, vertical line); similarly, for a given information rate R, one should be able to construct a more fine-grained clustering, thereby decreasing the scale S

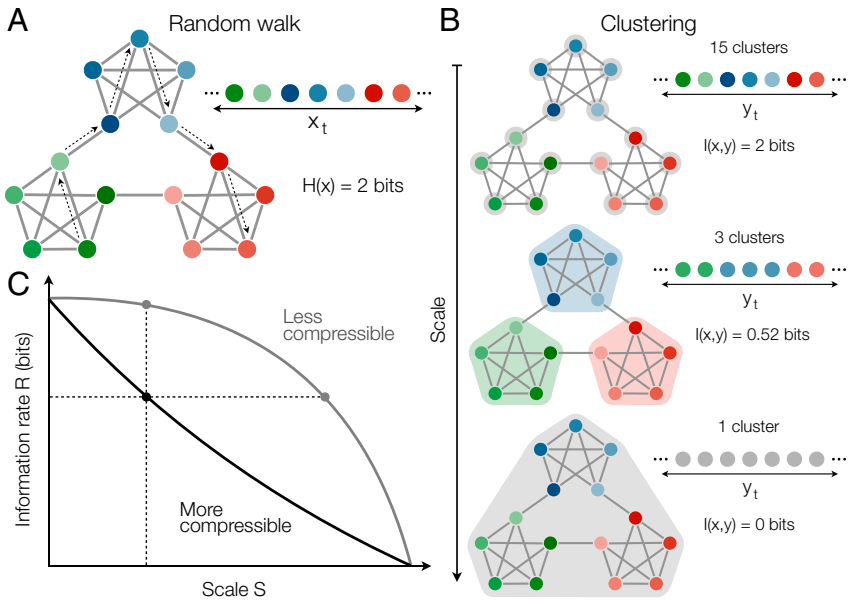

Fig. 1. Rate-distortion theory of random walks on networks. (*A*) A simple network with $N = 15$ nodes, each with constant degree $k = 4$. A random walk $\boldsymbol{x}$ generates information at a rate of $H(\boldsymbol{x}) = 2$ bits. (*B*) Network clusterings across various scales of description. (*B*, *Top*) For $n = 15$ clusters, each containing its own node, the sequence communicates $I(\boldsymbol{x},\boldsymbol{y}) = H(\boldsymbol{x}) = 2$ bits of information. (*B*, *Middle*) For $n = 3$ clusters, each corresponding to one of the three modules in the original network, the information rate is $I(\boldsymbol{x},\boldsymbol{y}) = 0.52$ bits. (*B*, *Bottom*) For $n = 1$ cluster containing the entire network, the sequence no longer communicates information. (*C*) Schematic of the optimal information rate R as a function of the scale of description S for networks that are either more compressible (black) or less compressible (gray). For more compressible networks, one can achieve a lower information rate at a given scale of description (vertical line), and one can achieve a more fine-grained description for a given information rate (horizontal line).

(Fig. 1*C*, horizontal line). Thus, in order to quantify the compressibility of a network, we must first be able to compute its rate-distortion curve.

Computing the Rate-Distortion Curve of a Network

Computing the rate-distortion curve $R(S)$ of a network—in particular, doing so efficiently to enable applications to large systems—poses two distinct challenges. First, we must estimate the mutual information $I(\boldsymbol{x},\boldsymbol{y})$ for different clusterings; and second, we must identify the clusterings that minimize this information rate across all scales.

Although estimating mutual information is generally difficult (18), the simplicity of our setup allows for tractable upper and lower bounds (*Materials and Methods*). Of particular interest is the upper bound $\bar{I}(\boldsymbol{x},\boldsymbol{y}) \geq I(\boldsymbol{x},\boldsymbol{y})$, which follows by approximating the clustered sequence $\boldsymbol{y}$ as Markovian [a property that we note is not guaranteed, even though the original random walk $\boldsymbol{x}$ is Markovian (13)]. Rather than minimizing the information rate $I(\boldsymbol{x},\boldsymbol{y})$ directly, we instead minimize the upper bound $\bar{I}(\boldsymbol{x},\boldsymbol{y})$, thereby yielding an upper bound $\bar{R}(S)$ on the rate-distortion curve. For simplicity, in what follows, we often refer to $\bar{I}(\boldsymbol{x},\boldsymbol{y})$ as the information rate and $\bar{R}(S)$ as the rate-distortion curve.

To compute $\bar{R}(S)$—that is, to find clusterings that minimize the information rate $\bar{I}(\boldsymbol{x},\boldsymbol{y})$—we employ a greedy clustering algorithm. Beginning with $n = N$ clusters, each containing its own node, we combine the pair of clusters that yields the largest reduction in the information rate $\bar{I}(\boldsymbol{x},\boldsymbol{y})$. Repeating this agglomerative process across all scales S (until only one cluster remains), we arrive at an estimate for the rate-distortion curve $\bar{R}(S)$. To speed up the calculation, rather than searching through all $\binom{n}{2}$ pairs of clusters at each step, we only consider a limited number of pairs chosen via principled heuristics (*Materials and Methods*). Importantly, these heuristics do not affect the definitions of information-theoretic quantities, such as the rate $I(\boldsymbol{x},\boldsymbol{y})$ and upper bound $\bar{I}(\boldsymbol{x},\boldsymbol{y})$. In practice, not only do these heuristics enable applications to networks of approximately 10^3 nodes, they also improve the accuracy of the rate-distortion estimates themselves (*SI Appendix*, Fig. S1).

We are now prepared to compute the rate-distortion curve for a specific system. In Fig. 2*A*, we plot the upper and lower bounds on the rate-distortion curve $R(S)$ for Zachary's karate club network (19). As is true for all networks (*Materials and Methods*), the two bounds are exact at both the minimum scale $S = 1/N$ (when the information rate simply equals the entropy of random walks $H(\boldsymbol{x})$) and the maximum scale $S = 1$ (when the information rate is zero). Moreover, the two bounds remain close across all intermediate scales (Fig. 2*A*), demonstrating that the upper

