Think locally, act locally: Detection of small, medium-sized, and large communities in large networks

Lucas G. S. Jeub,[1] Prakash Balachandran,[2,3] Mason A. Porter,[1,4] Peter J. Mucha,[5] and Michael W. Mahoney[6,7]
[1]*Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom*
[2]*Morgan Stanley, Montreal, Quebec, H3C 3S4, Canada*
[3]*Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, USA*
[4]*CABDyN Complexity Centre, University of Oxford, Oxford OX1 1HP, United Kingdom*
[5]*Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27599-3250, USA*
[6]*International Computer Science Institute, Berkeley, California 94704, USA*
[7]*Department of Statistics, University of California at Berkeley, Berkeley, California 94720, USA*
(Received 15 March 2014; revised manuscript received 7 October 2014; published 26 January 2015)

It is common in the study of networks to investigate intermediate-sized (or "meso-scale") features to try to gain an understanding of network structure and function. For example, numerous algorithms have been developed to try to identify "communities," which are typically construed as sets of nodes with denser connections internally than with the remainder of a network. In this paper, we adopt a complementary perspective that communities are associated with bottlenecks of locally biased dynamical processes that begin at seed sets of nodes, and we employ several different community-identification procedures (using diffusion-based and geodesic-based dynamics) to investigate community quality as a function of community size. Using several empirical and synthetic networks, we identify several distinct scenarios for "size-resolved community structure" that can arise in real (and realistic) networks: (1) the best small groups of nodes can be better than the best large groups (for a given formulation of the idea of a good community); (2) the best small groups can have a quality that is comparable to the best medium-sized and large groups; and (3) the best small groups of nodes can be worse than the best large groups. As we discuss in detail, which of these three cases holds for a given network can make an enormous difference when investigating and making claims about network community structure, and it is important to take this into account to obtain reliable downstream conclusions. Depending on which scenario holds, one may or may not be able to successfully identify "good" communities in a given network (and good communities might not even exist for a given community quality measure), the manner in which different small communities fit together to form meso-scale network structures can be very different, and processes such as viral propagation and information diffusion can exhibit very different dynamics. In addition, our results suggest that, for many large realistic networks, the output of locally biased methods that focus on communities that are centered around a given seed node (or set of seed nodes) might have better conceptual grounding and greater practical utility than the output of global community-detection methods. They also illustrate structural properties that are important to consider in the development of better benchmark networks to test methods for community detection.

DOI: 10.1103/PhysRevE.91.012821 PACS number(s): 89.75.Fb, 89.75.Hc, 05.10.−a

I. INTRODUCTION

Many physical, technological, biological, and social systems can be modeled as networks, which in their simplest form are represented by graphs. A (static and single-layer) graph consists of a set of entities (called "vertices" or "nodes") and pairwise interactions (called "edges" or "links") between those vertices [1–3]. Network representations of data have led to numerous insights in the natural, social, and information sciences; and the study of networks has, in turn, borrowed ideas from all of these areas [4].

In general, networks can be described using a combination of local, global, and "meso-scale" perspectives. To investigate meso-scale structures—i.e., intermediate-sized structures that are responsible for "coupling" local properties, such as whether triangles close, and global properties, such as graph diameter—a fundamental primitive in many applications entails partitioning graphs into meaningful and/or useful sets of nodes [3]. The most popular form of such a partitioning procedure, in which one attempts to find relatively dense sets of nodes that are relatively sparsely connected to other sets, is known as "community detection" [5–7]. Myriad methods have been developed to algorithmically detect communities [5,6]; and these efforts have led to insights in applications such as committee and voting networks in political science [8–10], friendship networks at universities and other educational institutions [11–13], protein-protein interaction networks [14], granular materials [15], amorphous materials [16], brain and behavioral networks in neuroscience [17–19], collaboration patterns [20], human communication networks [21,22], human mobility patterns [23], and so on.

The motivation for the present work is the observation that it can be very challenging to find meaningful medium-sized or large communities in large networks [24–26]. Much of the large body of work on algorithmically identifying communities in networks has been applied successfully either to find communities in small networks or to find small communities in large networks [5,6,25], but it has been much less successful at finding meaningful medium-sized and large communities in large networks [27]. There are many reasons that it is difficult to find "good" large communities in large networks. We discuss several such reasons in the following paragraphs.

First, although it is typical to think about communities as sets of nodes with "denser" interactions among its members than between its members and the rest of a network, the literature contains neither a consensus definition of community nor a consensus on a precise formalization of what constitutes a "good" community [5,6].

Second, the popular formalizations of a "community" are computationally intractable, and there is little precise understanding or theoretical control on how closely popular heuristics to compute communities approximate the exact answers in those formulations [28,29]. Indeed, community structure itself is typically "defined" operationally via the output of a community-detection algorithm, rather than as the solution to a precise optimization problem or via some other mathematically precise notion [5,6].

Third, many large networks are extremely sparse [25] and thus have complicated structures that pose significant challenges for the algorithmic detection of communities via the optimization of objective functions [29]. This is especially true when attempting to develop algorithms that scale well enough to be usable in practice on large networks [6,25,30].

Fourth, the fact that it is difficult to visualize large networks complicates the validation of community-detection methods in such networks. One possible means of validation is to compare algorithmically obtained communities with known "ground truth" communities. However, notions of ground truth can be weak in large networks [12,25,31], and one rarely possesses even a weak notion of ground truth for most networks. Indeed, in many cases, one should not expect a real (or realistic) large network to possess a single feature that (roughly) dominates large-scale latent structure in a network. Thus, comparing the output of community-detection algorithms to ground truth in practice is most appropriate for obtaining coarse insights into how a network might be organized into social or functional groups of nodes [11]. Furthermore, different notions and/or formalizations of "community" concepts might be appropriate in different contexts [5,6,32–34], so it is desirable to formulate flexible methods that can incorporate different perspectives.

Fifth, community-detection algorithms often have subtle and counterintuitive properties as a function of the sizes of their inputs and/or outputs. For example, the community-size "resolution limit" of the popular modularity objective function is a fundamental consequence of the form (e.g., its additive nature is a key ingredient) of that objective function, but it only became obvious to people after it was explicitly pointed out [35].

Motivated by these observations, we consider the question of community quality as a function of the size (i.e., number of nodes) of a purported community. That is, we are concerned with questions such as the following. (1) What is the relationship between communities of different sizes in a given network? In particular, for a given network and a given community-quality diagnostic, are larger communities "better" or "worse" than smaller communities? (2) What is an appropriate way to think about medium-sized and large communities in large networks? In particular, how do smaller communities "fit together" into medium-sized and larger communities? (3) More generally, what effect do the answers to these questions have on downstream tasks that are of primary concern when modeling data using networks? For example, what effect do they have on processes such as viral propagation or the diffusion of information on networks?

By considering a suite of networks and using several related notions of community quality, we identify several scenarios that can arise in realistic networks.

(1) *Small communities are better than large communities.* In this scenario, for which there is an upward-sloping *network community profile* (NCP; see the discussion below), a network has small groups of nodes that correspond more closely than any large groups to intuitive ideas of what constitutes a good community.

(2) *Small and large communities are similarly good or bad.* In this scenario, for which an NCP is roughly flat, the most communitylike small groups of nodes in a network have similar community quality scores to the most communitylike large groups.

(3) *Large communities are better than small communities.* In this scenario, for which an NCP is downward sloping, a network has large groups of nodes that are more communitylike (i.e., "better" in some sense) than any small groups.

Although the third scenario is the one that has an intuitive isoperimetric interpretation and thus corresponds most closely with peoples' intuition when they develop and validate community-detection algorithms, one of our main conclusions is that most large realistic networks exhibit the first or second scenarios. This is consistent with recent results on network community structure that use related approaches [24–26] as well as somewhat different approaches [31,36], and it also helps illustrate the importance of considering community structures with groups that have large overlaps. For more on this, see our discussions below.

One of the main tools that we use to justify the above observations and to interpret the implications of community structure in a network is a *network community profile*, which was originally introduced in Ref. [25]. Given a community "quality" score—i.e., a formalization of the idea of a "good" community—an NCP plots the score of the best community of a given size as a function of community size. The authors of Refs. [24,25] considered the community-quality notion of conductance and employed various algorithms to approximate it. In subsequent work [26], many other notions of community quality have also been used to compute NCPs.

In the present paper, we compute NCPs using three different procedures to identify communities.

(1) *Diffusion-based dynamics.* First, we consider a diffusion-based dynamics (called the ACLCUT method; see the discussions in Sec. III C and Appendix B) from the original NCP analysis [25] that has an interpretation that good communities correspond to bottlenecks in the associated dynamics.

(2) *Spectral-based optimization.* Second, we consider a spectral-based optimization rule (called the MOVCUT method; see the discussions in Sec. III C and Appendix B) that is a locally biased analog of the usual global spectral graph-partitioning problem [37].

(3) *Geodesic-based dynamics.* Finally, we consider a geodesic-based spreading process (called the EGONET method; see the discussions in Sec. III C and Appendix B) that has an

interpretation that nodes in a good community are connected by short paths that emanate from a seed node [38].

We describe these three procedures in detail in Appendix B. For now, we note that the first and third procedures have a natural interpretation as defining communities operationally as the output of an underlying dynamical process, and the first and second procedures allow us to compare this operational approach with an optimization-based approach.

Viewed from this perspective, the computation of network community structure depends fundamentally on three things: (1) actual network structure, (2) the dynamical process or application of interest, and (3) the initial conditions or network region of interest. The perspective in point (3) contrasts with the prevalent view of community structure as arising simply from network structure [5,6], but it is consistent with the notion of dynamical systems depending fundamentally on their initial conditions, and it is crucial in many applications (e.g., both social [39,40] and biological contagions [41–43]).

Facebook's Data Science Team and its collaborators have demonstrated that one can view Facebook as a collection of egocentric networks that have been patched together into a network whose global structure is very sparse [44,45]. The above three community-identification methods have the virtue of combining the prevalent structural perspective with the idea that one is often interested in structure that is located "near" (in terms of both network topology and edge weights) an exogenously specified "seed set" of nodes [46]. The perspective in point (2) underscores the fact that one should not expect answers to be "universal." The differences between the aforementioned three methods lie in the specific dynamical processes that underlie them. We also note that, although we focus on the measure of community quality known as "conductance" (which is intimately related to the problem of characterizing the mixing rates of random walks [47]), one can view other quality functions (e.g., based on nonconservative dynamics [48–50] or geodesic-based dynamics [38]) as solving other problems, and they thus can reveal different aspects of community structure in networks.

