Editors' Suggestion

Identifying hubs in directed networks

Alec Kirkley [*]

Institute of Data Science, University of Hong Kong, Hong Kong, China;
Department of Urban Planning and Design, University of Hong Kong, Hong Kong, China;
and Urban Systems Institute, University of Hong Kong, Hong Kong, China

(Received 13 December 2023; accepted 6 February 2024; published 20 March 2024)

Nodes in networks that exhibit high connectivity, also called "hubs," play a critical role in determining the structural and functional properties of networked systems. However, there is no clear definition of what constitutes a hub node in a network, and the classification of network hubs in existing work has either been purely qualitative or relies on ad hoc criteria for thresholding continuous data that do not generalize well to networks with certain degree sequences. Here we develop a set of efficient nonparametric methods that classify hub nodes in directed networks using the Minimum Description Length principle, effectively providing a clear and principled definition for network hubs. We adapt our methods to both unweighted and weighted networks, and we demonstrate them in a range of example applications using real and synthetic network data.

DOI: 10.1103/PhysRevE.109.034310

I. INTRODUCTION

Highly connected "hub" nodes play an important role in the structure and function of networks across a wide range of applications [1]. Hub regions in brain networks are central for communication and information integration [2,3]. Hub stations in transportation networks are important for resilience to failures and efficient routing [4,5]. Hub proteins in protein interaction networks are often essential for the survival and reproduction of an organism [6]. And hub locations in human mobility networks can be hot spots for congestion, economic activity, and the spread of disease [7,8].

The natural way to define a hub node in a network is to use degree centrality as an indicator—the more connections a node has, the more critical it is for network connectivity, and above some degree threshold (typically at or above the average degree in the network) we consider a node to be a "hub" [9]. Often the label of "hub" is reserved for nodes with an "unusually high" degree [1], as these nodes have a disproportionate influence on many processes that take place on the network, such as epidemics or information spreading [10,11]. The labeling of a node as a hub can be based on its in-degree and/or out-degree when the network is directed, depending on the application of interest. For example, in human mobility networks, one is often interested in targeting locations (nodes) with high population in-flows (weighted in-degree) for interventions to reduce congestion or the spread of disease, making the in-degree a relevant criterion for hub classification.

The concept of a hub node can also be extended to capture more global notions of centrality in a network. For example, the HITS algorithm [12] assigns a hub and authority score to each node in the network in a self-consistent manner: nodes are given a high authority score if they are pointed to by nodes with high hub scores, and nodes are given a high hub score if they point towards nodes with high authority scores. The stationary solution for the hub scores in the HITS algorithm are the entries in the leading eigenvector of the adjacency matrix of the network multiplied by its transpose, which indicates that these hub scores are capturing global information about the graph. Any number of other global centrality indices such as closeness, betweenness, eigenvector centrality, or Katz centrality can also in principle be used to define hub nodes at a global level, but ultimately many of these measures are highly correlated with degree in a large number of network models and real-world systems [13–18].

Identifying hub nodes as nodes with "unusually high" degrees can be thought of as performing outlier detection on the degree sequence listing the degree of each node in the network. But existing information-theoretic outlier detection methods either require the number of outliers to be known ahead of time [19], are formulated for general graph databases [20], or have parametric forms for the model likelihood which must be inferred [21], making them poorly suited for direct application to network degree sequences for classifying hub nodes.

The identification of hubs in weighted, directed networks also directly relates to the idea of identifying "hot spots" with high flows in human mobility networks, for which the "Loubar" method of [22] is an elegant and widely used method. The Loubar method utilizes the Lorenz curve of the flows in and/or out of nodes in a human mobility network to identify nodes with high flows as hot spots of activity. The Loubar method has been used in a range of applications to understand epidemic spreading [8,23], commuting structure [24], and economic growth [25] using human mobility data. This method has the desirable property of being completely nonparametric—it automatically selects the number of hot-spot nodes from the data itself, without any user-controlled input parameters. However, it does not specifically look at the pairwise nature of network structure in its formulation, so it cannot be compared with other network models using rigorous

[*] alec.w.kirkley@gmail.com

model selection criteria. The Loubar method also does not depend on the full distribution of flows, but only the mean and maximum of the flows, which we show in a range of experiments can lead to undesirable behavior where a large fraction of nodes in the network are classified as hubs.

Methods directly aimed at network compression using Bayesian or information-theoretic criteria—including blockmodeling [26], configuration models [27], core-periphery modeling [28], and other nonparametric information-theoretic methods for summarizing network structure [29–31]—are well-suited for identifying hubs in network data, since these methods allow for the automatic selection of the number of hubs and comparison with other network models using the MDL principle. However, none of the aforementioned methods explicitly aims to identify hub nodes, and so they may only identify hub nodes as a separate group by chance if grouping these nodes happens to provide compression with respect to the mixing structure assumed by the model. For example, high degree nodes are often grouped together in the (non-degree-corrected) stochastic blockmodel [32], and the core nodes in a core-periphery blockmodel are often of high degree [28]. To identify hub nodes, one only needs to consider the number of edges incident on the hub nodes, and not the connectivity among the hub nodes nor among the nonhub nodes which are key factors in the description length of a network under a blockmodel.

In this paper, we develop principled, efficient, nonparametric methods to classify hub nodes in directed networks based on the Minimum Description Length (MDL) principle, which states that the best model among a set of candidate models for a data set is the one that results in the lowest description length for the data [33]. In our approach, the MDL principle allows us to select the optimal configuration of hub nodes in a network by minimizing the description length of data encodings that exploit degree discrepancies between hub nodes and nonhub nodes. We adapt our formulation to multiple encoding schemes applicable to unweighted and weighted networks, and we describe a simple, fast algorithm to identify the hub nodes in these networks. We apply our method to a variety of synthetic network models, finding that the extent to which we can compress networks with hub nodes depends on the heterogeneity of the degree distribution and that our encodings can give more intuitive summaries of the hub structure than existing methods in these cases. We also apply our method to growing random graph models, finding that we can identify a hub transition at which it becomes most compressive to describe the network using hub nodes, and that this transition depends on the parameters of the growth model in a physically meaningful way. Finally, we apply our method to a corpus of directed network data sets from a wide range of disciplines, finding that many real networks do not have hub structure according to our more conservative encoding and that the information in many of the networks with a discernible hub structure can be effectively compressed by focusing on the high degree hub nodes when transmitting the network.

II. METHODS

To identify a particular type of structural and/or dynamical regularity in network data—for example, communities [26] or temporal change-points [30]—one can first construct an information encoding that is designed to exploit this regularity given an input data classification—for example, a partition of the nodes into communities or a segmentation of a time series of networks. Then, given an appropriate encoding, one can minimize its description length over all data classifications to find a representation that succinctly describes the data by exploiting the desired property. This process is equivalent to Maximum A Posteriori (MAP) estimation with hierarchical Bayesian generative models [26] but can often provide a more intuitive problem framing when choosing among various data encodings (models). In this section, we describe how to apply this line of reasoning for identifying hubs in network data.

A. Compressing network data

Let $G = (V, E)$ be a directed graph with N nodes in the node set V, and M edges in the edge set E. We will first treat the case in which G is a simple graph—in other words, G has no edges from a node to itself and has at most one edge going from any node i to any other node j. We will generalize our method to the multigraph case with self-edges in Sec. II C. We will let $\boldsymbol{k}$ be the in-degree sequence such that k_i is the in-degree of node i and $\sum_{i=1}^{N} k_i = M$.

Now, suppose we aim to transmit the network G—or, equivalently, the source node $i \in V$ and target node $j \in V$ of all the edges $(i, j) \in E$—in binary to a receiver through some communication channel. We assume that the receiver knows N and M, which would be of comparatively negligible cost to communicate and can be ignored anyway. Since there are $N(N-1)$ distinct ordered node pairs, and M of these pairs contains an edge, then there are $\binom{N(N-1)}{M}$ possible graphs G, and the same number of possible binary messages we could end up transmitting to the receiver. $\lceil \log_2 \binom{N(N-1)}{M} \rceil$ bits will be enough to encode all such messages uniquely when establishing a code book for our transmission ahead of time with the receiver, and so the information content or *description length* of this naive encoding of the graph G is

$$\mathcal{L}_0^{(\mathrm{ERs})} = \log \binom{N(N-1)}{M} \tag{1}$$

bits. Here we have ignored the ceiling function as it will provide a negligible change to the description length for $N \gg 1$, we have used the notation convention $\log_2 \equiv \log$, and we have used the subscript "0" to indicate that no hub nodes were used to aid in the network transmission process—we will describe how this works shortly. We use the superscript "ERs" to refer to the Erdos-Renyi model for simple directed graphs, $G(N, M)$, which selects uniformly at random from all directed simple graphs with N nodes and M edges. We can also derive Eq. (1) as the negative log-probability of picking any particular graph G from this ensemble.

Alternatively, we can transmit the graph G using a multipart encoding, where each step involves utilizing a different code book with the receiver that is constructed while keeping in mind the constraints imposed by information transmitted earlier in the transmission process. This process involves transmitting G in increasing levels of granularity until the entire edge set E is known.

One very simple multipart transmission scheme would involve first transmitting the in-degrees $\boldsymbol{k} = \{k_1, \ldots, k_N\}$, and then transmitting the entire edge set E from the set of all edge sets consistent with the in-degrees $\boldsymbol{k}$. There are at most $\left(\!\binom{N}{M}\!\right) = \binom{M+N-1}{N-1}$ (where (( )) is the multiset coefficient) unique ways to assign the N nodes as targets in the M directed edges—allowing nodes to potentially have in-degree 0—in order to fully specify the in-degrees $\boldsymbol{k}$. This is the total number of messages that will be in the code book for the first step of the process. (Depending on M and N, not all of the $\binom{M+N-1}{N-1}$ in-degree sequences will correspond to valid directed simple graphs, so we are technically accounting for some "nongraphical" in-degree sequences in our encoding in addition to all valid in-degree sequences.) Then, once $\boldsymbol{k}$ is known, there are $\binom{N-1}{k_i}$ possible source nodes for the k_i edges containing node i as a target, resulting in $\prod_{i=1}^{N}\binom{N-1}{k_i}$ possible messages for the second step of the process. Adding the information content of these two steps, the description length of this alternative encoding is

$$\mathcal{L}_0^{(\mathrm{CMs})} = \log\binom{M+N-1}{N-1} + \sum_{i=1}^{N}\log\binom{N-1}{k_i} \tag{2}$$

bits. We use the superscript "CMs" to refer to the Configuration Model—with only in-degree constraints—for simple directed graphs, and again use a subscript "0" to indicate that no hub nodes were used to aid in the network transmission process. This model will select uniformly at random from all directed simple graphs with N and M edges and a specific in-degree sequence $\boldsymbol{k}$. We can alternatively derive Eq. (2) as the negative log-probability of picking any particular graph G from this ensemble, given a uniform prior over all degree sequences $\boldsymbol{k}$ of length N that sum to M.