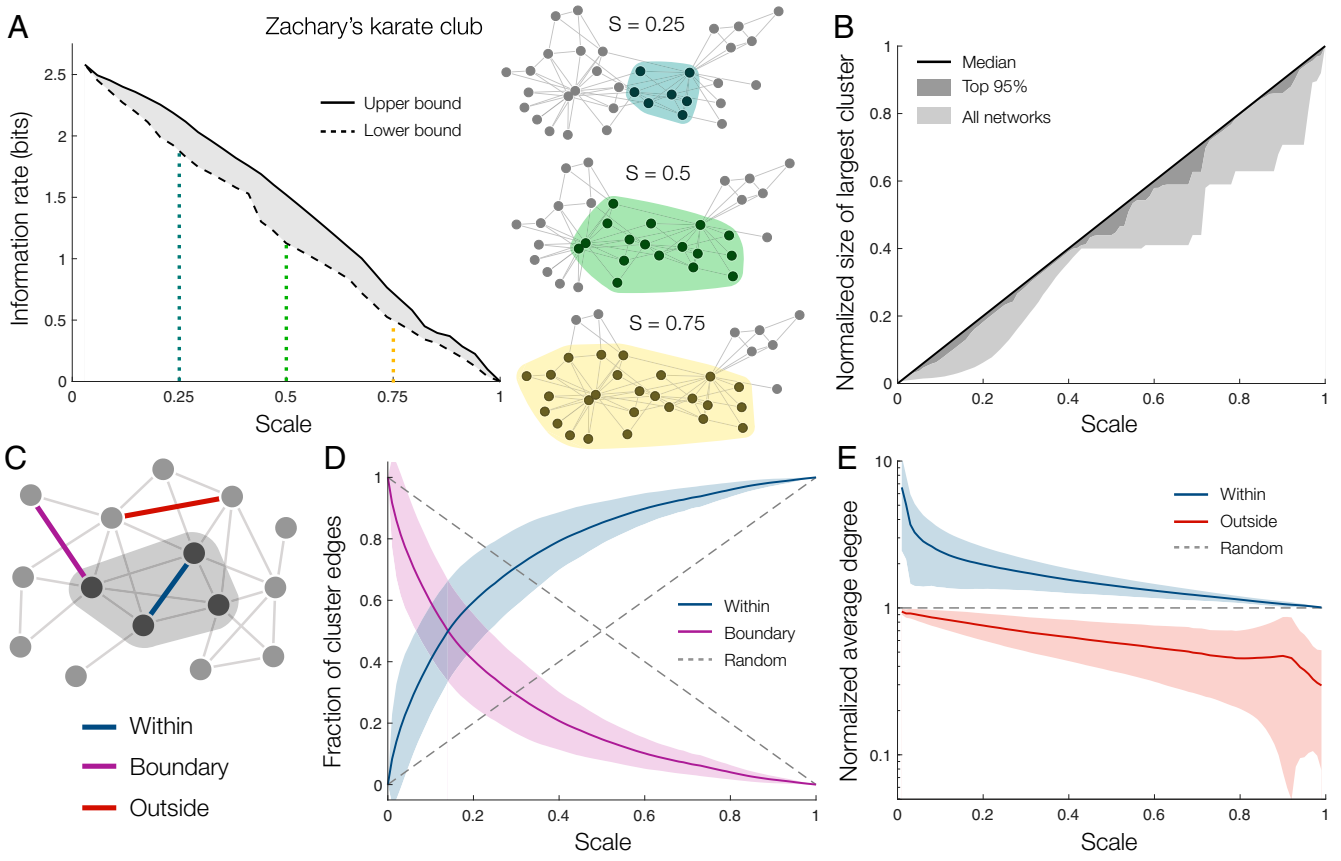

Fig. 2. Properties of optimal clusterings. (*A*, *Left*) Upper bound (solid line) and lower bound (dashed line) on the optimal information rate $R(S)$ as a function of the scale of description S for Zachary's karate club network (19). (*A*, *Right*) Across all scales, the optimal compression includes one large cluster, which we illustrate for $S = 0.25$, 0.5, and 0.75. (*B*) Size of the largest cluster in a compression, normalized by the size of the network N, as a function of the scale S for the real networks in *SI Appendix*, Table S1 (20–23). The median over real networks (solid line) matches the largest possible normalized cluster size, $(N - n + 1)/N = S$, indicating that (across all scales) most networks admit one large cluster of maximal size. (*C*) Illustration of edges within the one large cluster (blue), on the boundary of the cluster (purple), and outside the cluster (red). (*D*) Fraction of the k_c edges emanating from the large cluster that either connect to nodes outside the cluster $1 - G_{cc}/k_c$ (purple) or remain within the cluster G_{cc}/k_c (blue) as a function of the scale S. (*E*) Average degree of nodes inside (blue) and outside (red) the large cluster, normalized by the average degree of the network, as a function of the scale S. In *D* and *E*, solid lines and shaded regions represent averages and one-SD error bars, respectively, over the real networks (*SI Appendix*, Table S1) (20–23), and dashed lines correspond to clusters with nodes selected at random.

bound $\bar{R}(S)$ provides a good approximation to the true rate-distortion curve $R(S)$. To understand how the rate-distortion curve depends on the structure of a network, however, it helps to examine the properties of optimal compressions themselves.

Properties of Optimal Compressions

Using the framework developed above, we are ultimately interested in studying compression in real systems. The networks chosen for analysis span from communication networks (including semantic, language, and music networks) and information networks (including hyperlinks on the web and citations in science) to social networks, animal and protein interactions, transportation networks, and structural and functional connections in the brain (*Materials and Methods*; *SI Appendix*, Table S1) (20–23). Although these networks encompass a wide range of systems bridging several orders of magnitude in size, they are all encoded biologically, either in genetic material or in the neural code.

Emergence of One Large Cluster. To begin, we compute the rate-distortion curve $\bar{R}(S)$ for each of the above networks, and we confirm that these upper bounds provide good approximations to the true rate-distortion curves $R(S)$ (*SI Appendix*, Fig. S2). In the process of computing $\bar{R}(S)$, our compression algorithm also provides estimates for the optimal clusterings over all scales. Examining the structure of these compressions, we find a striking consistency across different networks. As can be observed in Zachary's karate club (Fig. 2 *A, Right*), rather than dividing the network into multiple clusters of moderate size, optimal compressions tend to comprise one large cluster containing $N-n+1=SN$ nodes and $n-1$ minimal clusters each containing one node. In fact, among the networks studied, this tendency to form one large cluster is a nearly ubiquitous feature of optimal compressions (Fig. 2*B*).

We remark that the clustering that minimizes the information rate need not (and, indeed, does not) provide a faithful characterization of a network's community structure, as is the goal in community detection (14–16). Instead, we find that optimal compressions seek to identify the group of nodes that can be combined to maximally reduce the information rate. By dividing the network into two parts—one inside the large cluster and the other outside—the challenge of compressing random walks thus resembles the graph-partitioning problem (24), which has generated key insights about the modular structure of networks across scales (17). This simplification, in turn, allows us to develop analytic predictions about the properties of optimal compressions and the structures of compressible networks.