The global NCPs that we compute from the three community-identification procedures are rather similar in some respects, suggesting that the characteristic features of NCPs are actual features of networks and not just artifacts of a particular way of sampling local communities. However, we observe significant differences in their local behaviors because they are based on different dynamical processes. In concert with other recent work (e.g., [34,51,52]), our results with these three procedures suggest that "local" methods that focus on finding communities centered around an exogenously specified seed node (or seed set of nodes) might have better theoretical grounding and more practical utility than other methods for community detection.

Our "local" (and "size-resolved") perspective on community structure also yields several other interesting insights. By design, it allows us to discern how community structure depends both on the seed node and on the size scales and time scales of a dynamical process running on a network. Similar perspectives were discussed in recent work on detecting communities in networks using Markov processes [9,32–34,53–55], and our approach is in the spirit of research on dynamical systems more generally, as bottlenecks to diffusion and other dynamics depend fundamentally on initial conditions. Local information algorithms are also an important approach for many other optimization problems and for practical purposes such as friend recommendation systems in online social networks [56]. Moreover, taking a local perspective on community structure is also consistent with the sociological idea of egocentric networks (and with real-world experiences of individuals, such as Facebook users [44,45], who interact with their personal neighborhoods of a social network). The local community of a given node should be similarly locally biased, and we demonstrate this feature quantitatively for several real networks. Using our perspective, we also demonstrate subtle yet fundamental differences between different networks: Some networks have high-quality communities of different sizes (especially small ones), whereas others do not possess communities of any size that give bottlenecks to diffusion-based dynamics. This is consistent with, and helps explain, prior direct observations of networks in which algorithmically computed communities seemed to have little or no effect on several dynamical processes [57]. More generally and importantly, whether small or large communities are "better" with respect to some measure of community quality has significant consequences not only for algorithms that attempt to identify communities but also for the dynamics of processes such as viral propagation and information diffusion.

The rest of this paper is organized as follows. Because our approach to examining network communities is uncommon in the physics literature, we start in Sec. II with an informal description of our approach. We then introduce NCPs in Sec. III. In Sec. IV, we present our main empirical results on community quality as a function of size, and we provide a detailed comparison of our three community-identification procedures when applied to real networks. This illustrates the three distinct scenarios of community quality versus community size that we described above. In Sec. V, we illustrate the behavior of these methods on the well-known Lancichinetti-Fortunato-Radicchi (LFR) benchmark networks that are commonly used to evaluate the performance of community-detection techniques. We find that their NCPs have a characteristic shape for a wide range of parameter values and are unable to reproduce the different scenarios that one observes for real networks. We then conclude in Sec. VI with a discussion of our results. In Appendix A, we provide a brief discussion of expander graphs (aka "expanders"). In Appendix B, we give detailed descriptions of the three specific procedures that we use to identify communities. Appendices C and D contain additional empirical results.

II. BACKGROUND AND PRELIMINARIES

In this section, we describe some background and preliminaries that provide the framework that we use to interpret our results on size-resolved community structure in Secs. IV and V. We start in Sec. II A by defining the notation that we use throughout this paper, and we continue in Sec. II B with a brief discussion of possibilities for what a network might "look like" if one is interested in its meso-scale or large-scale structure. To convey the basic idea of our approach, much of our discussion in this section is informal. In later sections, we make these ideas more precise.

A. Definitions and notation

We represent each of the networks that we study as an undirected graph. We consider both weighted and unweighted graphs.

Let $G = (V, E, w)$ be a connected and undirected graph with node set V, edge set E, and set w of weights on the edges. Let $n = |V|$ denote the number of nodes, and let $m = |E|$ denote the number of edges. The edge $\{i, j\}$ has weight w_{ij}. Let $A = A_G \in \mathbb{R}^{n\times n}$ denote the (weighted) adjacency matrix of G. Its elements are $A_G(i, j) = w_{ij}$ if $\{i, j\} \in E$ and $A_G(i, j) = 0$ otherwise. The matrix $D = D_G \in \mathbb{R}^{n\times n}$ denotes the diagonal degree matrix of G. Its elements are $D_G(i,i) = d_i = \sum_{\{i,j\}\in E} w_{ij}$, where d_i is called the "strength" or "weighted degree" of node i. The combinatorial Laplacian of G is $L_G = D_G - A_G$, and the normalized Laplacian of G is $\mathcal{L}_G = D_G^{-1/2} L_G D_G^{-1/2}$.

A path P in G is a sequence of edges $P = \{\{i_k, j_k\}\}_{k=1}^{s}$, such that $j_k = i_{k+1}$ for $k = 1, \ldots, s-1$. The length of path P is $|P| = \sum_{\{i,j\}\in P} l_{ij}$, where l_{ij} is the length of the edge that connects nodes i and j. For an unweighted network, $l_{ij} = 1$ for all edges. For weighted networks, we use w_{ij} as a measure of closeness for the tie between nodes i and j, and a common choice for length is $l_{ij} = \frac{1}{w_{ij}}$. Let $\mathcal{P}_{ij}$ be the set of all paths between i and j. The geodesic distance $\Delta_{ij} = \min_{P\in\mathcal{P}_{ij}} |P|$ between nodes i and j is the length of a shortest path between i and j. The k-neighborhood $N_k(i) = \{j \in V : \Delta_{ij} \leqslant k\}$ of i is the set of all nodes that are at most a distance k away from i, and the k-neighborhood of a set of nodes S is $N_k(S) = \bigcup_{i\in S} N_k(i)$.

The *egocentric network* (i.e., *ego network* or *ego-net*) [125] of a node (the ego) is the subgraph induced by the neighborhood of the ego. That is, the ego-net is the network that consists of all nodes that are in the ego's 1-neighborhood and all edges between those nodes. We use the term *k-ego-net* for the subgraph induced by the k-neighborhood of a node. The traditional definition of an ego-net excludes the ego and its edges, but we specifically include them.

B. What can networks look like?

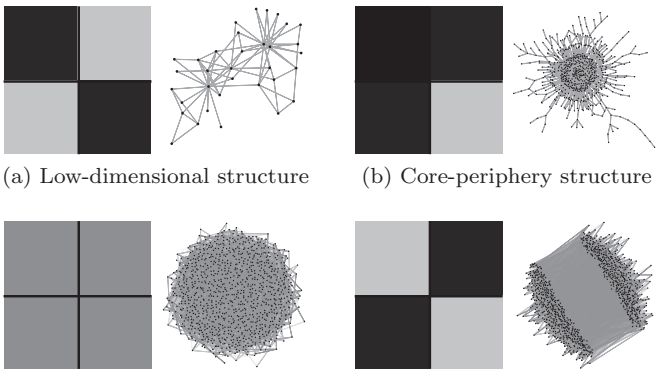
Before examining real networks, we start with the following question: What can a network look like, very roughly, if one "squints" at it? This question is admittedly vague, but the answer to it governs how small-scale network structure "interacts" with large-scale network structure, and it informs researchers' intuitions and the design decisions that they make when analyzing networks (and when developing methods to analyze networks). As an example of this idea, it should be intuitively clear that if one squints at the nearest-neighbor $\mathbb{Z}^2$ network (i.e., the uniform lattice of pairs of integers on the Euclidean plane), then they look like the Euclidean plane $\mathbb{R}^2$. Distances are approximately preserved and, up to boundary conditions and discretization effects, dynamical processes on one approximate the analogous dynamical processes on the other. In the fields of geometric group theory and coarse geometry, this intuitive connection between $\mathbb{Z}^2$ and $\mathbb{R}^2$ has been made precise using the notions of coarse embeddings and quasi-isometries [58].

Establishing quasi-isometric relationships on networks that are expander graphs (aka "expanders"; see Appendix A) is technically brittle [60]. Thus, for the present informal discussion, we rely on a simper notion. Suppose that we are interested in the "best fit" of the adjacency matrix A to a 2×2 block matrix,

$$A = \begin{pmatrix} A_{11} & A_{12}^T \\ A_{12} & A_{22} \end{pmatrix},$$

where $A_{ij} = \alpha_{ij}\vec{1}\vec{1}^T$, where $\alpha_{ij} \in \mathbb{R}^+$ (i.e., they are non-negative) and the "1-vector" $\vec{1}$ is a column vector of the appropriate dimension that contains a 1 in every entry. Thus, each block in A has uniform values for all of its elements, and larger values of α_{ij} correspond to stronger interactions between nodes. The structure of A is then determined based on the relative sizes of α_{11}, α_{12}, and α_{22}. The various relative sizes of these three scalars have a strong bearing on the structure of the network associated with A. We illustrate several examples in Fig. 1. For the block models that we use for three of its panels, one block has n_1 nodes and the second block has n_2 nodes, and a node in block i is connected to a node in block j with probability α_{ij} [61].

(1) *Low-dimensional structure.* In Fig. 1(a), we illustrate the case in which $\alpha_{11} \approx \alpha_{22} \gg \alpha_{12}$. In this case, each half of the network interacts with itself more densely than it interacts with the other half of the network. This "hot dog" or "pancake" structure corresponds to the situation in which there are two (or any number, in the case of networks more generally) dense communities of nodes that are reasonably well balanced in the sense that each community has roughly the same number of nodes. In this case, the network embeds relatively well in a one-dimensional, two-dimensional, or



(a) Low-dimensional structure (b) Core-periphery structure

(c) Expander or complete graph (d) Bipartite structure

FIG. 1. Idealized block models of network adjacency matrices; darker blocks correspond to denser connections among its component nodes. Panel (a) illustrates a low-dimensional "hot dog" or "pancake" structure; panel (b) illustrates a "core-periphery" structure; panel (c) illustrates an unstructured expander or complete graph; and panel (d) illustrates a bipartite graph. Our example networks are the Zachary Karate Club [59] in panel (a) and a realization of a random-graph block model in panels (b)–(d). For panel (b), we only show the largest connected component (LCC); the networks in panels (c) and (d) are connected. The parameters for the block models are as follows: (b) $\alpha_{11} = 0.3$, $\alpha_{22} = 0.001$, $\alpha_{12} = 0.005$, $n_1 = 50$ nodes, and $n_2 = 950$ nodes (the LCC has 615 nodes); (c) $\alpha_{11} = \alpha_{22} = \alpha_{12} = 0.01$, and $n_1 + n_2 = 1000$ nodes (this is an Erdős-Rényi graph); (d) $\alpha_{11} = \alpha_{22} = 0$, $\alpha_{12} = 0.02$, and $n_1 = n_2 = 500$ nodes.

other low-dimensional space. Spectral clustering or other clustering methods often find meaningful communities in such networks, and one can often readily construct meaningful and interpretable visualizations of network structure.