One can show that for $N \gg \langle k\rangle \gg 1$ we will always achieve superior data compression—in other words, a lower description length for the network G—using the two-step encoding corresponding to Eq. (2) compared with the one-step encoding corresponding to Eq. (1) (see Appendix A). This suggests that multistep encodings that exploit in-degree heterogeneity can provide improved network compression.

B. Compressing simple directed graphs with hub nodes

We can consider going one step further and transmitting the edges incident to a small set of high in-degree nodes—the "hub" nodes in the encoding—independently from the rest of the edges in the network. This will provide good compression when a large portion of the edges in E are pointing towards a small set of hub nodes, because in this case there are comparatively few potential configurations of edges incident to the hubs—there are not many target nodes to choose from—and the total number of binary messages needed for encoding the specific edge configuration E is reduced.

Consider the case in which we have h hub nodes $V_h \subseteq V$ and $N-h$ nonhub nodes $V \setminus V_h$, and we wish to transmit the edges incident to these two node sets independently to achieve improved data compression for the total edge set E. (We will address the issue of identifying the optimal value of h later on.) Let E_h denote the set of edges incident to the hub nodes V_h as targets, and $M_h = |E_h|$ be the number of such edges. Since there are N possibilities for the value of h, and M possibilities for the value of M_h, we will need $\log N + \log M = \log NM$ bits of information to transmit these initial quantities.

Next we need to transmit the identities of the h hub nodes V_h, which will require $\log\binom{N}{h}$ bits of information since it requires specifying a subset of h nodes from the total set of N nodes. Then, we can perform two transmission steps reminiscent of the one used to derive Eq. (1)—one for the hub-incident edges E_h, and one for the rest of the edges $E \setminus E_h$. Specifying E_h will require us to specify M_h edge positions out of $h(N-1)$ total (ordered) node pairs, and specifying $E \setminus E_h$ will require us to specify $M - M_h$ edge positions out of the $(N-h)(N-1)$ remaining node pairs. The total information cost of this encoding is then given by

$$\begin{aligned}\mathcal{L}^{(\mathrm{ERs})}(V_h) = \log NM + \log\binom{N}{h} + \log\binom{h(N-1)}{M_h}\\ + \log\binom{(N-h)(N-1)}{M-M_h},\end{aligned} \tag{3}$$

where we have used the "ERs" notation as before to denote the use of an encoding that transmits the edge positions for each node set in one step, as in Eq. (2). We have also made explicit that this description length of our hub-based encoding is a function of the number of hubs h that we choose and which h nodes V_h we choose to be the hubs. For the cases $h = 0$ and $h = N$, we let $\mathcal{L}^{(\mathrm{ERs})} = \mathcal{L}_0^{(\mathrm{ERs})}$, since the initial transmission costs are no longer needed.

We can view Eq. (3) as an objective function that, when minimized over node subsets V_h, finds an optimal set of nodes to classify as hubs in our network. In other words, the optimal subset of nodes to identify as hubs is the subset of nodes that allows us to most parsimoniously describe (i.e., best compress) the network structure using an encoding aimed at exploiting structural heterogeneity between hubs and nonhubs. In Appendix B we show that the optimal choice for the hub nodes at any given value of h is the set of nodes that maximizes M_h—in other words, the nodes with the h highest in-degrees—which confirms the intuition behind the construction of the hub-based encoding. This implies that we can identify the globally optimal configuration of hub nodes in a network G using the following simple algorithm:

(i) Order the node indices in G so that $k_1 \geqslant k_2 \geqslant \cdots \geqslant k_N$.

(ii) Initialize $V_h \leftarrow \{\}$, $h \leftarrow 0$, $M_h \leftarrow 0$, $h^*_{\mathrm{ERs}} \leftarrow 0$, and set $\mathcal{L}^*_{\mathrm{ERs}} \leftarrow \mathcal{L}_0^{(\mathrm{ERs})}$ using Eq. (1).

(iii) For $i \in \{1, \ldots, N\}$:

  (a) Add i to V_h and update $h \leftarrow h+1$, $M_h \leftarrow M_h + k_i$.

  (b) Compute the new description length $\mathcal{L}_h = \mathcal{L}^{(\mathrm{ERs})}(V_h)$.

  (c) If $\mathcal{L}_h < \mathcal{L}^*_{\mathrm{ERs}}$, set $\mathcal{L}^*_{\mathrm{ERs}} \leftarrow \mathcal{L}_h$ and $h^*_{\mathrm{ERs}} \leftarrow h$. Otherwise, do nothing.

(iv) After the loop terminates, the optimal set of hubs will be the set of node indices $\{1, \ldots, h^*_{\mathrm{ERs}}\}$, and these hubs will result in a description length of $\mathcal{L}^*_{\mathrm{ERs}}$ in Eq. (3).

In the case of ties—i.e., if at a certain degree cutoff k^* it is information theoretically optimal to only consider some fraction $0 < f < 1$ of the nodes i with $k_i = k^*$ as hubs—we will check the cases $f = 0$ (include all nodes with $k_i \geqslant k^* + 1$) and $f = 1$ (include all nodes with $k_i \geqslant k^*$) and choose the case with the lower description length. One can alternatively add all nodes with each unique degree k at once during the greedy optimization process, to ensure that all nodes at or above the optimal threshold degree k^* are included as hubs. These modifications remove the need to randomly break the tie to assign only some nodes of degree k^* as hubs, since all such nodes are treated equivalently in our scheme. In principle, one can ignore this step, and a random subset of nodes of degree k^* will be chosen as hubs based on the initial node ordering. This will often result in a slight compression gain at the cost of arbitrarily choosing the lowest-degree hubs.

The above method has an $O(N \log N)$ time complexity if the in-degrees $\boldsymbol{k}$ of the network G are known, the bottleneck being sorting the degree sequence. Therefore, it is equally as simple in practice to compute as hot-spot identification methods such as the Loubar method and average degree method discussed in [22]. This method will also select for the number of hubs h^* automatically based on the number that results in the best compression: too few hubs means we have not fully exploited the heterogeneity of the hub in-degrees for compression, and too many hubs means we have too little separation in the in-degrees of hubs and nonhubs to provide any meaningful compression. One key advantage of this approach over existing methods is that it allows for the result $h^* = 0$ if there is no information-theoretic justification to include any nodes as hubs according to Eq. (3). We will see in Sec. III that this situation is quite common for networks with homogeneous in-degree distributions.

There are a number of alternative ways we can construct an encoding that exploits a hub/nonhub dichotomy and allows for identifying an optimal set of network hubs. For example, one can transmit the hub nodes' incident edges individually using a two-step encoding inspired by the one used to compute Eq. (2), but transmit the remaining edges $E \setminus E_h$ using the same one-step encoding as in Eq. (1). This results in a description length of

$$\mathcal{L}^{(\mathrm{CMs})}(V_h) = \log NM + \log\binom{N}{h} + \log\binom{M_h + h - 1}{h - 1} + \sum_{i \in V_h} \log\binom{N-1}{k_i} + \log\binom{(N-h)(N-1)}{M - M_h}. \tag{4}$$

Here we also use the convention $\mathcal{L}^{(\mathrm{CMs})} = \mathcal{L}_0^{(\mathrm{CMs})}$ for $h = 0, N$ for convenience.

In Appendix C, we show that this description length can be optimized over hub node sets V_h using an analogous greedy algorithm, but in this case we only have a guarantee of local optimality. (One can in principle simply enforce the constraint that any hub node in-degree must be greater than or equal to any nonhub node in-degree, in which case this distinction is irrelevant since we will always add nodes in decreasing order of in-degree regardless of their effect on the description length.)

After running the hub identification algorithm using either the ERs or CMs encoding, we can determine the extent to which a hub/nonhub dichotomy allows us to compress the network data G in the first place—this gives us an alternative aggregate measure of heterogeneity in a network's in-degree distribution. To do this, we compare the optimal level of compression $\mathcal{L}^*$ achieved with the hub-based encoding (either ERs or CMs) to the baseline compression levels in Eqs. (1) and (2) when no hubs are utilized. The resulting quantity, which we call the inverse compression ratio, is given by

$$\eta^{(s)} = \frac{\mathcal{L}^*}{\max\big(\mathcal{L}_0^{(\mathrm{ERs})}, \mathcal{L}_0^{(\mathrm{CMs})}\big)}. \tag{5}$$

We can see that $\eta^{(s)} \in [0, 1]$, since each encoding will give a minimum description length $\mathcal{L}^*$ that is at least as low as the description length for the encoding with no hub nodes. An inverse compression ratio near 0 indicates that the hub nodes account for a large portion of the in-degrees and provide efficient compression of the network data, while an inverse compression ratio near 1 indicates that the hubs do not contribute a significant portion of the in-degrees of the network. The case $\eta^{(s)} = 1$ occurs when we achieve no compression using hubs, which happens when the optimal number of hubs according to the encoding is $h^* = 0$.

As was done with Eqs. (1) and (2), we show in Appendix D that the compression of the encoding corresponding to Eq. (4) is generally better than the encoding corresponding to Eq. (3) in the high in-degree regime. In applications with networks G that do not satisfy the conditions of Appendix D, one can perform model selection among the two hub-based encodings [corresponding to Eqs. (3) and (4)] by identifying which description length is smaller. The more compressive encoding can then be used as the method for identifying the hub nodes in G.

In addition to the in-degrees, one can consider using the *out*-degrees to aid in network compression. This amounts to the same process except it identifies nodes with high out-degrees rather than in-degrees. This characterization is sensible, for example, in applications aiming to identify potential "superspreaders" in epidemic and misinformation modeling [34,35]. We explore both the in- and out-degree versions of these methods in Sec. III.