Information Rate of Optimal Compressions. Although our framework is general, applying to any weighted, directed network (*Materials and Methods*), in order to make analytic progress, here, we focus on the special case of an unweighted, undirected network with adjacency matrix G_{ij}. For such a network, the entropy of random walks takes the simple form $H(\boldsymbol{x})=\frac{1}{2E}\sum_i k_i \log k_i$, where $k_i=\sum_j G_{ij}$ is the degree of node i, $E=\frac{1}{2}\sum_{ij} G_{ij}$ is the number of edges in the network, and $\log(\cdot)$ is base two such that information is measured in bits.

Now consider forming one large cluster c. One can show (*Materials and Methods*) that the information rate of the clustered network is given by

$$\bar{I}(\boldsymbol{x},\boldsymbol{y})=\frac{1}{2E}\left[\sum_{i\notin c} k_i\log k_i + k_c \log k_c - 2\sum_{i\notin c} G_{ic}\log G_{ic} - G_{cc}\log G_{cc}\right], \quad [1]$$

where $k_c=\sum_{i\in c} k_i$ is the sum of the degrees of the nodes in c, $G_{ic}=\sum_{j\in c} G_{ij}$ is the number of edges connecting nodes in c to a given node i, and $G_{cc}=\sum_{ij\in c} G_{ij}$ is the number of edges connecting nodes within c.

Information Content of Different Edges. Using Eq. **1**, can we predict the properties of the optimal cluster c? More broadly, can we anticipate the types of network topologies that facilitate compression? To answer these questions, it helps to group the edges in a network into three distinct categories (Fig. 2*C*): those connecting nodes within c, those connecting nodes outside of c, and those on the boundary of c (connecting nodes within c to nodes outside of c). We can gauge which type of edge is preferred over the others by comparing their contributions to the information rate (Eq. **1**). An optimal compression will maximize the number of edges that are informationally preferred (contributing only weakly to the information rate), while limiting edges that are informationally costly.

For example, adding an edge within c increases the information rate by $\Delta\bar{I}^{\text{within}}\approx\frac{1}{2E}(2\log k_c-2\log G_{cc})$. By contrast, adding an edge on the boundary of c (say, connecting c to a node $i\notin c$) yields an increase of roughly $\Delta\bar{I}^{\text{boundary}}\approx\frac{1}{2E}(\log k_i+\log k_c-2\log G_{ic})$. For a large cluster c, we have $k_c, G_{cc}\gg k_i, G_{ic}$, from which one can show that $\Delta\bar{I}^{\text{within}}\lesssim\Delta\bar{I}^{\text{boundary}}$ (*SI Appendix*). Thus, edges within the large cluster are informationally preferred to those on the boundary, suggesting that the large cluster will seek to combine groups of nodes that are tightly connected to one another and sparsely connected to the rest of the network. Indeed, in real networks, we find that among the k_c edges emanating from the large cluster, the proportion $1-G_{cc}/k_c$ that connects to the rest of the network is much smaller than chance (Fig. 2*D*). This proportion of edges leaving the cluster is a well-studied quantity, known as the conductance or Cheeger constant of a network (17). Thus, networks with low conductance—such as those with modular structure and strong transitivity (the tendency for nodes to form triangles, also known as clustering)—should be highly compressible (17, 25). This is our first hypothesis about the impact of network structure on compressibility.

We now consider an edge connecting two nodes i and j outside of c, which increases the information rate by approximately $\Delta\bar{I}^{\text{outside}}\approx\frac{1}{2E}(\log k_i+\log k_j)$. As before, one can show that $\Delta\bar{I}^{\text{within}}\lesssim\Delta\bar{I}^{\text{outside}}$ (*SI Appendix*), hence demonstrating that edges within the large cluster are informationally preferred to those outside the cluster. In turn, this preference for the large cluster to include as many edges as possible suggests that c will favor high-degree nodes over low-degree nodes, which we confirm in real networks (Fig. 2*E*). This result leads to our second hypothesis: Networks should be more compressible if they have heterogeneous degrees (or heavy-tailed degree distributions), containing "rich clubs" of high-degree hub nodes (26, 27). Given the predictions that modular and heterogeneous topologies facilitate compression, we now propose a quantitative definition for the compressibility of a network.

Quantifying Network Compressibility

Intuitively, a network should be compressible if one can achieve a large reduction in the information rate at a given scale (Fig. 1*C*). However, rather than choosing a specific scale S (equivalently, a specific number of clusters n), we would like our definition of compressibility to be a property of the network itself. We therefore define the compressibility of a network to be the amount of information that can be removed via compression, averaged across all scales,

$$C = H(\boldsymbol{x}) - \frac{1}{N}\sum_{S} R(S). \quad [2]$$

Visually, the compressibility represents the area above a network's rate-distortion curve (Fig. 3*A*). In practice, plugging our tractable upper bound on the rate-distortion curve $\bar{R}(S)$ into Eq. **2** yields a lower bound $\underline{C}$, which (for simplicity) we will refer to as compressibility.

To make the notion of compressibility concrete, consider the class of random k-regular networks (Fig. 3*B*). On average, these networks have no structure (besides the requirement that nodes have uniform degree k), which allows us to derive an analytic approximation for the rate-distortion curve (*SI Appendix*),

$$\bar{R}(S) \approx (1-S)^2 \log k + S(1-S)\log N - S\log S. \quad [3]$$

Each individual network, however, contains small structural variations, such as groups of nodes that are more tightly connected than expected. Generating random k-regular networks and computing their rate-distortion curves directly, we find that optimal compressions are able to capitalize on these structural variations (*SI Appendix*, Fig. S3), thereby achieving lower information rates than the approximation in Eq. **3** (Fig. 3*C*). By contrast, as the degree k increases, the networks become uniform in structure, and the analytic approximation becomes exact (Fig. 3*C*).

Using Eq. **3**, one can predict the compressibility of k-regular networks. Specifically, noting that the entropy of k-regular networks is $\log k$ (*Materials and Methods*), and approximating the average in Eq. **2** by an integral over S, we arrive at the analytic form

$$\underline{C} \approx \frac{2}{3}\log k - \frac{1}{6}\log N - \frac{1}{4\ln 2}, \quad [4]$$

which we verify numerically (Fig. 3*D*). We note that the compressibility grows logarithmically with the degree k, reflecting the fact that networks with larger degrees have more information to be removed via compression (*Materials and Methods*). Indeed, computing the compressibility of the real networks in *SI Appendix*, Table S1 (20–23), we find precisely the same logarithmic dependence on the average degree (Fig. 3*E*). Furthermore, we verify that this logarithmic dependence generalizes to directed versions of the networks (*SI Appendix*, Fig. S5) and is not simply due to our clustering heuristics (*SI Appendix*, Fig. S6). These results demonstrate that the compressibility of a network increases predictably with average degree. But how does compressibility depend on the topology of a complex network?