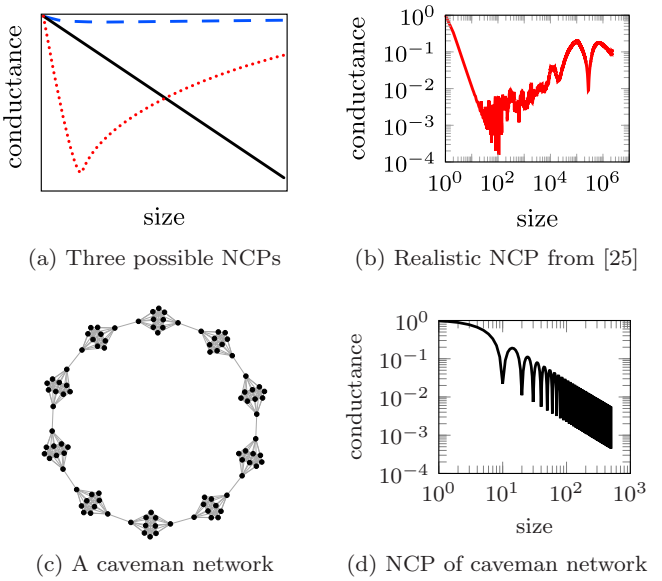
(2) *Core-periphery structure.* In Fig. 1(b), we illustrate the case in which $\alpha_{11} \gg \alpha_{12} \gg \alpha_{22}$. This is an example of a network with a density-based "core-periphery" structure [24,25,62–64]. There is a core set of nodes that are relatively well connected both among themselves and to a set of peripheral nodes that interact very little among themselves.

(3) *Expander or complete graph.* In Fig. 1(c), we illustrate the case in which $\alpha_{11} \approx \alpha_{12} \approx \alpha_{22}$. This corresponds to a network with little or no discernible structure. For example, if $\alpha_{11} = \alpha_{12} = \alpha_{22} = 1$, then the graph is a clique (i.e., the complete graph). Alternatively, if the graph is a constant-degree expander, then $\alpha_{11} \approx \alpha_{12} \approx \alpha_{22} \ll 1$. As discussed in Appendix A, constant-degree expanders yield the metric spaces that embed least well in low-dimensional Euclidean spaces. In terms of the idealized block model in Fig. 1, they look like complete graphs, and partitioning them would not yield network structure that one should expect to construe as meaningful. Informally, they are largely unstructured when viewed at large size scales.

(4) *Bipartite structure.* In Fig. 1(d), we illustrate the case in which $\alpha_{12} \gg \alpha_{11} \approx \alpha_{22}$. This corresponds to a bipartite or nearly bipartite graph. Such networks arise, e.g., when there are two different types of nodes, such that one type of node connects only to (or predominantly to) nodes of the other type [65].

Most methods for algorithmic detection of communities have been developed and validated using the intuition that networks have some sort of low-dimensional structure [5,25,36]. As an example, consider the infamous Zachary Karate Club network [59], which we show in Fig. 1(a). This well-known benchmark graph, which seems to be an almost obligatory example to include in papers that discuss community structure, clearly looks like it has a nice low-dimensional structure. For example, there is a clearly identifiable left half and right half, and two-dimensional visualizations of the network [such as that in Fig. 1(a)] highlight that bipartition. Indeed, the Zachary Karate Club network possesses well-balanced and (quoting Simon [66]) "nearly decomposable" communities; and the nodes in each community are more densely connected to nodes in the same community than they are to nodes in the other community. Relatedly, appropriately reordering the nodes of the Zachary Karate Club network yields an adjacency-matrix representation with an almost block-diagonal structure with two blocks [as typified by the cartoon in Fig. 1(a)]; and any reasonable community-detection algorithm should be able to find (exactly or approximately) the two communities.

Another well-known network that (slightly less obviously) looks like it has a low-dimensional structure is a so-called caveman network, which we illustrate later [in Fig. 2(c)]. Arguably, a caveman network has many more communities than the Zachary Karate Club, so details such as whether an algorithm "should" split it into two or a somewhat larger number of reasonably well-balanced communities might be different than in the Zachary Karate Club network. However, a caveman network also has a natural well-balanced partition that respects intuitive community structure. Reasonable



(a) Three possible NCPs (b) Realistic NCP from [25]

(c) A caveman network (d) NCP of caveman network

FIG. 2. (Color online) Illustration of network community profiles (NCPs) of conductance versus community size. (a) Stylized versions of possible shapes for an NCP: downward-sloping (black, solid curve), upward-sloping (red, dotted curve), and flat (blue, dashed curve). (b) NCP of a LIVEJOURNAL network that illustrates the characteristic upward-sloping NCP that is typical for many large empirical social and information networks [25]. (c) A toy "caveman network" with 10 cliques of 10 nodes each, where one edge from each clique has been rewired to create a ring [70]. (d) NCP for a similar caveman network with 100 cliques of 10 nodes each [the NCP for the network in panel (c) is identical for communities with fewer than 50 nodes], illustrating the characteristic downward-sloping NCP that is typical of networks that are embedded in a low-dimensional space.

two-dimensional visualizations of this network [such as the one that we present in Fig. 2(c)] shed light on that structure; and any reasonable community-detection algorithm can be adjusted to find (exactly or approximately) the expected communities. In this paper, we demonstrate that most realistic networks do *not* look like these small examples. Instead, realistic networks are often poorly approximated by low-dimensional structures (e.g., with a small number of relatively well-balanced communities, each of which is more densely connected internally than it is with the rest of the network). Realistic networks often include substructures that more closely resemble core-periphery graphs or expander graphs [see Figs. 1(b) and 1(c)]; and networks that partition into nice, nearly decomposable communities tend to be exceptional rather than typical [24,25,36].

III. NETWORK COMMUNITY PROFILES AND THEIR INTERPRETATION

Recall from Sec. I that an NCP measures the quality of the best possible community of a given size as a function of the size of the purported community [24–26]. In this section, we provide a brief description of NCPs and how we use them.

A. The basic NCP: Measuring size-resolved community quality

We start with the definition of conductance and the original conductance-based definition of an NCP from Ref. [25], and

we then discuss our extensions of such ideas. For more details on conductance and NCPs, see Refs. [25,37,67,68]. If $G = (V, E, w)$ is a graph with weighted adjacency matrix A, then the "volume" between two sets S_1 and S_2 of nodes (i.e., $S_i \subset V$) equals the total weight of edges with one end in S_1 and one end in S_2. That is,

$$\mathrm{vol}(S_1, S_2) = \sum_{i \in S_1} \sum_{j \in S_2} A_{ij}. \tag{1}$$

In this case, the "volume" of a set $S \subset V$ of nodes is

$$\mathrm{vol}(S) = \mathrm{vol}(S, V) = \sum_{i \in S} \sum_{j \in V} A_{ij}. \tag{2}$$

In other words, the set volume equals the total weight of edges that are attached to nodes in the set. The volume $\mathrm{vol}(S, \overline{S})$ between a set S and its complement $\overline{S}$ has a natural interpretation as the "surface area" of the "boundary" between S and $\overline{S}$. In this study, a set S is a hypothesized community. Informally, the conductance of a set S of nodes is the surface area of that hypothesized community divided by "volume" (i.e., size) of that community. From this perspective, studying community structure amounts to an exploration of the isoperimetric structure of G.

Somewhat more formally, the *conductance of a set of nodes* $S \subset V$ is

$$\phi(S) = \frac{\mathrm{vol}(S, \overline{S})}{\min\left(\mathrm{vol}(S), \mathrm{vol}(\overline{S})\right)}. \tag{3}$$

Thus, smaller values of conductance correspond to better communities. The *conductance of a graph* G is the minimum conductance of any subset of nodes:

$$\phi(G) = \min_{S \subset V} \phi(S). \tag{4}$$

Computing the conductance $\phi(G)$ of an arbitrary graph is an intractable problem (in the sense that the associated decision problem is NP-hard [69]), but this quantity can be approximated by the second-smallest eigenvalue λ_2 of the normalized Laplacian [67,68].

If the "surface area to volume" (i.e., isoperimetric) interpretation captures the notion of a good community as a set of nodes that is connected more densely internally than with the remainder of a network, then computing the solution to Eq. (4) leads to the "best" (in this sense) community of any size in the network.

Instead of defining a community quality score in terms of the best community of any size, it is useful to define a community quality score in terms of the best community of a given size k as a function of the size k. To do this, Ref. [25] introduced the idea of a *network community profile (NCP)* as the lower envelope of the conductance values of communities of a given size:

$$\phi_k(G) = \min_{S \subset V, |S| = k} \phi(S). \tag{5}$$

An NCP plots a community quality score (which, as in Ref. [25], we take to be the set conductance of communities) of the best possible community of size k as a function of k. Clearly, it is also intractable to compute the quantity $\phi_k(G)$ in Eq. (5) exactly. Previous work has used spectral-based and flow-based algorithms to approximate it [24–26].

To gain insight into how to understand an NCP and what it reveals about network structure, consider Fig. 2. In Fig. 2(a), we illustrate three possible ways that an NCP can behave. In each case, we use conductance as a measure of community quality. The three cases are the following ones.

(1) *Upward-sloping NCP.* In this case, small communities are "better" than large communities.
(2) *Flat NCP.* In this case, community quality is independent of size. (As illustrated in this figure, the quality tends to be comparably poor for all sizes.)
(3) *Downward-sloping NCP.* In this case, large communities are better than small communities.