C. Compressing directed multigraphs and weighted networks with hub nodes

One can also extend the method discussed in Sec. II B to directed multigraphs or directed, weighted networks with non-negative integer-valued weights. Such networks often arise in applications involving population "flows" from node to node, for example in transportation and human mobility modeling [4,5,7,8], which were the original motivating examples for the Loubar method of [22]. In this case, both the baseline encodings corresponding to Eqs. (1) and (2) as well as the hub-based encodings corresponding to Eqs. (3) and (4) must be modified to account for the potential of having more than one edge (or, equivalently, an edge weight greater than 1) between each pair of nodes. We will now let M be the total weight of all the edges, and k_i be the total weight of edges incident on node i—the latter is sometimes called the "in-strength" of node i, but we will continue to use the term "in-degree" for consistency with the simple graph case. We then have the same relation $M = \sum_{i=1}^{N} k_i$ as in the simple graph case.

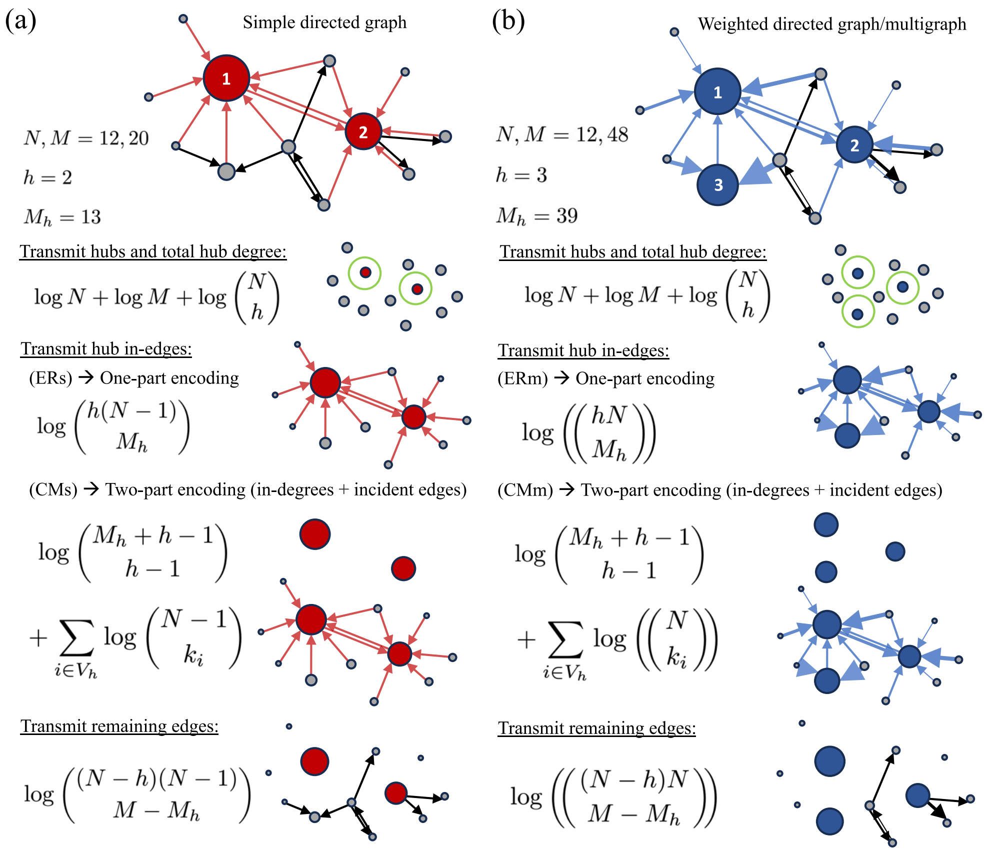
(a)
Simple directed graph
$N, M = 12, 20$
$h = 2$
$M_h = 13$
Transmit hubs and total hub degree:
$\log N + \log M + \log \binom{N}{h}$
Transmit hub in-edges:
(ERs) → One-part encoding
$\log \binom{h(N-1)}{M_h}$
(CMs) → Two-part encoding (in-degrees + incident edges)
$\log \binom{M_h + h - 1}{h - 1}$
$+ \sum_{i \in V_h} \log \binom{N-1}{k_i}$
Transmit remaining edges:
$\log \binom{(N-h)(N-1)}{M - M_h}$
(b)
Weighted directed graph/multigraph
$N, M = 12, 48$
$h = 3$
$M_h = 39$
Transmit hubs and total hub degree:
$\log N + \log M + \log \binom{N}{h}$
Transmit hub in-edges:
(ERm) → One-part encoding
$\log \left(\!\binom{hN}{M_h}\!\right)$
(CMm) → Two-part encoding (in-degrees + incident edges)
$\log \binom{M_h + h - 1}{h - 1}$
$+ \sum_{i \in V_h} \log \left(\!\binom{N}{k_i}\!\right)$
Transmit remaining edges:
$\log \left(\!\binom{(N-h)N}{M - M_h}\!\right)$
1
2
3

FIG. 1. Diagram of hub-based encodings. (a) Schematic of the simple directed graph encoding described in Sec. II B, along with the description length of each step. (b) Schematic of the weighted directed graph/multigraph encoding described in Sec. II C, along with the description length of each step. The equations above use the convention $\log \equiv \log_2$ as in the remainder of the text. We define the hub nodes of a network $G = (V, E)$ according to a given encoding (ERs, CMs, ERm, or CMm) as the node subset $V_h \subseteq V$ that minimizes the information required to transmit the network (e.g., the positions of the edges E) when transmitting the positions of edges incident to the hubs first. This provides a principled, nonparametric criterion for identifying hubs in directed networks based on the Minimum Description Length (MDL) principle.

The resulting description lengths, for which we use the superscripts "m" to denote "multigraph," are given by

$$\mathcal{L}_0^{(\mathrm{ERm})} = \log \left(\!\binom{N^2}{M}\!\right), \tag{6}$$

$$\mathcal{L}_0^{(\mathrm{CMm})} = \log \binom{M+N-1}{N-1} + \sum_{i=1}^{N} \log \left(\!\binom{N}{k_i}\!\right), \tag{7}$$

$$\mathcal{L}^{(\mathrm{ERm})}(V_h) = \log NM + \log \binom{N}{h} + \log \left(\!\binom{hN}{M_h}\!\right) + \log \left(\!\binom{(N-h)N}{M-M_h}\!\right), \tag{8}$$

$$\mathcal{L}^{(\mathrm{CMm})}(V_h) = \log NM + \log \binom{N}{h} + \log \binom{M_h+h-1}{h-1} + \sum_{i \in V_h} \log \left(\!\binom{N}{k_i}\!\right) + \log \left(\!\binom{(N-h)N}{M-M_h}\!\right). \tag{9}$$

These can be derived by transforming the binomial coefficients in Eqs. (1), (2), (3), and (4) to multiset coefficients $\left(\!\binom{n}{k}\!\right)$, which count the number of ways to place k edges into n edge positions while allowing for repetitions. We also allow for self-edges, which requires the substitution $N-1 \to N$ for the number of valid edge positions incident on a single node and $N(N-1) \to N^2$ for the total number of potential edge positions. We also use the conventions $\mathcal{L}^{(\mathrm{ERm})} = \mathcal{L}_0^{(\mathrm{ERm})}$ and $\mathcal{L}^{(\mathrm{CMm})} = \mathcal{L}_0^{(\mathrm{CMm})}$ for $h \in \{0, N\}$, as for the simple graph description lengths.

Equations (8) and (9) can be optimized using the same greedy procedure as in Sec. II B for identifying the hubs from Eq. (3) (with appropriate transformation of the variables). The resulting hub identification schemes can be used to construct an inverse compression ratio analogous to Eq. (5), thus

$$\eta^{(m)} = \frac{\mathcal{L}^*}{\max\left(\mathcal{L}_0^{(\mathrm{ERm})}, \mathcal{L}_0^{(\mathrm{CMm})}\right)}, \tag{10}$$

where $\mathcal{L}^*$ is the minimum description length achieved with the method of interest (ERm or CMm). Similar to the simple graph case, these ratios are bounded within [0,1] and equal 1 in the extreme case when $h^* = 0$. $\mathcal{L}^*$ can be compared across the ERm and CMm encodings to select the more compressive encoding for a given multigraph G.

In Fig. 1 we show a schematic summarizing the hub-based encodings described in the last two sections. Code implement-

TABLE I. Four methods for identifying hub nodes in Sec. III. All methods can be applied to in-degrees/out-degrees as well as unweighted/weighted networks by defining the degrees $\boldsymbol{k}$ accordingly.

Name	Description
ER	Identifies hub nodes by iterating over $\boldsymbol{k}$ and greedily adding hubs to minimize the description length in Eq. (3) (for simple graphs) or Eq. (8) (for multigraphs and integer-weighted graphs).
CM	Identifies hub nodes by iterating over $\boldsymbol{k}$ and greedily adding hubs to minimize the description length in Eq. (4) (for simple graphs) or Eq. (9) (for multigraphs and integer-weighted graphs).
Average	Identifies hub nodes as all nodes $i \in V$ such that $k_i \geqslant \langle k \rangle = M/N$.
Loubar [22]	Identifies hub nodes as the nodes at or above the $[1 - \frac{\langle k \rangle}{\max(\boldsymbol{k})}]$th quantile in terms of degree.

ing the hub identification methods of this paper can be found in Ref. [36].

III. RESULTS

A. Hubs in networks with specified degree sequences

We first perform a range of experiments with synthetic network data in order to understand the conditions under which hub nodes will be found using the methods described in Sec. II. We begin by analyzing the hub properties of random networks with degree sequences $\boldsymbol{k}$ that vary in magnitude and variability. Since the hubs and compression achieved using our proposed methods only depend on the degree sequence, we may interpret $\boldsymbol{k}$ in these experiments as the in-degree sequence of a network that is otherwise completely randomized, and we need not actually generate any network for each simulation.

We choose three discrete probability distributions from which we generate the in-degrees $\boldsymbol{k}$: (i) A Poisson distribution, whose relative variance Variance/Mean$^2 = 1/\mu$ will vanish for large mean degree $\mu = \langle k \rangle$; (ii) a geometric distribution, whose relative variance $1 - \mu^{-1}$ will tend to a constant for a large mean degree; and a power-law (e.g., Zipf) distribution, whose relative variance can potentially diverge. To ensure that all generated degree sequences $\boldsymbol{k}$ are graphically realizable, we consider the generated graphs to be multigraphs (or, equivalently, integer-weighted graphs) and use the encodings corresponding to Eqs. (8) and (9).