Impact of Network Structure on Compressibility

Based on the properties of optimal compressions (Fig. 2), we hypothesized that the compressibility of a network should increase with both 1) transitivity and 2) degree heterogeneity. To investigate the impact of transitivity on compressibility, we consider a class of stochastic block networks (Fig. 4*A*), wherein nodes are grouped into modules of equal size, and a specified fraction f of the edges in the network connect nodes within the same module. We find that optimal compressions take advantage of this modular structure by clustering together nodes within the same module (*SI Appendix*, Fig. S3). Indeed, strengthening the modular structure—that is, increasing the fraction f of within-module edges—decreases the rate-distortion curve $\bar{R}(S)$ (Fig. 4*B*). We therefore find that compressibility increases with both modularity (Fig. 4*C*) and transitivity (Fig. 4*D*). Importantly, these results on stochastic block networks generalize to real networks, with increases in transitivity yielding significant improvements in network compressibility (Fig. 4*E*).

To examine the dependence of compressibility on degree heterogeneity, we study scale-free networks (Fig. 4*F*), which have heavy-tailed degree distributions $P(k) \sim k^{-\gamma}$ characterized by a power-law exponent γ (27). Optimal compressions exploit this heterogeneous structure by clustering together high-degree hub nodes (*SI Appendix*, Fig. S3). As γ decreases, accentuating the heterogeneity in node degrees, the rate-distortion curve $\bar{R}(S)$ increases at small scales and decreases at intermediate and large scales (Fig. 4*G*). Both of these rate-distortion effects serve to improve the compressibility of scale-free networks (Fig. 4*H*). Moreover, rather than indirectly investigating the impact of heavy-tailed structure via the scale-free exponent γ, we can directly quantify the degree heterogeneity of a given network $h = \langle |k_i - k_j| \rangle / \langle k \rangle$, where $\langle |k_i - k_j| \rangle$ is the absolute difference in degrees averaged over all pairs of nodes and $\langle k \rangle$ is the average degree. We find that the compressibility of scale-free networks grows linearly with degree heterogeneity (Fig. 4*I*), a result that

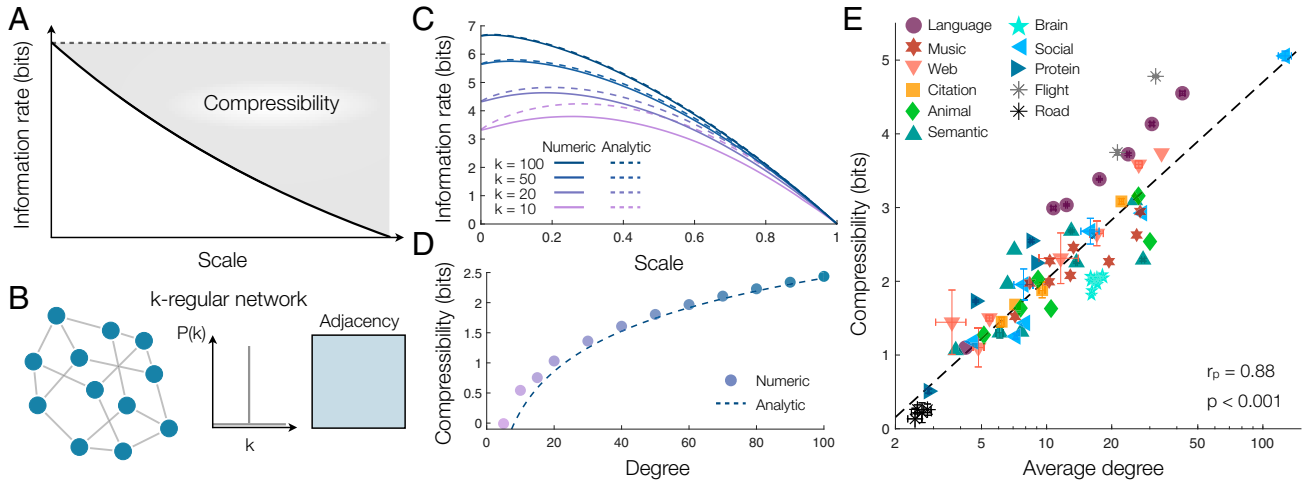

Fig. 3. Quantifying compressibility. (*A*) The compressibility of a network (shaded region) is the area between the rate-distortion curve (solid line) and the entropy of random walks (dashed line). (*B*) A k-regular network, characterized only by the requirement that all nodes have constant degree k. (*C*) Rate-distortion curves $\bar{R}(S)$ for k-regular networks with different degrees k. (*D*) Compressibility $\underline{C}$ of k-regular networks versus degree k. In *C* and *D*, solid lines and data points are averages over 50 randomly generated networks, each of size $N = 10^3$, and dashed lines indicate analytic predictions (Eqs. **3** and **4**). (*E*) Compressibility $\underline{C}$ versus average degree for the real networks in *SI Appendix*, Table S1 (20–23). We note that average degree is plotted on a log scale. Dashed line indicates a logarithmic fit. For networks of size $N > 10^3$, data points and error bars represent means and SDs over 50 randomly sampled subnetworks of 10^3 nodes each (*Materials and Methods*).