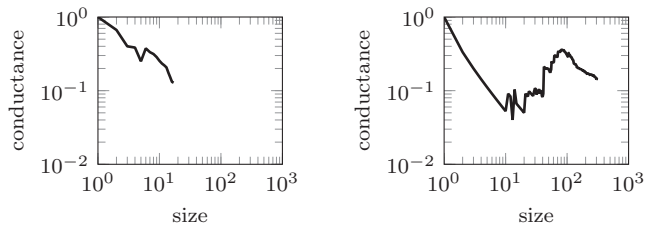
For ease of visualization and computational considerations, we only show NCPs for communities up to half of the size of a network. An NCP for very large communities, which we do not show in figures as a result of this choice, roughly mirrors that for small communities, as the complement of a good small community is a good large community because of the inherent symmetry in conductance [see Eq. (3)].

In Fig. 2(b), we show an NCP of a LIVEJOURNAL network from Ref. [25]. It demonstrates an empirical fact about a large variety of large social and information networks: There exist good small conductance-based communities, but there do not exist any good large conductance-based communities in many such networks. (See Refs. [24–26,37,67,68] for more empirical evidence that large social and information networks tend not to have large communities with low conductances.) On the contrary, Fig. 2(c) illustrates a small toy network—a so-called "caveman network"—formed from several small cliques connected by rewiring one edge from each clique to create a ring [70]. As illustrated by the downward-sloping NCP in Fig. 2(d), this network possesses good conductance-based communities, and large communities are better than small ones. One obtains a similar downward-sloping NCP for the Zachary Karate Club network [59] as well as for many other networks for which there exist meaningful visualizations [25]. The wide use of networks that have interpretable visualizations (such as the Zachary Karate Club and planted-partition models [71] with balanced communities) to help develop and evaluate methods for community detection and other procedures can lead to a strong selection bias when evaluating the quality of those methods.

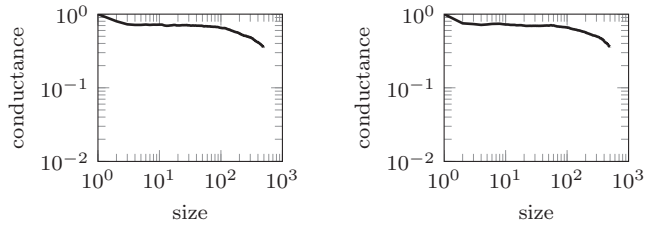
We now consider the relationship between the phenomena illustrated in Fig. 2 and the idealized block models of Fig. 1. As a concrete example, Fig. 3 shows the NCPs for the example networks in Fig. 1.

First, note that the best partitions consist roughly of well-balanced communities in the low-dimensional case of Figs. 1(a) and 3(a), and the "lowest" point on an NCP tends to be for large community sizes. Thus, an NCP tends to be downward sloping for low-dimensional examples.

Networks with pronounced core-periphery structure—e.g., networks that look like the example in Fig. 1(b)—tend to have many good small communities but no comparably good or better large communities. This situation arises in many large, extremely sparse networks [24–26]. The good small communities in such networks are sets of connected nodes in the extremely sparse periphery, and they do not combine to



(a) Low-dimensional structure (b) Core-periphery structure



(c) Expander or complete graph (d) Bipartite structure

FIG. 3. Network community profiles of the idealized example networks from Fig. 1. (a) NCP for the Zachary Karate Club network. (b) NCP for an example network generated from a block model with core-periphery structure. (c) NCP for an Erdős-Rényi graph. (d) NCP for an example network generated from a bipartite block model.

form good, large communities, as they are only connected via a set of core nodes with denser connections than the periphery. Thus, an NCP of a network with core-periphery structure tends to be upward sloping, as illustrated in Figs. 1(b) and 3(b). However, this observation does not apply to all networks with well-defined density-based core-periphery structure. If the periphery is sufficiently well connected (though still much sparser than the core), then one no longer observes good, small communities. Such networks act like expanders from the perspective of the behavior of random walkers, so they have a flat NCP. One can generate examples of such networks by modifying the parameters of the block model that we used to generate the example network in Fig. 1(b) [61].

For a complete graph or a degree-homogeneous expander [see Figs. 1(c) and 3(c)], all communities tend to have poor quality, so an NCP is roughly flat. (See Appendix A for a discussion of expander graphs.)

Finally, bipartite structure itself does not have any characteristic influence on an NCP. Instead, an NCP of a bipartite network reveals other structure present in the network. For the example network in Fig. 1(d), the two types of nodes are connected uniformly at random, so its NCP [see Fig. 3(d)] has the characteristic flat shape of an expander.

B. Robustness and information content of NCPs

It is important to discuss the robustness properties of NCPs. Such properties are not obvious *a priori*, as an NCP is an extremal diagnostic. Importantly, however, the qualitative property of being downward-sloping, upward-sloping, or roughly flat is very robust to the removal of nodes and edges, variations in data generation and preprocessing decisions, and similar sources of perturbation [24–26]. For example, upward-sloping NCPs typically have many small communities of good quality, so losing some communities via noise or some other perturbations has little effect on a realistic NCP. Naturally, whether a particular set of nodes achieves a local minimum is not robust to such modifications. In addition, one can easily construct pathological networks whose NCPs are not robust.

It is also important to consider the robustness of a network's NCP with respect to the use of conductance versus other measures of community quality. (Recall that many other measures have been proposed to capture the criteria that a good community should be densely connected internally but sparsely connected to the rest of a network [5,25].) Indeed, it has been shown that measures that capture both criteria of community quality (internal density and external sparsity) behave in a roughly similar manner to conductance-based NCPs, whereas measures that capture only one of the two criteria exhibit qualitatively different behavior (typically for rather trivial reasons) [26].

Although the basic NCP that we have been discussing yields numerous insights about both small-scale and large-scale network structure, it also has important limitations. For example, an NCP gives no information on the number or density of communities with different community quality scores. (This contributes to the robustness properties of NCPs with respect to perturbations of a network.) Accordingly, the communities that are revealed by an NCP need not be representative of the majority of communities in a network. However, the extremal features that are revealed by an NCP have important system-level implications for the behavior of dynamical processes on a network: They are responsible for the most severe bottlenecks for associated dynamical processes on networks [72].

Another property that is not revealed by an NCP is the internal structure of communities. Recall from Eq. (3) that the conductance of a community measures how well (relative to its size) it is separated from the remainder of a network, but it does not consider the internal structure of a community (except for size and edge density). In an extreme case, a community with good conductance might even consist of several disjoint pieces. Recent work has addressed how spectral-based approximations to optimizing conductance also approximately optimize measures of internal connectivity [73].

We augment the information from basic NCPs with some additional computations. To obtain an indication of a community's internal structure, we compute the internal conductance of the communities that form an NCP. The *internal conductance* $\phi_{\text{in}}(S)$ of a community S is

$$\phi_{\text{in}}(S) = \phi(G|_S), \tag{6}$$

where $G|_S$ is the subgraph of G induced by the nodes in the community S. The internal conductance is equal to the conductance of the best partition into two communities of the network $G|_S$ viewed as a graph in isolation. Because a good community should be well separated from the remainder of a network and also relatively well connected internally, we expect good communities to have low conductance but high internal conductance. We thus compute the *conductance ratio*

$$\Phi(S) = \frac{\phi(S)}{\phi_{\text{in}}(S)} \tag{7}$$

to quantify this intuition. A good community should have a small conductance ratio, and thus we also plot so-called *conductance ratio profiles (CRPs)* [25] to illustrate how conductance ratio depends on community size in networks.

C. Our application and extension of NCPs

In this paper, we examine small-scale, medium-scale, and large-scale community structure using conductance-based NCPs and CRPs. We employ three different methods, which we introduce in detail in Appendix B, for sampling an NCP: one based on local diffusion dynamics (the ACLCUT method), one based on a local spectral optimization (the MOVCUT method), and one based on geodesic distance from a seed node (the EGONET method). In each case, we find communities of different sizes, and we then plot the conductance of the best community for each size as a function of size. For each method, one can start from either a single seed node or a seed set of nodes. All of the numerical simulations in the present paper use a single seed node.

An NCP provides a signature of community structure in a network, and we can thereby compare community structure across different networks. This helps one to discern which properties are attributable predominantly to network structure and which are attributable predominantly to choice of algorithms for community detection. Our approach of comparing community structures in networks using NCPs and CRPs is very general: One can, of course, follow a similar procedure using other community-quality diagnostics on the vertical axis, other procedures for community generation, and so on.

IV. EMPIRICAL RESULTS ON REAL NETWORKS

In this section, we present the results of our empirical evaluation of the small-scale, medium-scale, and large-scale community structure in our example networks.

A. Example network data sets

We examine six empirical networks in depth. They fall into three classes: coauthorship networks, Facebook networks, and voting similarity networks. For each class, we consider two networks of different sizes.

(1) *Collaboration graphs.* The two (unweighted) coauthorship networks were constructed from papers submitted to the arXiv preprint server in the areas of general relativity and quantum cosmology (CA-GRQC) and astrophysics (CA-ASTROPH). In each case, two authors are connected by an edge if they coauthored at least one paper, so a paper with k authors appears as a k-clique (i.e., a complete k-node subgraph) in a network. These data sets are available as part of the Stanford Network Analysis Package (SNAP), and they were examined previously in Refs. [24–26].

(2) *Facebook graphs.* The two (unweighted) Facebook networks are anonymized data sets that consist of a snapshot of "friendship" ties on one particular day in September 2005 for two United States (U.S.) universities: Harvard (FB-HARVARD1) and Johns Hopkins (FB-JOHNS55). They form a subset of the FACEBOOK100 data set from Refs. [11,12]. In addition to the friendship ties, note that we possess node labels for gender and class year as well as numerical identifiers for student or some other (e.g., faculty) status, major, and high school.

(3) *Congressional voting graphs.* The two (weighted) Congressional voting networks represent similarities in voting patterns among members of the U.S. House of Representatives (US-HOUSE) and U.S. Senate (US-SENATE). Our construction follows prior work [9,74]. In particular, we represent these two data sets as "multilayer" temporal networks [9,75]. Each layer corresponds to a single two-year Congress, and an edge weight within a layer represents the voting similarity between two legislators during the corresponding Congress. In layer s, this yields adjacency elements of $A_{ij}^{(s)} = \frac{1}{b_{ij}(s)} \sum_k \gamma_{ijk}$, where $\gamma_{ijk} = 1$ if both legislators voted the same way on the kth bill, $\gamma_{ijk} = 0$ if they voted in different ways on that bill, $b_{ij}(s)$ is the number of bills on which both legislators voted during that Congress, and the sum is over bills. We identify different instances of the same legislator in different Congresses by connecting them using interlayer edges with weight ω [9]. (We use $\omega = 1$; the effect of changing ω has been investigated previously [74,76].) We represent each multilayer voting network using a single "supra-adjacency matrix" (see Refs. [75,77–79]) in which the different Congresses correspond to diagonal blocks and interlayer edges correspond to off-block-diagonal terms in the matrix. Throughout this paper, we treat the Congressional voting graphs as supra-adjacency matrices; we do not use any additional labeling or distinguished treatment of interlayer and intralayer edges (cf. [9]).