We compare our methods with two widely used methods for identifying hub nodes ("hot spots") using the weighted in-degrees of human mobility networks [8,22,24,25]. The first is the "average" method, which simply classifies all nodes with in-degree values higher than the network average M/N as hubs. However, the average method may not be conservative enough to give a useful guide for locating hot spots in human mobility applications, so the "Loubar" method is proposed in [22]. It uses the Lorenz curve to construct a threshold for hub nodes that depends on the mean in-degree and the maximum in-degree. Our measures and these alternative measures are summarized in Table I for convenience. There is no absolute way in which one can decide which of these measures is "best"—this may depend on the application of interest, and will require the consideration of multiple aspects of each measure—but our experiments highlight some potential intuitive advantages of the approaches based on the MDL principle.

In Fig. 2 we show the results of applying all four methods to degree sequences $\boldsymbol{k}$ randomly generated from the three distributions described above, for average in-degrees $\langle k \rangle \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$ and network sizes $N \in \{10^3, 10^5\}$. Each distribution only has a single parameter and can be specified uniquely given the desired mean in-degree. In each simulation, we generate $\boldsymbol{k}$ from the specified distribution, apply the four methods in Table I, and take the average result for the optimal number of hubs h^* and inverse compression ratio (for the ERm and CMm encodings) over 50 repeated trials. For easier visualization, we plot h^*/N to see what fraction of nodes are classified as hubs using each method.

In Fig. 2(a), we observe that for the Poisson-distributed in-degrees, the ER and CM encodings both find very few hubs—the ER encoding always finds zero hubs, while the CM encoding only finds a handful. This is consistent with the small relative variance of the Poisson distribution, which will rarely result in any nodes having substantially larger in-degrees than the rest of the nodes. On the other hand, we can see that the average and Loubar methods both indicate many hubs for networks with Poisson-distributed in-degrees. The average method produces many hubs because the distribution is relatively symmetric, so a little less than half of the nodes will have in-degrees above the average. The Loubar method also produces many hubs because by construction it will indicate that the fraction of nodes that are hubs is $\langle k \rangle/\max(\boldsymbol{k})$, and this quantity will be nearly equal to 1 for Poisson samples with large mean in-degrees. We can see that in general the mean degree and network size do not play a particularly important role, as the fraction of hubs h^*/N from each method is fairly constant across all simulation settings. The exception is the CM encoding, which produces only a handful (less than 10) of hubs for most simulations, which results in smaller values of h^*/N for larger N.

In Fig. 2(b), we can see that for Poisson-distributed in-degrees, neither method (ERm or CMm) provides any substantial compression of the network, as the inverse compression ratios are approximately equal to 1. This is because there is very little heterogeneity in the in-degrees that a hub-based information encoding can exploit to reduce the transmission cost of the data to a receiver. We also observe that the ER encoding has a slight edge over the CM encoding in terms of compression (indicated by the outlined markers), because transmitting the edges incident to hub nodes separately incurs an extra initial transmission cost but provides negligible additional compression.

In Figs. 2(c) and 2(d), we see a different story for geometrically distributed in-degrees. In this case, the ER and

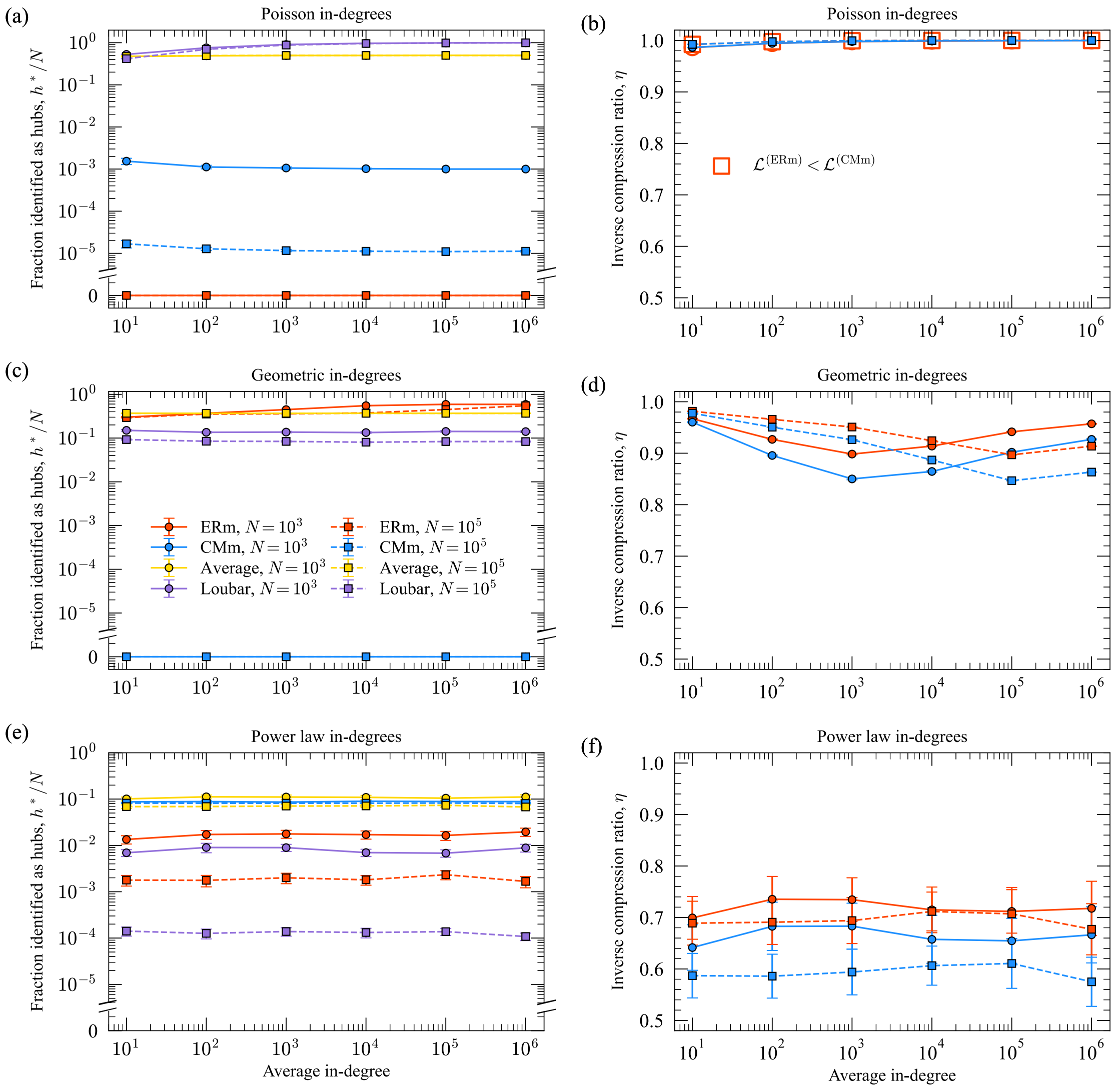

FIG. 2. Identifying hubs in networks with different in-degree distributions. (a) Fraction of nodes h^*/N identified as hubs using the four methods detailed in Table I, for Poisson-distributed weighted in-degrees. Experiments were performed over a broad range of average in-degree for $N = 10^3$ (solid lines) and $N = 10^5$ (dotted lines). (b) Inverse compression ratio η [Eq. (10)] for the ERm and CMm methods over the same set of experiments. The experiments were repeated for geometrically distributed in-degrees [panels (c) and (d)] and power–law (Zipf) distributed in-degrees [panels (e) and (f)], which exhibit progressively higher levels of relative variance. Error bars indicate two standard errors in the mean over 50 generated in-degree distributions, and large circles/squares around the data points in panel (b) indicate configurations for which the ERm model provided superior compression to the CMm model.

average methods indicate a substantial fraction of hub nodes, while the Loubar method is more conservative and only indicates that roughly 10% of nodes are hubs. Meanwhile, the CM encoding is still very conservative, this time classifying zero nodes as hubs in all cases. Here the CM encoding now compresses better than the ER encoding, and both methods provide substantially better compression than in the Poisson case. This is consistent with the greater relative variance of the geometric distribution, which will result in more effective compression using hub-based methods, and will benefit in particular from transmitting edges incident to the high degree hubs individually. There is also a greater dependence on the size N of the network, and optimal compression for both encodings appears to be achieved at roughly $\langle k \rangle \approx N$.

In Figs. 2(e) and 2(f), we plot the same results for the power-law in-degrees, which exhibit different behavior from the first two cases. Here we can see that the CM encoding is identifying a more substantial number of hubs (roughly 10% of all nodes), while the ER encoding is a bit more conservative and the Loubar method is the most conservative. In this case one may expect a greater fraction of nodes to be identified as hubs than for Poisson or geometric in-degrees, due to the highly skewed nature of the power-law distribution, which would suggest that the CM results overall are the most consistent with our expectations. Both the ER and CM encodings are most compressive for power-law in-degrees, reducing the information needed to transmit the network by roughly 30–40 %, and the CM encoding becomes even more heavily favored in terms of the inverse compression ratio. We also see a greater dependence on network size N and higher sample variability in the power-law results, as expected from the diverging relative variance in many cases.

Overall, these experiments suggest that the CM encoding of Sec. II is performing the most consistent with intuition for these randomized degree sequences, as it finds zero or only a handful of hub nodes for Poisson and geometric degree sequences, while finding that roughly 10% of nodes are hubs for the power-law degree sequences. Meanwhile, the ER method appears to be quite lenient for classifying hubs with geometrically distributed in-degrees, but identifies a sensible number of hubs for the Poisson and power-law cases—in particular, it identifies zero hubs for Poisson-distributed in-degrees. On the other hand—in contrast with our intuition—, the average and Loubar methods identify fewer hubs as the relative variance increases (from Poisson to geometric to power-law in-degrees).