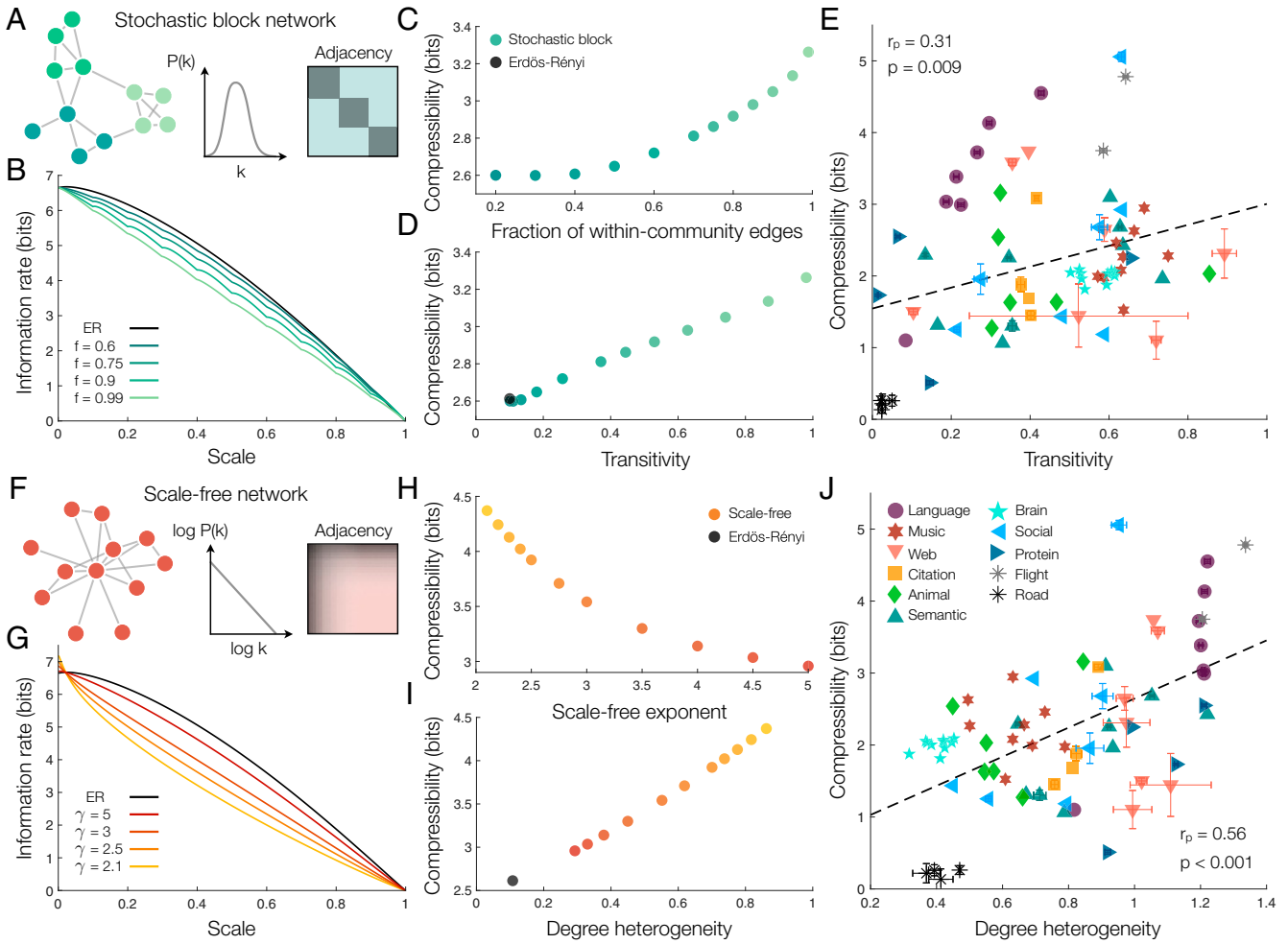

Fig. 4. Compressibility increases with transitivity and degree heterogeneity. (*A*) Stochastic block network, characterized by dense connectivity within modules and sparse connectivity between modules. (*B*) Rate-distortion curves $\bar{R}(S)$ for Erdös-Rényi (ER) networks (black line) and stochastic block networks (colored lines) with 10 modules and different fractions f of within-module edges. Undulations in the rate-distortion curves result from compressing each of the 10 modules (*SI Appendix*, Fig. S3). (*C*) Compressibility $\underline{C}$ of stochastic block networks versus the fraction of within-module edges f. (*D*) Compressibility $\underline{C}$ of stochastic block networks (colored points) and Erdös-Rényi networks (black point) versus transitivity (quantified by the average clustering coefficient). In *B*–*D*, data reflect averages over 50 randomly generated networks, each of size $N = 10^3$ and average degree $\langle k \rangle = 100$. (*E*) Compressibility $\underline{C}$ versus transitivity for the real networks in *SI Appendix*, Table S1 (20–23) with a linear best fit (dashed line). (*F*) Scale-free network, characterized by a power-law degree distribution and the presence of high-degree hubs. (*G*) Rate-distortion curves $\bar{R}(S)$ for Erdös-Rényi networks (black line) and scale-free networks (colored lines) with different scale-free exponents γ. (*H*) Compressibility $\underline{C}$ of scale-free networks versus the scale-free exponent γ. (*I*) Compressibility $\underline{C}$ of scale-free networks (colored points) and Erdös-Rényi networks (black point) versus degree heterogeneity h. In *G*–*I*, data reflect averages over 50 networks generated by using the static model (28), each of size $N = 10^3$ and average degree $\langle k \rangle = 100$. (*J*) Compressibility $\underline{C}$ versus degree heterogeneity for the real networks in *SI Appendix*, Table S1 (20–23) with a linear best fit (dashed line). In *E* and *J*, for networks of size $N > 10^3$, data points and error bars represent means and SDs over 50 randomly sampled subnetworks of 10^3 nodes each (*Materials and Methods*).

generalizes to real networks (Fig. 4*J*). Furthermore, we confirm that the dependencies of compressibility on both transitivity and degree heterogeneity extend to directed networks (*SI Appendix*, Fig. S5) and are robust to our choice of clustering heuristics (*SI Appendix*, Fig. S6).

The above results demonstrate that network compressibility increases with both transitivity and degree heterogeneity, the two defining features of hierarchical structure (29). Indeed, in networks with explicit hierarchical organization (such as those examined in ref. 29), we verify that optimal compressions capitalize on both modular structure and heterogeneous degrees in order to reduce the information rate (*SI Appendix*, Fig. S3). The high compressibility of hierarchical networks highlights a key distinction between lossy and lossless compression. In lossless compression, a network is more compressible if it has lower entropy $H(\boldsymbol{x})$, thereby admitting a more concise exact encoding (12, 13). The networks with the lowest entropies (and therefore the highest compressibilities from a lossless perspective) are those with homogeneous structure, such as Erdös-Rényi and k-regular networks (30). By contrast, lossy compression exploits structural regularities to remove redundant features of a network (Fig. 2), much like real-space renormalization (31). This direct coarse-graining renders hierarchical networks, which have strong structural regularities, highly compressible; similarly, it renders homogeneous networks, which have little to no structure, highly incompressible (Fig. 4 and *SI Appendix*, Fig. S3).

Finally, by focusing on specific families of networks, we discover variations in compressibility that reflect a network's specific function. Road networks, for example, exhibit the lowest transitivity and degree heterogeneity, and therefore the lowest compressibility, among the networks studied. This low compressibility is likely due to the fact that, unlike the other networks,

road networks are confined to exist in two dimensions, severely constraining their topology (32). Besides road networks, we find that protein interactions have the lowest transitivity and brain networks have the lowest degree heterogeneity, leading both classes of networks to be relatively incompressible. Interestingly, these two families are unique among the networks studied in that they are only encoded genetically and need not be represented cognitively by a human or animal. By contrast, language networks are highly compressible, perhaps reflecting the primary function of language as a means for encoding and communicating information. Thus, although many networks are encoded biologically, the pressure for these encodings to be efficient manifests to varying degrees in different families of networks, yielding a spectrum of compressibilities.