We chose these three sets of networks because (as we will see in later sections) they have *very* different properties with respect to their large-scale versus small-scale community structures. We thus emphasize that, with respect to the topic of this paper, these six networks are representative of several broad classes of previously studied networks: CA-GRQC and CA-ASTROPH are representative of the SNAP networks that were examined previously in Refs. [24–26]; both FB-HARVARD1 and FB-JOHNS55 (aside from a few very small communities in FB-HARVARD1) are representative of the FACEBOOK100 networks that were examined previously in Refs. [11,12]; and US-HOUSE and US-SENATE give examples of networks (that are larger than the Zachary Karate Club and caveman networks) on which conventional notions of and algorithms for community detection have been validated successfully [9,74].

In Table I, we provide summary statistics for each of the six networks. We give the numbers of nodes and edges in the largest connected component (LCC), the mean degree/strength ($\langle k_i \rangle$), the second-smallest eigenvalue (λ_2) of the normalized Laplacian matrix, and the mean clustering coefficient ($\langle C_i \rangle$). We use the local clustering coefficient $C_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{\frac{1}{3}}$, where $\hat{w}_{ij} = \frac{w_{ij}}{\max_{ij} w_{ij}}$, which reduces to the usual expression for local clustering coefficients in unweighted networks [80–82]. The high values for mean clustering coefficient in both the U.S. Congress and the coauthorship networks are unsurprising, given how those networks have been constructed. However, the latter is noteworthy, as the coauthorship networks are much sparser than the Facebook networks.

TABLE I. Six medium-sized to large networks. For each network, we show the number of nodes and edges in the LCC, the mean degree/strength ($\langle k_i \rangle$), the second-smallest eigenvalue (λ_2) of the normalized Laplacian matrix, the mean clustering coefficient ($\langle C_i \rangle$), prior references that used these networks, and a brief description.

	Nodes	Edges	$\langle k_i \rangle$	λ_2	$\langle C_i \rangle$	Refs.	Description
CA-GRQC	4158	13 422	6.5	0.0019	0.56	[24–26]	Coauthorship network: arXiv general relativity
CA-ASTROPH	17 903	196 972	22.0	0.0063	0.63	[24–26]	Coauthorship network: arXiv astrophysics
FB-JOHNS55	5157	186 572	72.4	0.1258	0.27	[11,12]	Johns Hopkins Facebook network
FB-HARVARD1	15 086	824 595	109.3	0.0094	0.21	[11,12]	Harvard Facebook network
US-SENATE	8974	422 335	60.3	0.0013	0.50	[9,74,76]	Network of voting patterns in the U.S. Senate
US-HOUSE	36 646	6 930 858	240.5	0.0002	0.58	[9,74,76]	Network of voting patterns in the U.S. House

Recall that the second-smallest eigenvalue λ_2 of the normalized Laplacian provides a qualitative notion of connectivity that can be used to bound the mixing time of diffusion-based dynamics on networks [47] (where larger values of λ_2 imply that there are fewer bottlenecks to mixing), and its associated eigenvector can be used to partition a graph into communities [3,65,68] (where smaller values of λ_2 correspond to the existence of better communities). We show the values of λ_2 for our six networks in Table I. For comparison, we show in Fig. 4 a scatter plot of λ_2 versus the size of the network (i.e., the number of nodes in the network) for these six networks, the remaining networks from SNAP [83] (black circles) that were also studied in [24,25], and the remaining 98 networks from the FACEBOOK100 data set (red stars) [11,12].

The first point to note about Fig. 4 is that nearly all of the FACEBOOK100 graphs have a λ_2 value that is much larger than those for the two collaboration graphs and the two voting graphs. Figure 4 and previous empirical results (from Refs. [24,25]) clearly demonstrate that the λ_2 values for the two collaboration graphs are representative of (and, in many cases, higher than) those of the other SNAP graphs studied empirically in Refs. [24,25]. That is, nearly all of the networks have λ_2 values that are much smaller than those in the FACEBOOK100 graphs. This implies, in particular, that the SNAP graphs contain more substantial bottlenecks to mixing than the FACEBOOK100 graphs. (Note, however, that the value of λ_2 says nothing about the size of the set of nodes that achieves the minimum.) In order to understand these differences, we study two networks from the FACEBOOK100 data set in detail: one (FB-JOHNS55) with a typical value of λ_2 and another (FB-HARVARD1) that is an "outlier," in that it has the lowest value of λ_2 in the entire FACEBOOK100 data set. (Interestingly, the FB-CALTECH36 network is the smallest network in the FACEBOOK100 data set—it has 762 nodes in its LCC—but it has the largest value of λ_2.)

The second point to note about Fig. 4 and Table I is that they suggest that FB-JOHNS55 (and possibly also FB-HARVARD1) are better connected than the other four networks, and that the connectivity properties of the two collaboration graphs and the two voting graphs (and perhaps also FB-HARVARD1) might be very similar. As we will see below, however, the situation is considerably more subtle.

In Fig. 5, we visualize the adjacency matrices of each of our six networks using a sparsity-pattern (Spy) plot. We draw

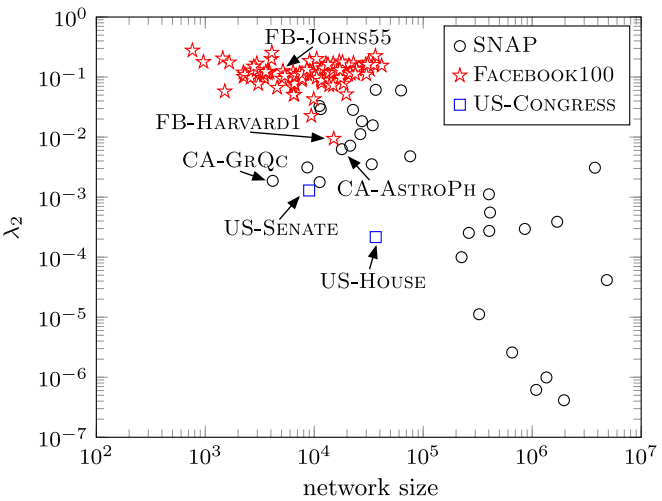


FIG. 4. (Color online) Scatter plot of the second-smallest eigenvalue (λ_2) of the normalized Laplacian versus size of the network for the networks from the SNAP data [83] that were studied in [24,25], all 100 networks in the FACEBOOK100 data set [11,12], and the two US-CONGRESS temporal networks [9,74,76].

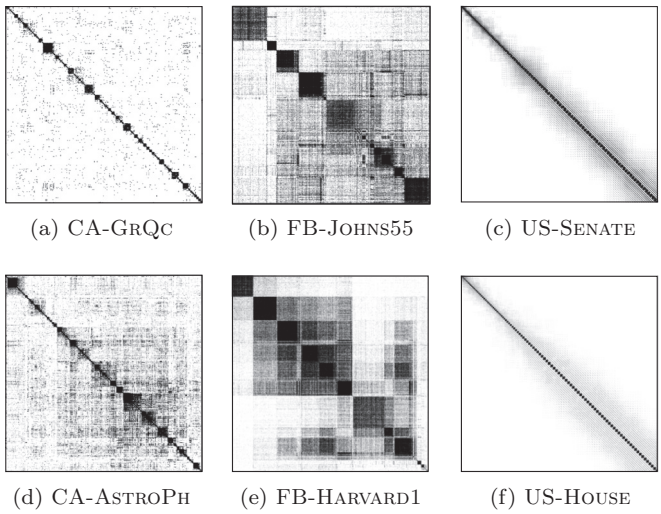


FIG. 5. Sparsity-pattern (Spy) plots for the LCC of each of our six example networks. We arrange the coauthorship networks (CA-GRQC and CA-ASTROPH) and Facebook networks (FB-JOHNS55 and FB-HARVARD1) according to communities that we obtain using an implementation [84] of a Louvain-like heuristic for modularity optimization [30]. For US-CONGRESS, we preserve the temporal order of the nodes starting with the first Congress in the top left and ending with the 110th Congress in the bottom right.

the nonzero entries of the adjacency matrices as black dots. The grayscale visualization in Fig. 5 is a result of coarsening the dpi resolution; it illustrates the density of connections in the corresponding areas of the adjacency matrices. This yields a visualization comparable to the idealized block models in Fig. 1. The node order in a Spy plot is arbitrary and, by permuting the nodes, can sometimes yield visualizations that are suggestive of structural features in a network. For the coauthorship and Facebook networks, we use results from a single run of an implementation [84] of a Louvain-like heuristic [30] for modularity optimization to partition these networks into communities. We then sort nodes by community assignment: We choose the order of communities manually to suggest potential large-scale structures. For the voting similarity networks, time provides a natural order for the nodes. We start with the nodes from the 1st Congress and end with the nodes from the 110th Congress. The small blocks on the diagonal are the individual Congresses, which are almost fully connected internally, and the off-diagonal blocks result from the interlayer coupling between the same individuals from different Congresses.

While certainly not definitive, Fig. 5 suggests several hypotheses about the relationship between small-scale structure and the large-scale structure—and, in particular, between small communities and large communities—in these six networks. First, from Figs. 5(c) and 5(f), it appears that the large-scale structure in US-SENATE and US-HOUSE corresponds to that of a "banded" matrix [85]. This banded structure is a result of the interlayer edges in these networks. Second, from Figs. 5(a) and 5(d), it appears that CA-GRQC and CA-ASTROPH both have many small-scale communities. It appears that they have a large-scale structure that is roughly banded, but there also appear to be many "long-range" off-diagonal interactions between distant nodes in the depicted ordering. Third, from Figs. 5(b) and 5(e), we observe that both FB-JOHNS55 and FB-HARVARD1 appear to have roughly ten communities that are both relatively large and relatively good.