B. Emergence of hubs in growing networks

Here we examine the four methods of Table I in a more dynamic context, applying these methods to track the evolution of hub nodes in growing networks. To simulate growing networks with varying levels of degree heterogeneity, we use the Price model for citation dynamics [37] with a variable attachment exponent. In this model, at each time step $t = 1, \dots, T$ a new node i arrives with an out-degree m and attaches each out-edge to an existing node j with probability

$$q_{i\to j}^{(t)} = \frac{\left(k_j^{(t)} + 1\right)^{\alpha}}{\sum_{j\in V^{(t)}} \left(k_j^{(t)} + 1\right)^{\alpha}}, \tag{11}$$

where $k_j^{(t)}$ is the in-degree of node j at time t, and $V^{(t)}$ are the nodes in the network at time t (excluding i). The model begins with m seed nodes of degree zero, which are the targets for the first incoming node's out-edges. In [38] they find that a similar model with $m = 1$ results in stretched exponential in-degree distributions for sublinear attachment ($\alpha < 0$) as $T \to \infty$, while for $\alpha > 1$ the single initial node becomes a hub that has connections from nearly every other node that is added. Meanwhile, for $\alpha = 1$ the in-degrees follow a power law asymptotically. These results suggest that as $\alpha \to 0$ we should see fewer and fewer hub nodes, while for $\alpha > 1$ we should see $h \approx m$ as $T \to \infty$ since nearly every connection is made to a seed node. On the other hand, for $\alpha = 1$ we expect to see a power-law distribution of in-degrees and will observe a point in time at which a large number of hubs emerge.

In Fig. 3(a), we show the average number of hubs h^* detected by the four methods as a function of the number of time steps t, for 50 simulations of the growth model in Eq. (11) with $T = 100$ and $\{m, \alpha\} = \{1, 0\}$. Under this parameter configuration, the network is a Random Recursive Tree, which will exhibit a highly homogeneous degree distribution and produce high degree hub nodes with very low probability [39]. We can see that in this case, the ER and CM methods identify very few hubs—the CM method identifies exactly zero hubs in all simulations—while the average and Loubar methods classify a sizable number of nodes as hubs (which increases steadily as the network grows). We do not see any sharp transition at which hubs emerge in the network according to any of the four methods.

In Fig. 3(b), we show the growth model simulation results for the sublinear case of $\{m, \alpha\} = \{18, 0.5\}$. We find again that the average and Loubar methods find a steadily increasing number of hubs starting with $h^* \approx m$, with some small oscillations in h^* for the average method due to slight changes in the number of nodes with in-degree above the average of m. The ER and CM methods in this case exhibit quite different trends from those for the Random Recursive Tree in panel (a). The ER curve exhibits a sharp phase transition-like jump at $t \approx 5$, then steadily increases as the network grows further. This indicates that, as expected, after roughly five attachment events it is information theoretically more efficient to describe the network using hub nodes, according to Eq. (3). The CM encoding, typically being more conservative in its classification of hubs, does not begin to find hubs in the network until much farther into the simulations at around $t \approx 60$. By $T = 100$, the CM method typically detects around $h^* \approx 18$–20 hubs, often finding that the seed nodes have high enough in-degrees to justify their existence as hubs under the CM encoding.

In Fig. 3(c), we increase the attachment exponent so that linear preferential attachment will occur with $\{m, \alpha\} = \{10, 1\}$, producing a power-law in-degree distribution asymptotically. Here we see largely the same trend as in Fig. 3(b) for the ER, average, and Loubar methods, with slower growth rates in h^* as the simulations progress. This is consistent with a greater fraction of the in-connections being concentrated at the seed nodes, whose early existence in the network has given them a strong cumulative advantage. The CM encoding now classifies a substantial fraction of nodes as hubs, and displays sharp phase transition-like behavior at $t \approx 20$. This is consistent with the behavior seen in Fig. 2(e) for the Zipf-distributed in-degrees, where the CM encoding identified many hubs.

In Fig. 3(d), we increase the attachment exponent to lie in the superlinear regime with $\{m, \alpha\} = \{4, 2.7\}$. In this case, all curves converge to $h^* \approx m = 4$, which reflects the fact that most incoming edges will attach to the four seed nodes due to the superlinear attachment process. We again see the ER and CM encodings producing a sharp hub transition, but at even earlier time steps, while the average and Loubar methods produce smooth curves as before.

Finally, in Figs. 3(e) and 3(f) we plot the hub transition—the time step t at which $h^* = 1$ in expectation over the simulations—for the ER and CM encodings, at various values of the simulation parameters $\{m, \alpha\}$. The four parameter configurations corresponding to panels (a)–(d) are indicated by small white squares. We can see that for no attachment preference ($\alpha = 0$) or a single outgoing edge ($m = 1$), the ER model indicates a late hub transition, which we find is due to a smooth ascent reminiscent of that in Fig. 3(a) [corresponding to the configuration in the bottom left corner of panel (e)]. We find a weak trend in the ER hub transition as a function of the simulation parameters outside of these cases, with the hub transition occurring very early in the growth simulations. Meanwhile, the CM model has a hub transition that exhibits a much stronger dependence on the attachment exponent α, with meaningful hub transitions occurring at roughly $\alpha = 0.5$ for most values of m, against which the CM results display little variation in panel (f).

Altogether the results of these simulations further suggest that the ER and CM encodings—and particularly the CM encoding—are identifying meaningful hub structure in controlled synthetic network data that is consistent with

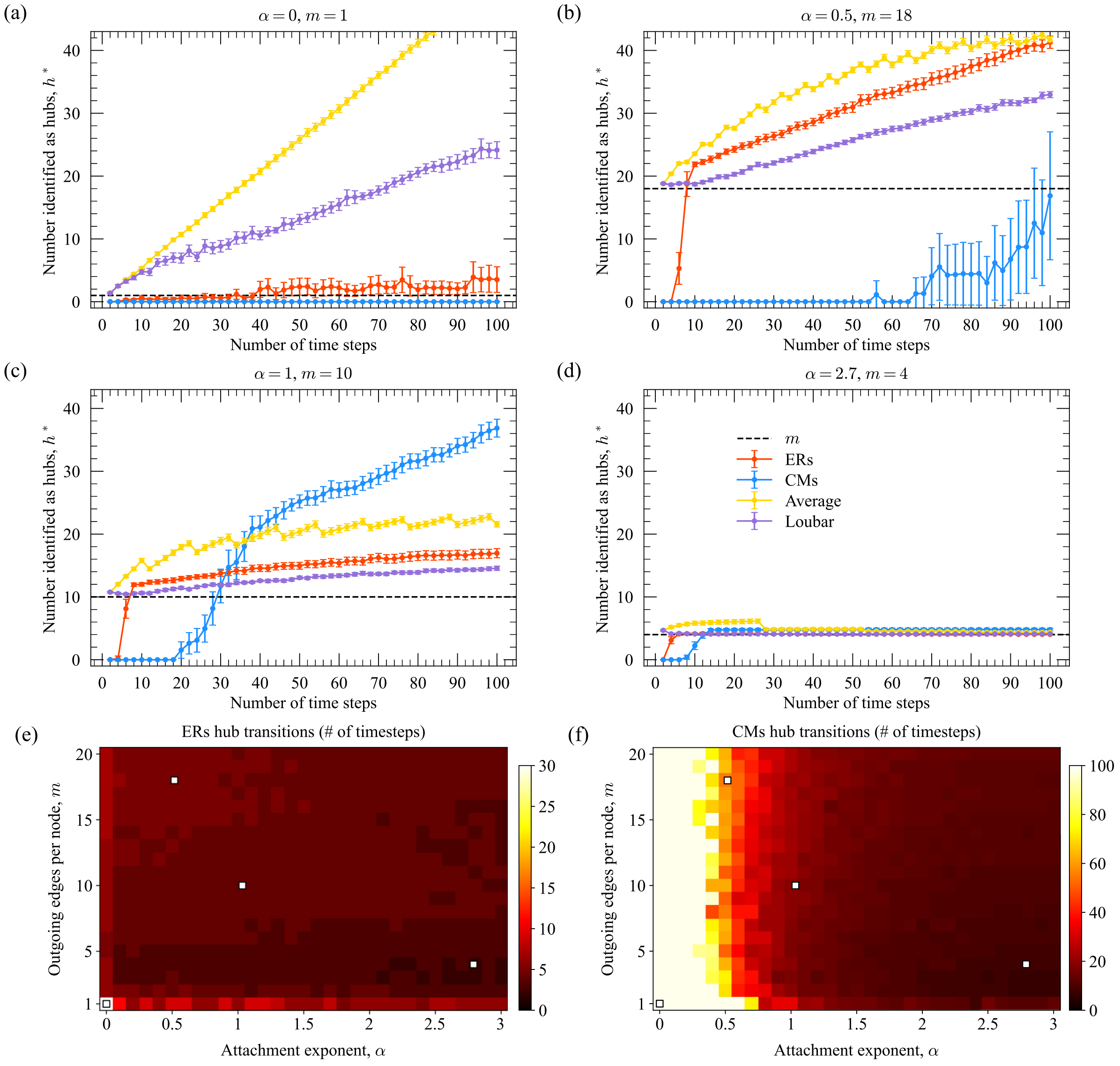

FIG. 3. Identifying hub transitions in Price's model with different attachment exponents and seed sets. (a)–(d) The number of hubs h^* identified by the four methods in Table I is shown as a function of the number of time steps in a generalization of Price's network growth model [Eq. (11)] [37,38] for various attachment exponents α and numbers of seed nodes m (dashed black lines). Error bars indicate two standard errors in the mean over 50 growth simulations with $T = 100$ time steps. (e),(f) Expected number of time steps until a single hub is detected (the "hub transition") over a range of attachment exponents and seed set sizes, for the ERs [Eq. (3)] and CMs [Eq. (4)] hub identification objectives. Small white squares indicate the parameter values corresponding to panels (a)–(d).

expectations based on the evolution of the networks' in-degree sequences. In the next section, we explore the application of these methods to a corpus of real networks from various disciplines.

C. Hubs in real-world networks

We collected 82 real-world network data sets from the Netzschleuder repository [40] by querying all networks for which "is_directed==True," "is_bipartite==False," and for which the number of edges M is less than 10^7. Networks with nonintegral weights were transformed to unweighted networks for the analyses, and after preprocessing there were 51 simple graphs—to which the ERs and CMs encodings were applied—and 31 weighted graphs/multigraphs—to which the ERm and CMm encodings were applied—for the analyses. The collected networks exhibit high variation in their size N and average degree M/N, and represent systems from a broad range of disciplines (see Table II in Appendix E for more details).