Discussion

Complex networks perform an astonishing array of functions, which are supported by a multitude of topological structures. Many networks, however, are unified by a common constraint: that they rely on biological entities to encode them and pass them on. Encoding a network efficiently—that is, striking an optimal balance between simplicity and accuracy—requires compression, an insight that has provided information-theoretic perspectives on network structure (14–16). Naturally, some networks should be more compressible than others, with structural regularities enabling efficient representations across multiple scales. To investigate this hypothesis, here, we introduce a rate-distortion theory of network compression (Fig. 1) and propose a quantitative definition for the compressibility of a network (Eq. **2**; Fig. 3*A*).

Applying our framework to a number of real and model networks, we demonstrate that network compressibility increases with both transitivity and degree heterogeneity (Fig. 4). Importantly, these two features are frequently observed across an array of real-world networks, from social, scientific, and biological interactions (29, 33, 34) to the internet (2), language (29), music (35), and the brain (36). Moreover, the combination of transitivity (with tightly connected modules) and heterogeneous degrees (with well-connected hubs) defines hierarchical organization (29), which has been shown to support multiscale representations of complex networks (37, 38) and enable efficient information processing in neural and communication systems (30, 39). In fact, when encoding information about the world, the brain itself often employs hierarchical representations (40–42). Our results lend to these perspectives an additional outlook on the role of hierarchical structure: that it supports the efficient compression of complex networks.

The interplay between network structure and compressibility opens the door for a number of future directions. For example, given that transitivity and heterogeneous degrees are nearly ubiquitous features of information, social, and biological networks (2, 29, 33–36), it is tempting to suspect that these networks have been shaped, at least in part, by the pressure to be compressed. Future work could directly address this hypothesis by investigating whether real-world networks, from language and music to protein interactions and the internet, have evolved over time to become more compressible. From a complementary perspective, one could develop methods for designing artificial networks that are optimally compressible. What might such optimally compressible networks look like? And how close to optimal are the networks that we observe in nature and society? The framework presented here provides the quantitative tools to begin answering these questions.

Materials and Methods

Entropy of Random Walks. Given a (possibly weighted, directed) network with adjacency matrix G_{ij}, the probability of one node i transitioning to another node j in a random walk is $P_{ij} = G_{ij}/k_i$, where $k_i = \sum_j G_{ij}$ is the (out) degree of node i (Fig. 1*A*). The entropy of random walks is given by

$$H(\boldsymbol{x}) = -\sum_i \pi_i \sum_j P_{ij} \log P_{ij}, \quad [5]$$

where π_i is the stationary distribution defined by the condition $\boldsymbol{\pi} = P^T\boldsymbol{\pi}$ (which we note is uniquely defined if the network is strongly connected and aperiodic). For undirected networks, Eq. **5** simplifies significantly. In this case, the stationary distribution is proportional to the node degrees $\pi_i = k_i/2E$, where $E = \frac{1}{2}\sum_{ij} G_{ij}$ is the number of edges in the network, and, thus, the entropy takes the form

$$H(\boldsymbol{x}) = \frac{1}{2E}\sum_i k_i \log k_i. \quad [6]$$

If, in addition, the nodes have uniform degree k (as in the k-regular networks in Fig. 3), then the entropy equals $\log k$. For example, in the simple network in Fig. 1, the nodes have uniform degree four, and thus the entropy is 2 bits.

Bounding the Information Rate. After clustering a network, a random walk $\boldsymbol{x} = (x_1, x_2, \ldots)$ gives rise to a new sequence $\boldsymbol{y} = (y_1, y_2, \ldots)$, where y_t is the cluster containing node x_t (Fig. 1*B*). The information rate of this sequence is given by the mutual information $I(\boldsymbol{x}, \boldsymbol{y})$, which for deterministic clusterings (such as those considered here) is equivalent to the entropy $H(\boldsymbol{y})$. However, even though the random walk $\boldsymbol{x}$ is Markovian (yielding a simple form for the entropy [Eq. **5**]), the clustered sequence $\boldsymbol{y}$ need not be (13), and, thus, it is generally difficult to derive an analytic form for $H(\boldsymbol{y})$.

Despite this hurdle, there exist simple bounds on the information rate $I(\boldsymbol{x}, \boldsymbol{y}) = H(\boldsymbol{y})$, summarized by the inequalities

$$H(y_{t+1} \mid x_t) \le H(\boldsymbol{y}) \le H(y_{t+1} \mid y_t), \quad [7]$$

where $H(y_{t+1} \mid x_t)$ and $H(y_{t+1} \mid y_t)$ are the conditional entropies of y_{t+1} on x_t and y_t, respectively (13). These bounds are tight at the minimum scale $S = 1/N$, when each cluster contains one node, and so $H(\boldsymbol{y}) = H(\boldsymbol{x}) = H(x_{t+1} \mid x_t)$. The bounds are also tight at the maximum scale $S = 1$, when there is one cluster, and so $H(\boldsymbol{y}) = H(y_{t+1} \mid x_t) = H(y_{t+1} \mid y_t) = 0$.

To compute the lower bound at intermediate scales, we begin with the conditional probability of node i in the random walk $\boldsymbol{x}$ transitioning to cluster c in the clustered sequence $\boldsymbol{y}$, $P_{ic} = \sum_{j\in c} P_{ij}$. Then, the lower bound is given by

$$I(\boldsymbol{x}, \boldsymbol{y}) \ge \underline{I}(\boldsymbol{x}, \boldsymbol{y}) = H(y_{t+1} \mid x_t) = -\sum_i \pi_i \sum_c P_{ic} \log P_{ic}, \quad [8]$$

where the second sum runs over all clusters c. Similarly, to compute the upper bound, we consider the probability of one cluster c transitioning to another cluster c',

$$P_{cc'} = \frac{1}{\pi_c}\sum_{i\in c} \pi_i \sum_{j\in c'} P_{ij}, \quad [9]$$

where $\pi_c = \sum_{i\in c} \pi_i$ is the stationary distribution over clusters. We then arrive at the following upper bound,

$$I(\boldsymbol{x}, \boldsymbol{y}) \le \overline{I}(\boldsymbol{x}, \boldsymbol{y}) = H(y_{t+1} \mid y_t) = -\sum_c \pi_c \sum_{c'} P_{cc'} \log P_{cc'}, \quad [10]$$

which is exact if the clustered sequence $\boldsymbol{y}$ is Markovian. In practice, when estimating the optimal information rate for a network, we minimize the upper bound in Eq. **10** over clusterings, resulting in an upper bound $\overline{R}(S)$ on the rate-distortion curve.