From these visuals, it appears that nearly all of these communities have dense internal connections and sparse connections to other communities. Given the usual notion that a community is a set of nodes with denser connections among its constituent nodes than with the rest of a network, the visualizations in Fig. 5 appear to suggest that there might be interesting large-scale structure in FB-JOHNS55 and FB-HARVARD1 but not in the other networks. In particular, FB-JOHNS55 and FB-HARVARD1 seem to be examples of the case $\alpha_{11} \approx \alpha_{22} \gg \alpha_{12}$ that we illustrated in Fig. 1(a).

The focus of the present investigation is to test the level of correctness of the above hypotheses about the relationship between small-scale structure and large-scale structure in these six networks. As we have discussed, intuition like we have illustrated in Fig. 5 is common in the development and validation of methods for community detection, so it is useful to delve into great depth on a set of networks to explore the connections between small-scale and large-scale structures in networks. As we see in the next several sections, the situation is considerably more subtle than these figures (and commonly employed intuition) might suggest. For example, with the exception of the small communities in CA-GRQC/CA-ASTROPH and the large-scale structure (i.e., the one-dimensional temporal ordering) in US-SENATE and US-HOUSE, these intuitive hypotheses about the relationship between the local structure and the global structure in these networks are not unambiguously supported by other evidence. Similarly, many communities that appear to be "good" based on the usual intuition and visualizations like that in Fig. 5 often are judged as largely artifactual from the perspective of quantitative measures of community quality.

B. Network community profiles

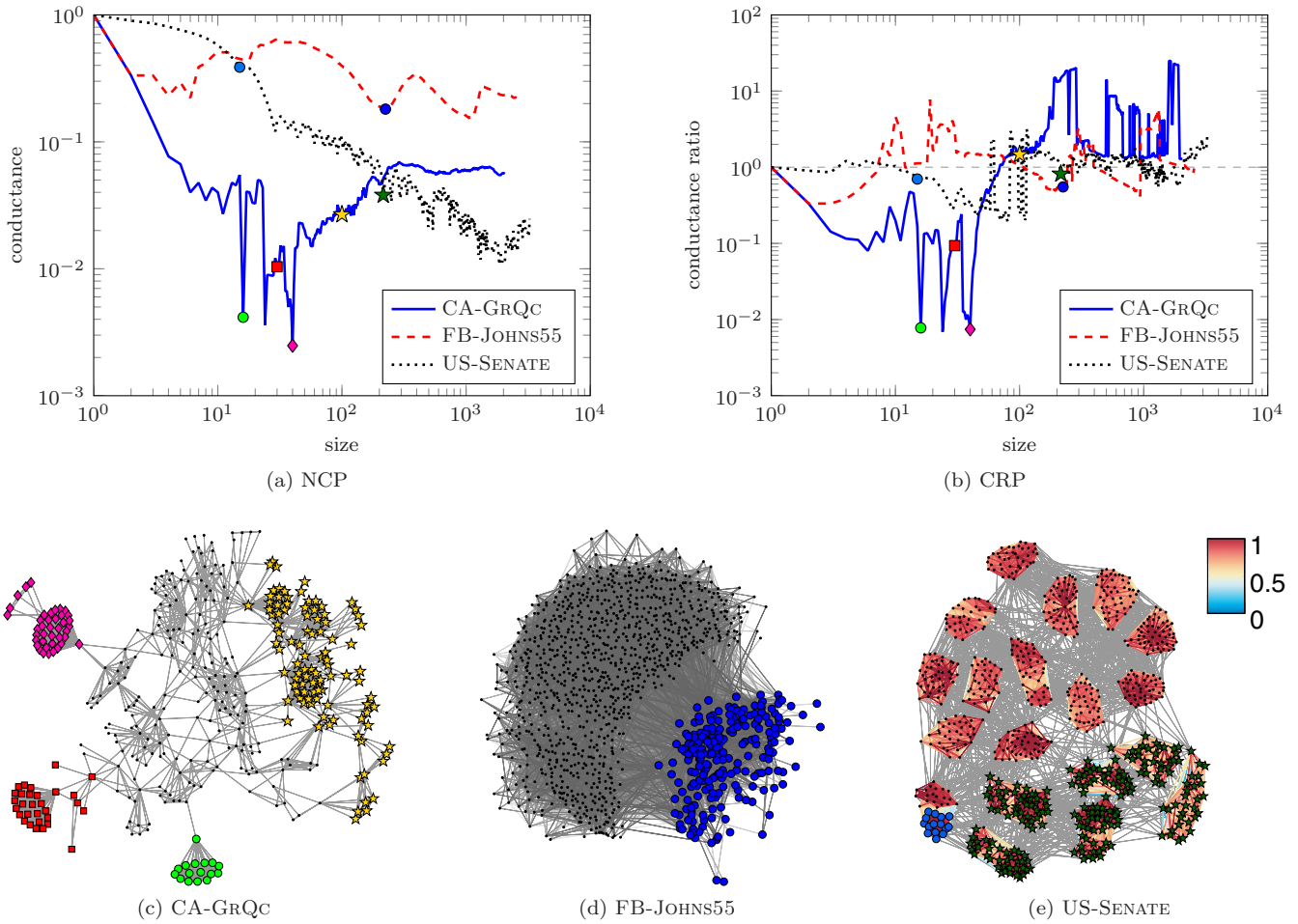
We start by presenting our main results from using the ACLCUT method (see Figs. 6 and 7). We obtain similar insights about global structure using the MOVCUT (see Appendix C) and EGONET (see Appendix D) methods, although they can exhibit rather different local behavior.

In Fig. 6, we show the NCPs and CRPs for the smaller network from each of the three pairs of networks from Table I. In Fig. 7, we show the results for the corresponding larger networks. Note the logarithmic scale for both the vertical and the horizontal axes in these figures as well as in subsequent NCP and CRP plots. Observe from Figs. 6(a) and 7(a) that the NCPs for networks of the same type are qualitatively similar, whereas NCPs for networks of different types have qualitatively distinct shapes.

For the coauthorship networks CA-GRQC and CA-ASTROPH, the NCPs have a mostly upward-sloping shape, except for the region with fewer than 100 nodes. We conclude that CA-GRQC and CA-ASTROPH have good small (e.g., consisting of tens of nodes) communities, but they do not have good large (e.g., consisting of hundreds or thousands of nodes) communities. These results are consistent with the NCPs of LIVEJOURNAL from Fig. 2(b) and with the results of [24–26]. Additionally, the high values for the CRPs for the coauthorship networks [see Figs. 6(b) and 7(b)] for communities with hundreds or thousands of nodes reveals that these large communities are loosely connected collections of good, small communities. This feature is also visible in Fig. 6(c), which shows selected communities and their neighborhoods for the CA-GRQC network.

For the Facebook networks FB-JOHNS55 and FB-HARVARD1, all of the communities at every size (except for two small "communities" with five and ten nodes in FB-HARVARD1 [88]) have very large conductances (greater than 10^{-1}). This indicates that the communities in this network all have very poor community quality, in sharp contrast with the coauthorship and voting networks. The essentially flat shape for the NCPs of the Facebook networks illustrate that these networks have strong expanderlike properties (see Appendix A) and, relatedly, that there are no substantial bottlenecks to the rapid mixing of random walks on these networks. Both Facebook networks have noticeable dips in their NCPs at larger community sizes (about 220 and 1100 nodes for FB-JOHNS55 and about 1500 nodes for FB-HARVARD1), and the sets of nodes associated with each of these dips correlate strongly with self-reported demographic information [89].

For the voting networks US-SENATE and US-HOUSE, the NCP has a predominantly downward-sloping shape. This is characteristic of "low-dimensional" networks in the sense that we described informally in Sec. II B. The informal



(a) NCP (b) CRP

(c) CA-GrQc (d) FB-Johns55 (e) US-Senate

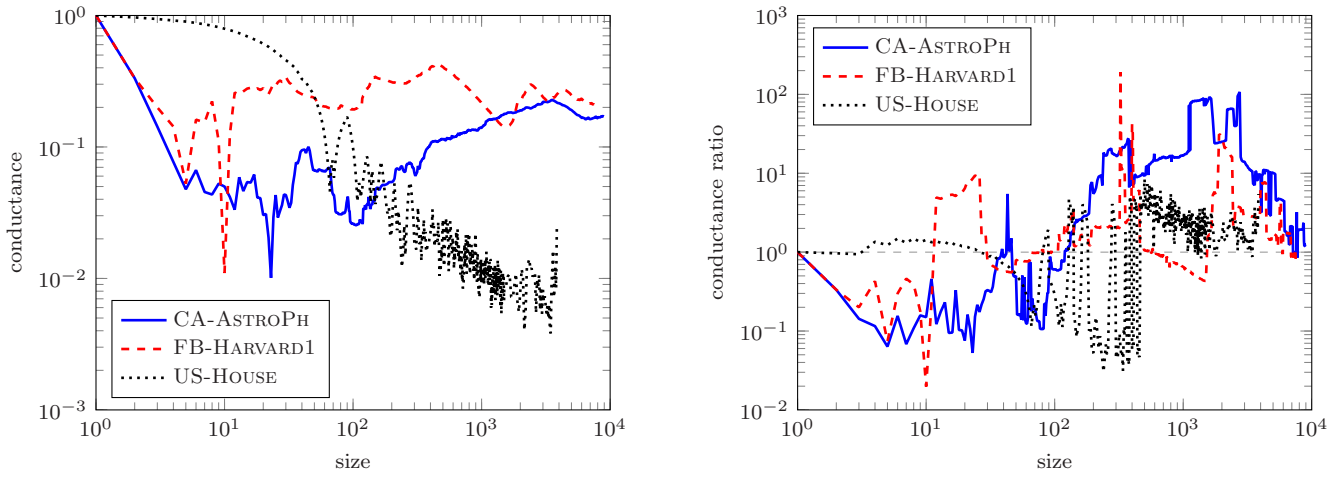
FIG. 6. (Color online) NCP plots [in panel (a)] and conductance ratio profile (CRP) plots [in panel (b)] for CA-GrQc, FB-Johns55, and US-Senate (i.e., the smaller network from each of the three pairs of networks from Table I) generated using the AclCut method. In panels (c)–(e), we show modified Kamada-Kawai [86] spring-embedding visualizations that emphasize community structure [87] of corresponding (color-coded) communities and their neighborhoods (a 2-neighborhood for CA-GrQc, a 1-neighborhood for FB-Johns55, and all Senates that have at least one senator in common with those in the communities for US-Senate). We find good small communities but no good large communities in CA-GrQc, some weak large-scale structure in FB-Johns55 that does *not* create substantial bottlenecks for the random-walk dynamics, and signatures of low-dimensional structure (i.e., good large communities but no good small communities) for US-Senate. The low-dimensional structure in US-Senate results from the multilayer structure that encapsulates the network's temporal properties. [The dashed line in panel (b) indicates a conductance ratio of 1.]

reason for the downward-sloping shape is that US-Senate and US-House each consist of a low-dimensional structure that is evolving along a one-dimensional scaffolding (i.e., time), upon which the detailed structure of individual Congresses (i.e., a good partition that is nearly along party lines) is superimposed. (One can examine such structures by using smaller values of the interlayer coupling parameter ω; see Ref. [74].) This is consistent with previous results [90].