In Fig. 4(a), we plot the fractional number of hubs h^*/N as a function of the number of nodes N for all networks studied (both weighted and unweighted), using both the in- and out-degree distributions (giving 164 data points for each of the four methods in Table I). We find a story consistent with the findings for synthetic networks in Sec. III A. In particular, the ER and CM encodings both assign zero hubs in a substantial fraction of cases, with the CM method assigning zero hubs in

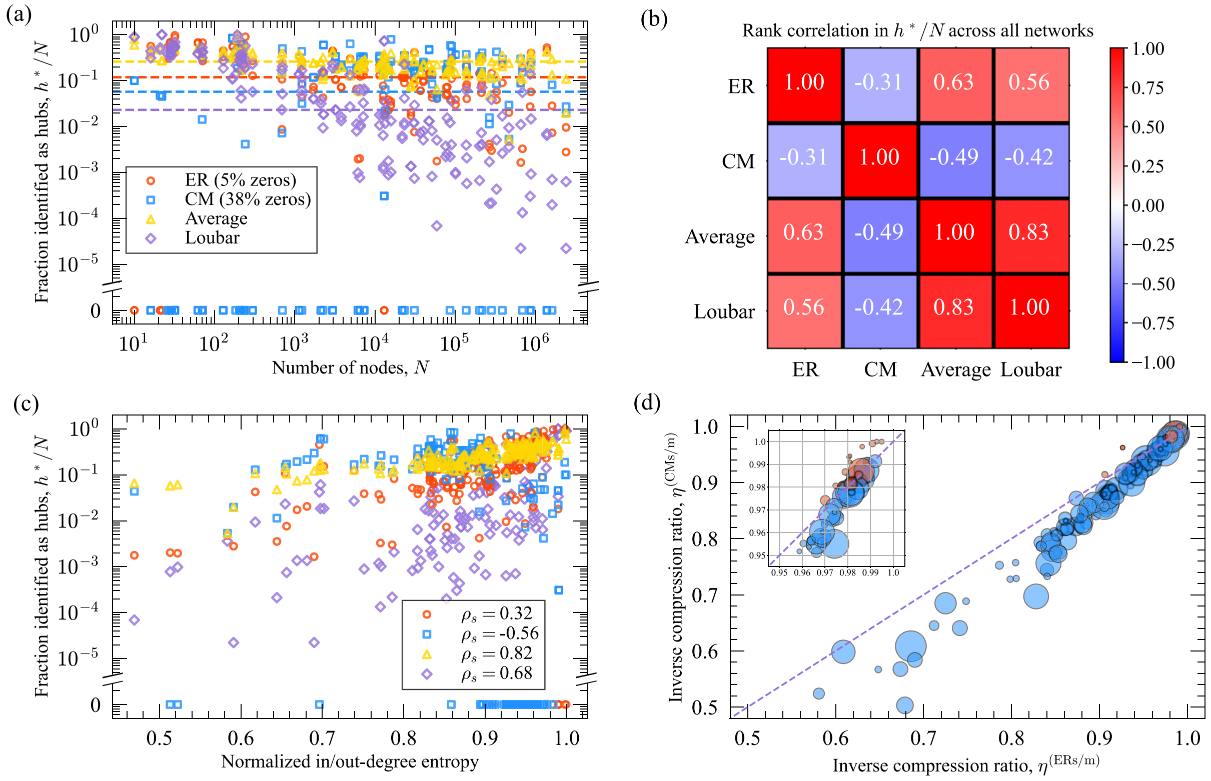

FIG. 4. Hub properties of real-world directed networks. (a) The fraction h^*/N of nodes identified as hubs using the four methods in Table I, for 82 real-world directed networks of various sizes collected from the Netzschleuder repository [40]. The median fraction of hubs found across all networks is shown with a dashed line for each method. See Appendix E for details on the networks studied. (b) Spearman rank correlation in the fraction of nodes identified as hubs across all networks in the corpus, for each pair of methods examined. (c) Fraction of nodes identified as hubs vs the normalized degree entropy [Eq. (12)] for the four methods. Spearman correlations between h^*/N and the normalized degree entropy values are reported in the legend, and the marker colors/styles correspond to the methods indicated in panel (a). (d) Inverse compression ratios [Eq. (5) for simple graphs and Eq. (10) for weighted graphs] across all networks when using the ER and CM encodings (x- and y-axes, respectively). The points are scaled monotonically with the size N of the network analyzed, and red (blue) markers indicate that the ER (CM) encoding was more compressive for the given network. The inset shows a zoomed-in view of the plot for $0.95 \leqslant \eta \leqslant 1$, and the line of equality is shown as a dashed line for reference.

over a third (38%) of the networks. In the cases where both methods detect hub nodes, the ER and CM methods detect a similar number of hubs, with the ER typically detecting slightly fewer. As seen in Fig. 2(c), the average method classifies a consistent *fraction* of nodes $h^*/N \approx 0.3$ as hubs across the range of network sizes N. Meanwhile the Loubar method is much more conservative and classifies a consistent *number* of nodes $h^* \approx 10$ as hubs as N varies in Fig. 4(a). The median values of h^*/N for the four methods vary by roughly an order of magnitude, with $\text{Median}_{\text{Average}} \approx 0.26 > \text{Median}_{\text{ER}} \approx 0.12 > \text{Median}_{\text{CM}} \approx 0.057 > \text{Median}_{\text{Loubar}} \approx 0.023$.

In Fig. 4(b), we compute the Spearman correlation coefficient between the h^*/N values produced by each pair of methods over all networks studied [in other words, the correlations in the y-values of panel (a)]. We find that while the average and Loubar methods are highly correlated with each other in terms of the fraction of hubs they identify across networks ($\rho \approx 0.83$), they are each more weakly correlated with the ER encoding's results ($\rho \approx 0.63$ and $\rho \approx 0.56$, respectively, for the average and Loubar methods). The CM method, meanwhile, is *negatively* correlated with the other three methods in terms of h^*/N—for a given network, when the CM encoding classifies a greater fraction of hubs than usual, other methods will tend to classify a smaller fraction of hubs than usual.

To investigate these correlations among the measures further, we plot the fraction of hubs versus the normalized entropy $H_{\text{norm}}(\boldsymbol{k})$ of the degree sequence $\boldsymbol{k}$, given by

$$H_{\text{norm}}(\boldsymbol{k}) = -\frac{1}{\log N}\sum_{i=1}^{N}\frac{k_i}{M}\log\left(\frac{k_i}{M}\right), \qquad (12)$$

for both in- and out-degrees $\boldsymbol{k}$ of each network in the corpus. The normalized degree entropy of Eq. (12) is a natural measure of the variability in the degrees $\boldsymbol{k}$ which is bounded in [0,1], with $H_{\text{norm}} = 0$ being the extreme where all edges point to a single node and $H_{\text{norm}} = 1$ being the other extreme where all nodes have identical degrees. We can see from Fig. 4(c) that the fraction of hubs h^*/N for the average and Loubar methods exhibits a fairly strong positive correlation with the normalized degree entropy of Eq. (12), while the ER encoding results are more weakly correlated with the degree entropy. In contrast, the CM encoding tends to assign a lower fraction of hub nodes as the degree entropy increases, as indicated by its strong negative correlation with H_{norm}. The behavior of the CM encoding is perhaps most aligned with expectations in that more heterogeneous degree sequences (those with lower entropy) should have more hubs, whereas degree sequences that are highly homogeneous (those with high entropy) should not have any hubs or only have relatively few hubs. This is also consistent with the observations in Fig. 2.

In Fig. 4(d), we plot the inverse compression ratio [Eq. (5) for simple graphs and Eq. (10) for weighted graphs] for the ER and CM encodings (x- and y-axis, respectively). We observe fairly substantial compression of the real networks using these hub-based encodings—more than 10% of the information required for transmission is reduced relative to the baseline encoding in many cases—and that the compression achieved with each method is relatively independent of network size N. We can also see that the CM encoding outperforms the ER encoding in most cases, but that the ER encoding is more compressive for small networks and networks where little compression is achievable (i.e., networks with homogeneous degree sequences). This is consistent with the results of Appendix D.

IV. CONCLUSION

Here we described a set of methods for identifying hub nodes in directed networks with weighted or unweighted edges whose goal is to extract the subset of high degree nodes that allows for the best compression of the network data. Our methods are nonparametric, selecting the number of hub nodes in the network automatically using the Minimum Description Length principle, and they can be run with a time complexity that is $O(N \log N)$ in the number of nodes N in the network. We apply these methods in a range of experiments involving real and synthetic network data, finding an intuitive dependence on the degree heterogeneity of networks and improved performance relative to existing methods that are not explicitly designed for compressing network data. These methods provide a simple, principled, and flexible toolkit for exploring the hub structure of network data in a range of applications.

There are a number of ways this work can be extended in future studies. First, methods for compressing network data are not limited to focusing on purely local structure [27,41], so one can in principle develop information-theoretic encodings that exploit hub structure at larger scales in order to classify nodes that are more globally central in the network as hub nodes. However, the description length of such global encodings may become very challenging to compute due to the combinatorial structure of the problem. One can also adapt the ideas in this work to undirected graphs, which are more challenging to deal with than the directed graphs considered in this paper because edges only need to be specified in a single direction. In this case, one may aim to find a set of nodes that constitutes a complete or nearly complete vertex cover of the graph as the hub nodes that provide the most efficient network compression. One can also compare the compression of the methods proposed here with other methods such as stochastic blockmodels [42] or various configuration models [27], as well as integrate hub-based priors in these models for improved compression. Finally, one can apply the proposed methods to human mobility networks in order to uncover hot-spot structure and compare empirical performance with the average and Loubar methods in the mobility context [22].

ACKNOWLEDGMENTS

This research was supported by the HKU-100 Start Up Grant and the HKU Urban Systems Institute Fellowship Grant.

APPENDIX A: RELATIVE COMPRESSION OF $\mathcal{L}_0^{(\text{ERs})}$ AND $\mathcal{L}_0^{(\text{CMs})}$

Before proceeding, we can establish the useful inequality

$$\log\binom{\sum_n x_n}{\sum_n y_n} - \sum_n \log\binom{x_n}{y_n} \geqslant 0 \tag{A1}$$

for non-negative integers $\{x_n\}, \{y_n\}$. Letting $Y = \sum_n y_n$, the Vandermonde identity gives

$$\binom{\sum_n x_n}{Y} = \sum_{\sum_n z_n = Y} \prod_n \binom{x_n}{z_n} \geqslant \prod_n \binom{x_n}{y_n}, \tag{A2}$$

and taking the logarithm of both sides of the inequality gives the desired result.