The upper bound $\overline{I}(\boldsymbol{x}, \boldsymbol{y})$ simplifies significantly for unweighted, undirected networks. In this case, the cluster transition probabilities take the form $P_{cc'} = G_{cc'}/k_c$, where $G_{cc'} = \sum_{i\in c}\sum_{j\in c'} G_{ij}$ is the induced network of clusters and $k_c = \sum_{i\in c} k_i$ is the sum of the degrees of the nodes in c. Recalling that the stationary distribution simplifies to $\pi_i = k_i/2E$, one can manipulate Eq. **10** into the form

$$\overline{I}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2E}\left[\sum_c k_c \log k_c - \sum_{cc'} G_{cc'} \log G_{cc'}\right]. \quad [11]$$

Under the further simplification of a clustering with one large cluster c and $n-1$ minimal clusters of one node each (Fig. 2), this upper bound can be fashioned into Eq. **1**.

Clustering Algorithm. To compute the rate-distortion curve $\bar{R}(S)$, we use an agglomerative clustering algorithm. Beginning with $n = N$ clusters (corresponding to the minimum scale $S = 1/N$), each containing an individual node, we iteratively combine pairs of clusters until we eventually arrive at one large cluster containing the entire network (corresponding to the maximum scale $S = 1$). At each step, we greedily select the pair of clusters to combine that minimizes the information rate $\bar{I}(\boldsymbol{x}, \boldsymbol{y})$ (Eq. **10**). However, rather than searching through all $\binom{n}{2}$ pairs of clusters at each iteration (which would limit applications to small networks), we instead focus on a subset of m pairs chosen through one of two heuristics.

The first heuristic, motivated by the observation that optimal clusterings tend to combine clusters with large degrees (Fig. 2*E*), selects the m pairs of clusters c and c' with the largest combined stationary probabilities $\pi_c + \pi_{c'}$. For unweighted, undirected networks, we note that this choice is equivalent to selecting the pairs of clusters with the largest combined degrees, since $\pi_c + \pi_{c'} = \frac{1}{2E}(k_c + k_{c'})$. The second heuristic, motivated by the fact that optimal compressions tend to form clusters with tight intracluster connectivity (Fig. 2*D*), selects the pairs of clusters c and c' with the largest combined joint transition probabilities $\pi_c P_{cc'} + \pi_{c'} P_{c'c}$. For unweighted, undirected networks, we remark that this second heuristic is equivalent to selecting the pairs of clusters with the largest number of connecting edges, since $\pi_c P_{cc'} + \pi_{c'} P_{c'c} = \frac{1}{2E}(G_{cc'} + G_{c'c})$. In practice, we consider $m = 100$ pairs of clusters at each iteration. In *SI Appendix*, Fig. S1, we compare these two heuristics to the brute-force approach that searches through all pairs of clusters at each iteration of the clustering algorithm. In addition to significantly speeding up the algorithm, we find that these two heuristics often yield more accurate estimates of the rate-distortion curve $R(S)$ than the brute-force implementation.

Network Datasets. The networks analyzed in this paper are listed and described in *SI Appendix*, Table S1 (20–23). While we study unweighted, undirected versions of the networks in Figs. 2, 3*E*, and 4 *E* and *J*, similar results hold for directed versions of the networks (*SI Appendix*, Figs. S2 and S3). For networks of size $N \leq 10^3$, we perform analyses directly. For larger networks with $N > 10^3$, we analyze 50 subnetworks of 10^3 nodes each. Each subnetwork is generated by performing a random walk beginning at a randomly selected node until 10^3 nodes have been reached. This sampling method has been shown to give accurate estimates of network statistics (43).

Data and Code Availability. The data analyzed in this paper and the code used to perform the analyses are openly available at GitHub (https://github.com/ChrisWLynn/Network_compressibility).

Citation Diversity Statement. Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minorities are undercited relative to the number of such papers in the field (44–49). Here, we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, and other factors. We obtained predicted gender of the first and last author of each reference by using databases that store the probability of a name being carried by a woman (48, 50). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 16% woman(first)/woman(last), 17% man/woman, 18% woman/man, and 50% man/man. This method is limited in that 1) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity; and 2) it cannot account for intersex, nonbinary, or transgender people. Second, we obtained the predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (51, 52). By this measure (and excluding self-citations), our references contain 9% author of color(first)/author of color(last), 14% white author/author of color, 15% author of color/white author, and 62% white author/white author. This method is limited in that 1) names, Census entries, and Wikipedia profiles used to make the predictions may not be indicative of racial/ethnic identity; and 2) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.

ACKNOWLEDGMENTS. We thank Christopher Kroninger, Dr. Lia Papadopoulos, Dr. Pragya Srivastava, Mathieu Ouellet, and Dale Zhou for helpful feedback on earlier versions of this manuscript. C.W.L. was supported by the James S. McDonnell Foundation 21st Century Science Initiative Understanding Dynamic and Multiscale Systems–Postdoctoral Fellowship Award. This work was also supported by the John D. and Catherine T. MacArthur Foundation; the Institute for Scientific Interchange Foundation; the Paul G. Allen Family Foundation; Army Research Laboratory Grant W911NF-10-2-0022; Army Research Office Grants Bassett-W911NF-14-1-0679, Falk-W911NF-18-1-0244, Grafton-W911NF-16-1-0474, and DCIST-W911NF-17-2-0181; the Office of Naval Research; National Institute of Mental Health Grants 2-R01-DC-009209-11, R01-MH112847, R01-MH107235, and R21-M MH-106799; and NSF Grants PHY-1554488, BCS-1631550, and NCS-FO-1926829.