These results, which illustrate that community quality changes very differently with size in each of the three pairs of networks, also indicate that these three types of networks have very different properties with respect to large-scale versus small-scale community structure. Moreover, the qualitative similarity in behavior between the two networks in each pair suggests that the coarse behavior of an NCP (downward-sloping, upward-sloping, or flat) is indicative of large classes of networks and not an artifact of our particular choice of example networks. One obtains similar insights about global structure using the MovCut (see Appendix C) and EgoNet (see Appendix D) methods, although they can exhibit rather different local behavior. We investigate these differences in local behavior in Sec. IV C.

C. Comparison of results from AclCut, MovCut, and EgoNet

The NCPs generated using either AclCut or MovCut (see Appendix C), and to a somewhat lesser extent those generated using EgoNet (see Appendix D), have similar global features—i.e., they exhibit the same general trends and have dips at small size scales that correspond to nearly identical communities—indicating that we obtain a broadly similar picture of the large-scale community structure by using any of the methods. However, the detailed local behavior of the three methods can differ considerably. Such behavior depends sensitively on the choice of seed node, the choice of the



(a) NCP (b) CRP

FIG. 7. (Color online) NCP plots [in panel (a)] and CRP plots [in panel (b)] for CA-AstroPh, FB-Harvard1, and US-House (i.e., the larger network from each of the three pairs of networks from Table I) generated using the AclCut method.

parameters in the different methods, and the specific details of each method. In this section, we discuss the similarities and differences in the results from these methods. We only show the results of calculations for the smaller network from each of the three network pairs in Table I (but we observe similar results for the larger networks).

All three methods return a vector that can be used to rank the nodes, and it is this ranking that determines the community memberships. Both the AclCut and the MovCut methods return a variant of a personalized PageRank (PPR) vector, which we parametrize using a teleportation parameter α. Additionally, the approximate calculation performed by the AclCut method depends on a truncation parameter ϵ. The EgoNet method returns an EgoRank vector that only depends on the seed node and does not have any parameters. We describe the methods and their parameters in detail in Appendix B.

To compare different methods, we note that any meaningful difference between the three methods should manifest as a difference in the rank order of nodes, as this determines the assignment of nodes to local communities. We quantify rank differences by computing the Spearman rank correlation [91] between the (exact for MovCut and approximate for AclCut) personalized PageRank (PPR) and EgoRank ranking vectors. We also restrict all comparisons to the support of the corresponding approximate PPR vector that we obtain using the AclCut method. This induces an indirect dependency of the results from MovCut and EgoNet on α and ϵ (in addition to the direct dependency of MovCut on α).

In Tables II–IV, we show the results of our calculations of Spearman rank correlations. For each of the three networks, we select 50 seed nodes by sampling uniformly without replacement. (Recall that a single run of a method uses a single seed node.) We then compute PPR vectors for these seed nodes using the AclCut and MovCut method for different values of the truncation parameter ϵ and teleportation parameter α, and we also compute the EgoRank vector for each of the seed nodes. As we discuss in Appendix B, *smaller* values of α correspond to *more local* versions of the procedures, but *larger* values of ϵ correspond to *more local* versions of the procedures.

The AclCut and MovCut methods give very similar results for most of the 50 seed nodes in our sample, although (as we discuss below) some seed nodes do yield noticeable differences. The two methods give the most similar results for FB-Johns55 (a mean Spearman correlation of 0.92; a minimum of 0.43), whereas we find larger deviations in both CA-GrQc (mean 0.85; minimum -0.13) and US-Senate (mean 0.86; minimum -0.44). Note that we calculate the mean, maximum, and minimum over all sampled seed nodes and parameter values.

Interestingly, the larger deviations between the two methods for CA-GrQc and US-Senate occur at different values of the truncation parameter ϵ. For CA-GrQc (and, to a lesser extent, for FB-Johns55), we obtain the largest deviations for smaller values (e.g., $\epsilon = 10^{-6}$). For US-Senate, however, we obtain the largest deviations for $\epsilon = 10^{-4}$. See the bold values in Tables II–IV. This is consistent with the very different isoperimetric properties of these three networks, as revealed by their NCPs, as well as with well-known connections between conductance and random walks.

There are two potential causes for the differences between our results for the AclCut and MovCut methods. First, there is a truncation effect, governed by the parameter ϵ, in approximating the PPR vector using the AclCut method. As ϵ becomes smaller, the approximation in AclCut becomes more accurate and this effect diminishes. Second, the two methods differ in the precise way that they use a seed vector to represent a seed node. As we discuss in Appendix B, the AclCut method uses an indicator vector $\vec{s}$ to represent a seed node i; thus, we use $\vec{s}_i = 1$ whenever i is a seed node, and we set all other entries in that vector to 0. In contrast, the MovCut method projects the indicator vector onto the orthogonal complement of the strength vector to ensure that $\vec{s}^T D\vec{1} = 0$ (see Appendix B). This effect decreases as $\alpha \to 1$.

The largest deviations between the two methods occur for smaller values of ϵ in CA-GrQc and FB-Johns55; for these values of ϵ, the truncation effect is small, suggesting that

TABLE II. Pairwise comparison of the three methods using the Spearman rank correlation between the ranking vectors from the ACLCUT (A), MOVCUT (M), and EGONET (E) methods for CA-GRQC. We use a uniform random sample of 50 nodes for each of several values for the teleportation parameter α and truncation parameter ϵ. We take the maximum, mean, and minimum over the seed nodes. Bold values highlight the largest deviations between the ACLCUT and MOVCUT methods for a given value of α.

		α														
		0.6			0.7			0.8			0.9			0.99		
		A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E
ϵ																
10^{-3}	Max	1.00	0.99	0.99	1.00	0.98	0.98	1.00	0.97	0.97	1.00	0.95	0.95	0.98	0.89	0.86
	Mean	0.98	0.78	0.77	0.98	0.76	0.73	0.98	0.72	0.68	0.97	0.68	0.62	0.91	0.60	0.48
	Min	0.92	0.26	0.23	0.91	0.18	0.14	0.94	0.21	0.15	0.85	−0.01	−0.05	0.74	0.25	0.06
10^{-4}	Max	1.00	0.97	0.97	1.00	0.97	0.97	1.00	0.94	0.94	0.99	0.89	0.85	**0.92**	0.69	0.53
	Mean	0.99	0.74	0.72	0.98	0.73	0.68	0.97	0.70	0.64	0.95	0.63	0.54	0.89	0.51	0.36
	Min	0.96	0.16	0.10	0.91	−0.05	−0.19	0.84	0.35	−0.02	0.88	0.43	0.25	0.85	0.32	0.18
10^{-5}	Max	1.00	0.96	0.95	0.97	0.92	0.87	0.94	0.81	0.69	0.89	0.73	0.55	0.93	0.75	0.60
	Mean	0.91	0.74	0.58	0.89	0.69	0.51	0.85	0.65	0.42	0.78	0.62	0.33	**0.84**	0.63	0.36
	Min	0.24	0.21	−0.20	0.42	0.30	−0.10	0.42	0.39	−0.13	0.25	0.43	−0.10	**0.49**	0.44	0.05
10^{-6}	Max	**0.84**	0.85	0.68	**0.79**	0.81	0.49	**0.70**	0.79	0.39	**0.80**	0.81	0.47	0.93	0.75	0.60
	Mean	**0.62**	0.72	0.25	**0.57**	0.69	0.17	**0.51**	0.69	0.12	**0.51**	0.72	0.13	0.85	0.63	0.37
	Min	**0.01**	0.57	−0.36	**0.07**	0.52	−0.30	**−0.06**	0.50	−0.27	**−0.13**	0.57	−0.32	0.52	0.46	0.08

the different way of representing a seed node is partially responsible for the difference between the results of the two methods for these networks. For larger values of ϵ (in particular, for $\epsilon \geqslant 10^{-4}$), where the support of the approximate PPR vector from the ACLCUT method is small, the behavior of the two methods is very similar. Consequently, the differences in the choice of seed vector become more important for nodes that are "far away" from the seed node, in the sense that they are rarely visited by the PPR dynamics that underlie these methods. As a result, the "local NCPs" for the two methods in Figs. 8(a) and 9(a) are largely identical for small community sizes but differ for large community sizes. (We use the term *local NCP* to refer to an NCP that we compute using only a single seed node without optimizing over the results from multiple seed choices; see Ref. [37] for details on the construction of local NCPs.)

For US-SENATE, the two methods behave almost identically for small ϵ (see Table IV), so we conclude that the different ways of representing a seed node have only a small effect on this network. However, the truncation effect is more pronounced in this network compared with CA-GRQC or FB-JOHNS55. This feature manifests as larger deviations between ACLCUT and MOVCUT in Table IV for large ϵ and small α (i.e., where the truncation has the strongest impact). The discrepancy occurs because the ACLCUT method initially pushes a large amount of probability to the interlayer

TABLE III. Pairwise comparison of the three methods using the Spearman rank correlation between the ranking vectors from the ACLCUT (A), MOVCUT (M), and EGONET (E) methods for FB-JOHNS55. We use a uniform random sample of 50 nodes for each of several values for the teleportation parameter α and truncation parameter ϵ. We take the maximum, mean, and minimum over the seed nodes. Bold values highlight the largest deviations between the ACLCUT and MOVCUT methods for a given value of α.