Applying Stirling's approximation to the first term in Eq. (2) gives, for $N \gg 1$, the following expression:

$$\log\binom{M+N-1}{N-1} \approx (M+N-1)H_b\left(\frac{N-1}{M+N-1}\right) \tag{A3}$$

$$= [N(\langle k\rangle+1)-1]H_b\left(\frac{N\langle k\rangle}{N(\langle k\rangle+1)-1}\right) \tag{A4}$$

$$\approx N(\langle k\rangle+1)H_b\left(\frac{\langle k\rangle}{\langle k\rangle+1}\right) \tag{A5}$$

$$\approx N\log(\langle k\rangle+1), \tag{A6}$$

where $H_b(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function. We can then see that for $N \gg \langle k\rangle \gg 1$, this term will vanish relative to the second term in Eq. (2), and so we have

$$\mathcal{L}_0^{(\text{CMs})} \approx \sum_{i=1}^{N} \log\binom{N-1}{k_i}. \tag{A7}$$

Now, applying the identity in Eq. (A1) we can see that

$$\mathcal{L}_0^{(\text{ERs})} - \mathcal{L}_0^{(\text{CMs})} \approx \log\binom{N(N-1)}{M} - \sum_{i=1}^{N} \log\binom{N-1}{k_i} \geqslant 0. \tag{A8}$$

Therefore, in the regime $N \gg \langle k\rangle \gg 1$, we will always achieve better compression using the two-step encoding with description length in Eq. (2) than the one-step encoding used for Eq. (1). However, for small and/or very sparse networks we do not have any guarantee that $\mathcal{L}_0^{(\text{CMs})} \leqslant \mathcal{L}_0^{(\text{ERs})}$ since the above approximation is no longer valid. In this regime, the one-step encoding may compress better than the two-step encoding since the initial transmission cost for the degrees is non-negligible.

APPENDIX B: OPTIMIZATION OF $\mathcal{L}^{(\mathrm{ERs})}(V_h)$ AND $\mathcal{L}^{(\mathrm{ERm})}(V_h)$

We can show that the global optimum of $\mathcal{L}^{(\mathrm{ERs})}(V_h)$ is obtained using the greedy hub identification process outlined in Sec. II B. This demonstrates that the optimal set of h hub nodes is the set of h nodes with the highest in-degrees. For any fixed $h \geqslant 1$, we have that substituting $M_h \to M_h + 1$—in other words, an increase in the cumulative in-degree M_h of the hubs [see Eq. (3)]—induces a change $\Delta_h^{(\mathrm{ERs})}$ in the description length $\mathcal{L}^{(\mathrm{ERs})}$ of

$$\Delta_h^{(\mathrm{ERs})} = \log \frac{[h(N-1) - M_h][M - M_h]}{[M_h + 1][(N-1)(N-h) - (M - M_h) + 1]} \tag{B1}$$

$$< \log \frac{[h(N-1) - M_h][M - M_h]}{M_h[(N-1)(N-h) - (M - M_h)]} \tag{B2}$$

$$= \log \frac{h(N-1)M - h(N-1)M_h - MM_h + M_h^2}{M_hN(N-1) - h(N-1)M_h - MM_h + M_h^2}. \tag{B3}$$

Now, as long as $h(N-1)M \leqslant M_hN(N-1)$—equivalent to the condition $M_h/h \geqslant (M/N)$, or that the average degree of the hubs is greater than or equal to the average in-degree of the network as a whole—then we have that $\Delta_h^{(\mathrm{ERs})} < 0$, which in turn implies that the optimal set of h nodes to choose as hubs are those with the h highest in-degrees, since this set of nodes will maximize M_h. Therefore, if we force the first hub node (i.e., for $h = 1$) to be the node of maximum in-degree, then the greedy scheme from Sec. II must produce a globally optimal set of hubs with respect to the description length in Eq. (3).

Similarly, we can show that the global optimum of $\mathcal{L}^{(\mathrm{ERm})}(V_h)$ is also obtained using the greedy hub identification process outlined in Sec. II B (with quantities appropriately mapped from the ERs encoding to the ERm encoding). The analogous expression to Eq. (B1) for the multigraph description length in Eq. (8) is

$$\Delta_h^{(\mathrm{ERm})} = \log \frac{[M_h + hN][M - M_h]}{[M_h + 1][M - M_h + N(N-h) - 1]} \tag{B4}$$

$$= \log \frac{[MM_h - M_h^2] + hN[M - M_h]}{[MM_h - M_h^2] + [M - M_h] + N(N-h)[M_h + 1] - [M_h + 1]} \tag{B5}$$

$$< \log \frac{[MM_h - M_h^2] + hN[M - M_h]}{[MM_h - M_h^2] + [N(N-h) - 1][M_h + 1]}. \tag{B6}$$

In this case, we need $hN[M - M_h] \leqslant [N(N-h) - 1][M_h + 1]$ to hold in order for $\Delta_h^{(\mathrm{ERm})} < 0$ (and, hence, global optimality of the greedy method). Rearranging the inequality, we can see that for $M_h, N(N-h) \gg 1$, the same condition $M_h/h > M/N$ will guarantee $\Delta_h^{(\mathrm{ERm})} < 0$. Therefore, forcing the highest in-degree node as the hub for $h = 1$ allows for global optimality of the greedy hub identification algorithm in this case as well, so long as the maximum in-degree is much greater than 1.

APPENDIX C: OPTIMIZATION OF $\mathcal{L}^{(\mathrm{CMs})}(V_h)$ AND $\mathcal{L}^{(\mathrm{CMm})}(V_h)$

We can guarantee local optimality of the greedy algorithm for $\mathcal{L}^{(\mathrm{CMs})}(V_h)$ by showing that the description length resulting from adding a node of degree $k + 1$ at step h is always less than the description length resulting from adding a node of degree k at step h. In other words, at any step h, we should add the remaining nonhub node with the highest in-degree as the new hub. The difference $\Delta_{k|h}^{(\mathrm{CMs})}$ in Eq. (4) due to adding a node of degree $k + 1$ at step h instead of a node of degree k at step h is given by

$$\Delta_{k|h}^{(\mathrm{CMs})} = \log \frac{[M_{h-1} + h + k][N - k - 1][M - M_{h-1} - k]}{[M_{h-1} + k + 1][k + 1][(N-h)(N-1) - (M - M_{h-1}) + k + 1]}, \tag{C1}$$

where here we have let M_{h-1} be the cumulative in-degree of whatever set of hubs we have chosen after step $h - 1$. Setting $h/N \to \gamma$, $M/N \to \langle k\rangle$, and $M_{h-1}/(h-1) \to \langle k\rangle_{h-1}$, we have that in the limit $N \gg 1$ the expression can be simplified to

$$\Delta_{k|h}^{(\mathrm{CMs})} \approx \log \frac{[\langle k\rangle_{h-1} + 1][\langle k\rangle - \gamma\langle k\rangle_{h-1}]}{[k + 1][\langle k\rangle_{h-1} - \gamma\langle k\rangle_{h-1}]}. \tag{C2}$$

From here, we can see that $\Delta_{k|h}^{(\mathrm{CMs})} \leqslant 0$ as long as

$$\langle k\rangle_{h-1}[k - \langle k\rangle] + \gamma\langle k\rangle_{h-1}[\langle k\rangle_{h-1} - k] + [\langle k\rangle_{h-1} - \langle k\rangle] \geqslant 0, \tag{C3}$$

which is satisfied for $\langle k\rangle_{h-1} \geqslant k \geqslant \langle k\rangle$. The first inequality is satisfied for the greedy scheme, since nodes are added in order of decreasing in-degree. Therefore, we have a guarantee of local optimality under the greedy scheme when $k \geqslant \langle k\rangle$. Repeating the above argument for $\mathcal{L}^{(\mathrm{CMm})}(V_h)$ gives the same final condition $\langle k\rangle_{h-1} \geqslant k \geqslant \langle k\rangle$ for local optimality using the CMm encoding. Outside of this regime—i.e., when the node being considered has a degree k that is less than the network average $\langle k\rangle$—we no longer have a proof of local optimality for the greedy scheme. However, as discussed in Sec. II, we can simply enforce the constraint that every hub node must have a degree that is at least as large as the highest nonhub node degree, in which case the greedy scheme is the only scheme that will produce hub sets V_h consistent with this constraint at all values of h.

APPENDIX D: RELATIVE COMPRESSION OF “ER” VERSUS “CM” HUB-BASED ENCODINGS

In the regime $\langle k\rangle_h \equiv M_h/h \gg 1$, we have that

$$\log \binom{M_h + h - 1}{h - 1} \approx [h(\langle k\rangle_h + 1) - 1]H_b\left(\frac{h\langle k\rangle_h}{h(\langle k\rangle_h + 1) - 1}\right)$$
$$\approx h(\langle k\rangle_h + 1)H_b\left(\frac{\langle k\rangle_h}{\langle k\rangle_h + 1}\right)$$
$$\approx h\log(\langle k\rangle_h + 1). \tag{D1}$$

Now, using a similar argument to the one in Appendix A, we have that this term will vanish relative to $\sum_{i\in V_h}\log\binom{N-1}{k_i}$, and subtracting Eq. (4) from Eq. (3) gives

$$\begin{aligned}
&\mathcal{L}^{(\mathrm{ERs})}(V_h) - \mathcal{L}^{(\mathrm{CMs})}(V_h)\\
&\quad= \log \binom{h(N-1)}{M_h} - \log \binom{M_h + h - 1}{h-1}\\
&\qquad - \sum_{i \in V_h} \log \binom{N-1}{k_i} && \text{(D2)}\\
&\approx \log \binom{h(N-1)}{M_h} - \sum_{i \in V_h} \log \binom{N-1}{k_i} && \text{(D3)}\\
&\geqslant 0, && \text{(D4)}
\end{aligned}$$

where we used the identity in Eq. (A1). Therefore, in the regime $\langle k \rangle_h \gg 1$, we will always achieve superior compression using the "CMs" encoding over the "ERs" encoding. However, for very sparse networks or networks with no hub nodes, we do not have any guarantee that $\mathcal{L}^{(\mathrm{CMs})} \leqslant \mathcal{L}^{(\mathrm{ERs})}$ since the above approximation is no longer valid. In this regime, the ERs encoding may compress better than the CMs encoding since the transmission cost for the degrees is non-negligible. The same argument can be used to establish that in the same regime we will achieve superior compression using the CMm encoding over the ERm encoding.