1. A. E. Sizemore, E. A. Karuza, C. Giusti, D. S. Bassett, Knowledge gaps in the early growth of semantic feature networks. *Nat. Hum. Behav.* **2**, 682 (2018).
2. A. Vázquez, R. Pastor-Satorras, A. Vespignani, Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E* **65**, 066130 (2002).
3. X. F. Liu, K. T. Chi, M. Small, Complex network structure of musical compositions: Algorithmic generation of appealing music. *Physica A* **389**, 126–132 (2010).
4. M. Girvan, M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7821–7826 (2002).
5. E. R. Brush, D. C. Krakauer, J. C. Flack, Conflicts of interest improve collective computation of adaptive social structures. *Sci. Adv.* **4**, e1603311 (2018).
6. V. Kalakoski, P. Saariluoma, Taxi drivers' exceptional memory of street names. *Mem. Cognit.* **29**, 634–638 (2001).
7. C. W. Lynn, D. S. Bassett, How humans learn and represent networks. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29407–29415 (2020).
8. A.-C. Gavin *et al.*, Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
9. C. W. Lynn, D. S. Bassett, The physics of brain network structure, function and control. *Nat. Rev. Phys.* **1**, 318–332 (2019).
10. P. E. Vértes *et al.*, Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philos. Trans. R. Soc. B* **371**, 20150362 (2016).
11. K. J. Whitaker *et al.*, Adolescence is associated with genomically patterned consolidation of the hubs of the human brain connectome. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9105–9110 (2016).
12. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
13. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Hoboken, NJ, 2012).
14. M. Rosvall, C. T. Bergstrom, Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118–1123 (2008).
15. M. Rosvall, C. T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7327–7331 (2007).
16. N. Slonim, G. S. Atwal, G. Tkačik, W. Bialek, Information-based clustering. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18297–18302 (2005).
17. J. Leskovec, K. J. Lang, A. Dasgupta, M. W. Mahoney, Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.* **6**, 29–123 (2009).
18. E. Archer, I. M. Park, J. W. Pillow, Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy* **15**, 1738–1755 (2013).
19. W. W. Zachary, An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).
20. C. W. Lynn, D. S. Bassett, Network compressibility. Github. https://github.com/ChrisWLynn/Network_compressibility. Deposited 10 November 2020.
21. J. Kunegis, KONECT: The Koblenz network collection. KONECT. http://konect.cc/. Accessed 1 August 2020.
22. J. Leskovec, A. Krevl, Stanford Large Dataset Collection. SNAP. https://snap.stanford.edu/data. Accessed 1 August 2020.
23. V. Batagelj, A. Mrvar, Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/. Accessed 1 August 2020.
24. A. Buluç, H. Meyerhenke, I. Safro, P. Sanders, C. Schulz, "Recent advances in graph partitioning" in *Algorithm Engineering*, L. Kliemann, P. Sanders, Eds. (Lecture Notes in Computer Science, Springer, Cham, Switzerland, 2016), vol. 9220, pp. 117–158.
25. M. E. J. Newman, G. Reinert, Estimating the number of communities in a network. *Phys. Rev. Lett.* **117**, 078301 (2016).
26. A. R. Benson, D. F. Gleich, J. Leskovec, Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
27. A.-L. Barabási, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
28. K.-I. Goh, B. Kahng, D. Kim, Universal behavior of load distribution in scale-free networks. *Phys. Rev. Lett.* **87**, 278701 (2001).
29. E. Ravasz, A.-L. Barabási, Hierarchical organization in complex networks. *Phys. Rev. E* **67**, 026112 (2003).
30. C. W. Lynn, L. Papadopoulos, A. E. Kahn, D. S. Bassett, Human information processing in complex networks. *Nat. Phys.* **16**, 965–973 (2020).

31. E. Efrati, Z. Wang, A. Kolan, L. P. Kadanoff, Real-space renormalization in statistical mechanics. *Rev. Mod. Phys.* **86**, 647 (2014).
32. M. M. Sperry, Q. K. Telesford, F. Klimm, D. S. Bassett, Rentian scaling for the measurement of optimal embedding of complex networks into physical space. *J. Complex Netw.* **5**, 199–218 (2017).
33. M. Tomassini, L. Luthi, Empirical analysis of the evolution of a scientific collaboration network. *Physica A* **385**, 750–764 (2007).
34. E. Ravasz, "Detecting hierarchical modularity in biological networks" in *Computational Systems Biology*, R. Ireton, K. Montgomery, R. Bumgarner, R. Samudrala, J. McDermott, Eds. (Methods in Molecular Biology, Humana Press, Totowa, NJ, 2009), vol. 541, pp. 145–160.
35. M. M. Farbood, D. J. Heeger, G. Marcus, U. Hasson, Y. Lerner, The neural processing of hierarchical structure in music and speech at different timescales. *Front. Neurosci.* **9**, 157 (2015).
36. D. S. Bassett *et al.*, Hierarchical organization of human cortical networks in health and schizophrenia. *J. Neurosci.* **28**, 9239–9248 (2008).
37. M. Sales-Pardo, R. Guimera, A. A. Moreira, L. A. N. Amaral, Extracting the hierarchical organization of complex systems. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15224–15229 (2007).
38. M. Rosvall, C. T. Bergstrom, Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS One* **6**, e18209 (2011).
39. D. S. Bassett *et al.*, Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* **6**, e1000748 (2010).
40. J. Balaguer, H. Spiers, D. Hassabis, C. Summerfield, Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron* **90**, 893–903 (2016).
41. A. O. Diaconescu *et al.*, Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Comput. Biol.* **10**, e1003810 (2014).
42. K. Friston, Hierarchical models in the brain. *PLoS Comput. Biol.* **4**, e1000211 (2008).
43. J. Leskovec, C. Faloutsos, "Sampling from large graphs" in *KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computer Machinery, New York, NY, 2006), pp. 631–636.
44. S. M. L. Mitchell, S. Lange, H. Brus, Gendered citation patterns in international relations journals. *Int. Stud. Perspect.* **14**, 485–492 (2013).
45. M. L. Dion, J. L. Sumner, S. M. L. Mitchell, Gendered citation patterns across political science and social science methodology fields. *Polit. Anal.* **26**, 312–327 (2018).
46. N. Caplar, S. Tacchella, S. Birrer, Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nat. Astron.* **1**, 1–5 (2017).
47. D. Maliniak, R. Powers, B. F. Walter, The gender citation gap in international relations. *Int. Organ.* **67**, 889–922 (2013).
48. J. D. Dworkin *et al.*, The extent and drivers of gender imbalance in neuroscience reference lists. *Nat. Neurosci.* **23**, 918–926 (2020).
49. M. A. Bertolero *et al.*, Racial and ethnic imbalance in neuroscience reference lists and intersections with gender. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.10.12.336230 (Accessed 1 November 2020).
50. D. Zhou *et al.*, Diversity statement and code notebook (v1.1. 2020). https://github.com/dalejn/cleanBib. Accessed 1 November 2020.
51. A. Ambekar, C. Ward, J. Mohammed, S. Male, S. Skiena, "Name-ethnicity classification from open sources" in *KDD'09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2009), pp. 49–58.
52. G. Sood, S. Laohaprapanon, Predicting race and ethnicity from the sequence of characters in a name. arXiv [Preprint] (2018). https://arxiv.org/abs/1805.02109 (Accessed 1 November 2020).