APPENDIX E: REAL-WORLD NETWORK DETAILS

TABLE II. Details on network datasets studied in Sec. III C.

No.	Name [40]	***N***	***M***	Weighted
0	packet_delays	10	45	True
1	rhesus_monkey	16	120	True
2	high_tech_company	21	210	True
3	moreno_taro	22	231	False
4	bison	26	325	True
5	moreno_sheep	28	378	True
6	cattle	28	378	True
7	7th_graders	29	406	True
8	hens	32	496	False
9	college_freshmen	32	496	True
10	macaques	62	1891	True
11	highschool	70	2415	True
12	law_firm	71	2485	True
13	foodweb_baywet	128	8128	False
14	email_company	167	13861	True
15	foodweb_little_rock	183	16653	False
16	psi	192	18336	True
17	cintestinalis	205	20910	False
18	fao_trade	214	22791	True
19	residence_hall	217	23436	True
20	un_migrations	232	26796	False
21	physician_trust	241	28920	False
22	celegansneural	297	43956	False
23	yeast_transcription	690	295283	True
24	messal_shale	700	244650	False
25	uni_email	1133	641278	False
26	polblogs	1224	934396	True
27	faa_routes	1226	750925	False
28	interactome_stelzl	1706	1454365	False
29	at_migrations	2115	2235555	True
30	interactome_figeys	2239	2505441	False
31	us_air_traffic	2278	2593503	True
32	fly_larva	2952	4355676	False
33	openflights	3214	5163291	False
34	bitcoin_alpha	3783	7153653	False
35	fediverse	4860	11807370	False
36	bitcoin_trust	5881	17290140	False
37	jung	6120	18724140	False
38	jdk	6434	20694961	False
39	advogato	6539	21379008	True
40	elec	7118	25329403	False
41	chess	7301	26648650	False
42	wiki_rfa	11381	64757890	False
43	dblp_cite	12590	79247755	False
44	anybeat	12645	79941690	False
45	chicago_road	12979	84224356	True
46	foldoc	13356	89184690	True
47	inploid	14629	106996506	False
48	google	15763	124228203	False
49	fly_hemibrain	21739	236281191	True
50	word_assoc	23132	267533146	True
51	cora	23166	268320195	False
52	lkml_reply	27927	840498910	True
53	digg_reply	30398	462004003	False
54	linux	30837	475444866	False
55	email_enron	36692	673133086	False
56	pgp_strong	39796	791840910	False
57	facebook_wall	46952	1102221676	False
58	slashdot_threads	51083	1304710903	False
59	python_dependency	58743	1725340653	False
60	epinions_trust	75879	2879167673	True
61	slashdot_zoo	79116	3129687706	True
62	twitter_15m	85712	3741488617	True
63	prosper	89269	3984432546	False
64	wiki_link_dyn	100312	5031198516	False
65	lastfm_aminer	136409	9303639436	False
66	wiki_users	138592	9603801936	False
67	academia_edu	200169	20033714196	False
68	google_plus	211187	22299868891	False
69	flickr_aminer	214626	23032052625	False
70	email_eu	265214	35169100291	False
71	stanford_web	281903	39734791656	True
72	notre_dame_web	325729	53049527856	False
73	citeseer	384413	73886485078	False
74	twitter	465017	108120172636	False
75	yahoo_ads	653260	213373987170	False
76	berkstan_web	685230	234770419065	True
77	myspace_aminer	854498	365082988753	False
78	google_web	875713	402754552837	True
79	wikitree	1382751	955999472625	False
80	trec_web	1601787	1282859995791	False
81	wikipedia-en-talk	2394385	2866538566920	False

[1] M. Newman, *Networks*, 2nd ed. (Oxford University Press, Oxford, 2018).

[2] M. P. van den Heuvel and O. Sporns, Network hubs in the human brain, Trends Cognit. Sci. **17**, 683 (2013).

[3] D. S. Bassett and O. Sporns, Network neuroscience, Nat. Neurosci. **20**, 353 (2017).

[4] T. Verma, N. A. Araújo, and H. J. Herrmann, Revealing the structure of the world airline network, Sci. Rep. **4**, 5638 (2014).

[5] C. Roucolle, T. Seregina, and M. Urdanoz, Measuring the development of airline networks: Comprehensive indicators, Transp. Res. Pt. A **133**, 303 (2020).

[6] X. He and J. Zhang, Why do hubs tend to be essential in protein networks?, PLoS Genetics **2**, e88 (2006).

[7] S. Mimar, D. Soriano-Paños, A. Kirkley, H. Barbosa, A. Sadilek, A. Arenas, J. Gómez-Gardeñes, and G. Ghoshal, Connecting intercity mobility with urban welfare, PNAS Nexus **1**, pgac178 (2022).

[8] J. Aguilar, A. Bassolas, G. Ghoshal, S. Hazarie, A. Kirkley, M. Mazzoli, S. Meloni, S. Mimar, V. Nicosia, J. J. Ramasco *et al.*, Impact of urban structure on infectious disease spreading, Sci. Rep. **12**, 3816 (2022).

[9] A.-L. Barabási, *Network Science* (Cambridge University Press, Cambridge, 2016).

[10] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, Epidemic thresholds in real networks, ACM Trans. Inf. Syst. Secur. **10**, 1 (2008).

[11] K. T. Gradoń, J. A. Hołyst, W. R. Moy, J. Sienkiewicz, and K. Suchecki, Countering misinformation: A multidisciplinary approach, Big Data Soc. **8**, 20539517211013848 (2021).

[12] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM **46**, 604 (1999).

[13] S. Oldham, B. Fulcher, L. Parkes, A. Arnatkeviciūtė, C. Suo, and A. Fornito, Consistency and differences between centrality measures across distinct classes of networks, PLoS ONE **14**, e0220061 (2019).

[14] J. R. F. Ronqui and G. Travieso, Analyzing complex networks through correlations in centrality measurements, J. Stat. Mech.: Theor. Exp. (2015) P05030.

[15] F. Grando, D. Noble, and L. C. Lamb, An analysis of centrality measures for complex and social networks, in *2016 IEEE Global Communications Conference (GLOBECOM)* (IEEE, Piscataway, NJ, 2016), pp. 1–6.

[16] D. Schoch, T. W. Valente, and U. Brandes, Correlations among centrality indices and a class of uniquely ranked graphs, Soc. Netw. **50**, 46 (2017).

[17] F. Bloch, M. O. Jackson, and P. Tebaldi, Centrality measures in networks, in *Social Choice and Welfare* (Springer-Verlag, Berlin, 2023), pp. 1–41.

[18] C. Shao, P. Cui, P. Xun, Y. Peng, and X. Jiang, Rank correlation between centrality metrics in complex networks: an empirical study, Open Phys. **16**, 1009 (2018).

[19] S. Wu and S. Wang, Information-theoretic outlier detection for large-scale categorical data, IEEE Trans. Knowl. Data Eng. **25**, 589 (2011).

[20] L. Akoglu, H. Tong, and D. Koutra, Graph based anomaly detection and description: a survey, Data Mining Knowledge Discov. **29**, 626 (2015).

[21] C. Böhm, K. Haegler, N. S. Müller, and C. Plant, Coco: coding cost for parameter-free outlier detection, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2009), pp. 149–158.

[22] T. Louail, M. Lenormand, O. G. Cantu Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, From mobile phone data to the spatial structure of cities, Sci. Rep. **4**, 5276 (2014).

[23] S. Hazarie, D. Soriano-Paños, A. Arenas, J. Gómez-Gardeñes, and G. Ghoshal, Interplay between population density and mobility in determining the spread of epidemics in cities, Commun. Phys. **4**, 191 (2021).

[24] T. Louail, M. Lenormand, M. Picornell, O. Garcia Cantu, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, Uncovering the spatial structure of mobility networks, Nat. Commun. **6**, 6007 (2015).

[25] W. Xu, H. Chen, E. Frias-Martinez, M. Cebrian, and X. Li, The inverted u-shaped effect of urban hotspots spatial compactness on urban economic growth, R. Soc. Open Sci. **6**, 181640 (2019).

[26] T. P. Peixoto, Bayesian stochastic blockmodeling, *Advances in Network Clustering and Blockmodeling* (Wiley, Hoboken, 2019), pp. 289–332.

[27] L. Hébert-Dufresne, J.-G. Young, A. Daniels, and A. Allard, Network onion divergence: Network representation and comparison using nested configuration models with fixed connectivity, correlation and centrality patterns, arXiv:2204.08444.

[28] R. J. Gallagher, J.-G. Young, and B. F. Welles, A clarified typology of core-periphery structure in networks, Sci. Adv. **7**, eabc9800 (2021).

[29] A. Kirkley, Spatial regionalization based on optimal information compression, Commun. Phys. **5**, 249 (2022).

[30] A. Kirkley, A. Rojas, M. Rosvall, and J.-G. Young, Compressing network populations with modal networks reveal structural diversity, Commun. Phys. **6**, 148 (2023).

[31] A. Kirkley, Constructing hypergraphs from temporal data, arXiv:2308.16546.

[32] B. Karrer and M. E. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E **83**, 016107 (2011).

[33] J. Rissanen, Modeling by the shortest data description, Automatica **14**, 465 (1978).

[34] A. Madotto and J. Liu, Super-spreader identification using meta-centrality, Sci. Rep. **6**, 38994 (2016).

[35] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, Fake news on twitter during the 2016 US presidential election, Science **363**, 374 (2019).

[36] https://github.com/aleckirkley/Network-hubs.

[37] D. d. S. Price, A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inf. Sci. **27**, 292 (1976).

[38] P. L. Krapivsky, S. Redner, and F. Leyvraz, Connectivity of growing random networks, Phys. Rev. Lett. **85**, 4629 (2000).

[39] S. Janson, Asymptotic degree distribution in random recursive trees, Random Struct. Alg. **26**, 69 (2005).

[40] T. P. Peixoto, The Netzschleuder network catalogue and repository (2020). Accessible at https://networks.skewed.de.

[41] T. P. Peixoto, Hierarchical block structures and high-resolution model selection in large networks, Phys. Rev. X **4**, 011047 (2014).

[42] T. P. Peixoto and A. Kirkley, Implicit models, latent compression, intrinsic biases, and cheap lunches in community detection, Phys. Rev. E **108**, 024309 (2023).