		α														
		0.6			0.7			0.8			0.9			0.99		
		A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E
ϵ																
10^{-3}	Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Mean	1.00	0.85	0.84	1.00	0.82	0.81	0.99	0.79	0.79	0.98	0.78	0.77	0.98	0.78	0.75
	Min	0.95	0.61	0.56	0.94	0.54	0.47	0.80	0.55	0.53	0.80	0.55	0.49	0.77	0.40	0.39
10^{-4}	Max	1.00	0.89	0.89	1.00	0.93	0.93	1.00	0.91	0.91	1.00	0.89	0.89	1.00	0.88	0.88
	Mean	0.98	0.48	0.47	0.97	0.45	0.44	0.96	0.43	0.41	0.95	0.41	0.37	0.94	0.38	0.33
	Min	0.81	0.06	0.05	0.78	0.05	0.04	0.74	0.07	0.06	0.69	0.07	0.07	0.69	0.07	0.00
10^{-5}	Max	1.00	0.85	0.85	1.00	0.86	0.85	1.00	0.84	0.83	0.99	0.84	0.82	1.00	0.81	0.78
	Mean	0.98	0.58	0.54	0.97	0.59	0.54	0.97	0.63	0.57	0.96	0.61	0.54	0.95	0.53	0.46
	Min	0.90	0.24	0.18	0.91	0.26	0.17	0.91	0.23	0.14	0.89	0.22	0.07	0.88	0.18	0.04
10^{-6}	Max	**0.99**	0.75	0.46	**0.97**	0.69	0.46	**0.91**	0.71	0.39	**0.90**	0.73	0.47	**0.97**	0.69	0.57
	Mean	**0.79**	0.41	0.20	**0.75**	0.45	0.13	**0.72**	0.56	0.20	**0.77**	0.62	0.32	**0.88**	0.59	0.41
	Min	**0.57**	0.23	−0.07	**0.49**	0.24	−0.06	**0.43**	0.32	−0.02	**0.49**	0.32	0.01	**0.60**	0.26	0.07

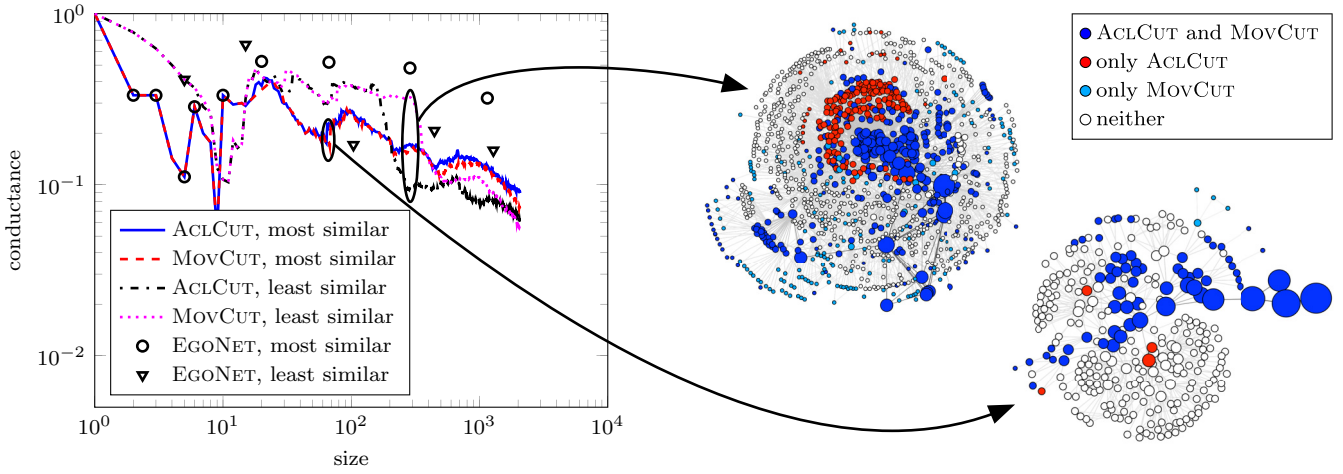
TABLE IV. Pairwise comparison of the three methods using the Spearman rank correlation between the ranking vectors from the ACLCUT (A), MOVCUT (M), and EGONET (E) methods for US-SENATE. We use a uniform random sample of 50 nodes for each of several values for the teleportation parameter α and truncation parameter ϵ. We take the maximum, mean, and minimum over the seed nodes. Bold values highlight the largest deviations between the ACLCUT and MOVCUT methods for a given value of α.

		α														
		0.6			0.7			0.8			0.9			0.99		
		A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E	A-M	A-E	M-E
ϵ	Max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10^{-3}	Mean	**0.65**	0.66	0.40	0.68	0.71	0.30	0.74	0.77	0.46	**0.74**	0.85	0.57	**0.80**	0.80	0.56
	Min	0.02	0.50	−0.05	0.38	0.32	−0.24	0.46	0.53	−0.07	0.50	0.68	−0.02	**0.51**	0.62	−0.00
	Max	1.00	0.93	0.87	1.00	0.92	0.77	1.00	0.93	0.82	1.00	0.89	0.75	**0.99**	0.81	0.61
10^{-4}	Mean	0.75	0.61	0.36	**0.54**	0.59	0.17	**0.67**	0.62	0.34	0.86	0.50	0.34	0.92	0.36	0.23
	Min	**−0.03**	−0.16	−0.51	**−0.32**	−0.11	−0.63	**−0.44**	0.15	−0.36	**0.50**	−0.08	−0.26	0.77	−0.20	−0.38
	Max	1.00	0.91	0.90	1.00	0.89	0.87	1.00	0.85	0.83	1.00	0.80	0.78	1.00	0.85	0.84
10^{-5}	Mean	0.99	0.58	0.54	0.99	0.48	0.46	0.99	0.42	0.40	0.96	0.44	0.40	0.95	0.43	0.39
	Min	0.86	0.03	0.01	0.88	−0.02	−0.02	0.88	−0.09	−0.13	0.80	−0.17	−0.20	0.80	0.07	0.01
	Max	1.00	0.84	0.84	1.00	0.84	0.84	1.00	0.84	0.85	1.00	0.87	0.86	1.00	0.96	0.96
10^{-6}	Mean	0.94	0.54	0.48	0.98	0.55	0.50	0.99	0.62	0.60	1.00	0.69	0.68	0.99	0.91	0.89
	Min	0.85	0.30	0.22	0.89	0.08	0.19	0.98	0.25	0.21	0.98	0.30	0.29	0.96	0.83	0.73

neighbors of the seed node (i.e., to the same senator in different Congresses). This probability does not diffuse to other nodes for sufficiently large values of ϵ.

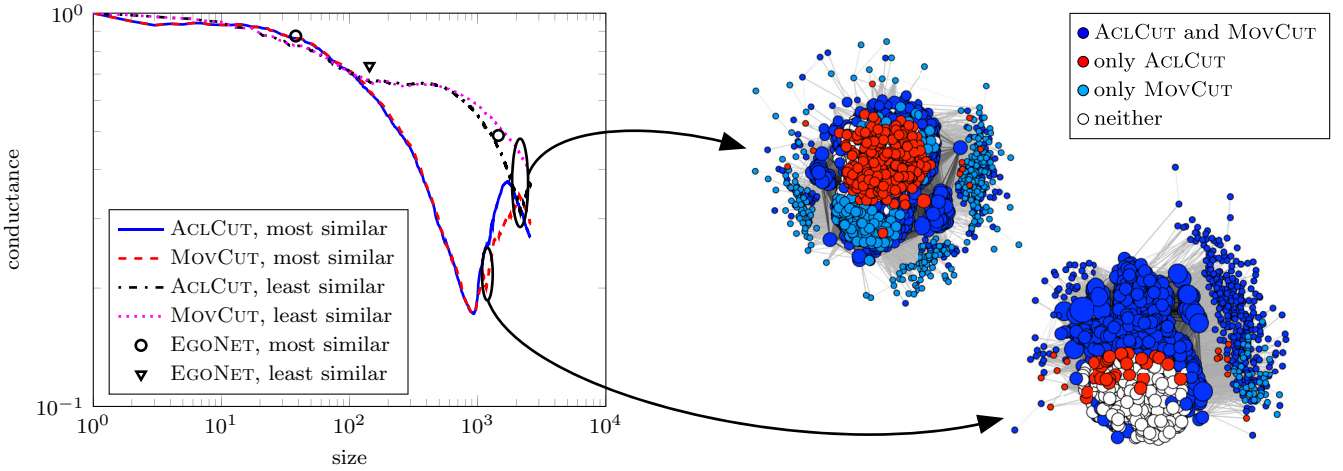
In Figs. 8–10, we illustrate the results from Tables II–IV. In these figures, we plot the local NCPs for CA-GRQC, FB-JOHNS55, and US-SENATE for the seed nodes (from the sample of 50) that yield the largest and smallest mean Spearman rank correlation between the ACLCUT and MOVCUT methods. (That is, one seed node yields the largest mean Spearman correlation, and the other yields the smallest mean Spearman correlation.) In these figures, we also include visualizations of example communities that we obtain from the ACLCUT and MOVCUT methods using a Kamada-Kawai-like spring-embedding visualization [87] of the k-ego-nets of these seed nodes.

From the visualizations of the local communities, it seems for CA-GRQC (see Fig. 8) and FB-JOHNS55 (see Fig. 9) that nodes included in local communities obtained from ACLCUT tend to be closer in geodesic distance to the seed node than those obtained from MOVCUT. (To see this, observe that red nodes tend to be larger than light blue nodes in the visualization of the k-neighborhoods.) If this observation holds



(a) Local NCP (b) Spring-embedding visualizations

FIG. 8. (Color online) Comparison of methods for CA-GRQC. (a) Local NCPs for the seed nodes (of the 50 nodes that we sample) with the highest and lowest mean Spearman correlation over the sampled parameter values. These NCPs highlight the difference in behavior of the ACLCUT and MOVCUT methods for large communities. (b) Kamada-Kawai-like spring-embedding visualization [87] of (bottom right) the 9-neighborhood of the seed node with the smallest difference between the two methods and (top left) the 6-neighborhood of the seed node with the largest difference. In these two visualizations, the node size decreases as a function of geodesic distance from the seed node. We color the nodes according to whether they belong to the local community that we obtain using the ACLCUT method, the one that we obtain using the MOVCUT method, the intersection of the two communities, or neither community.

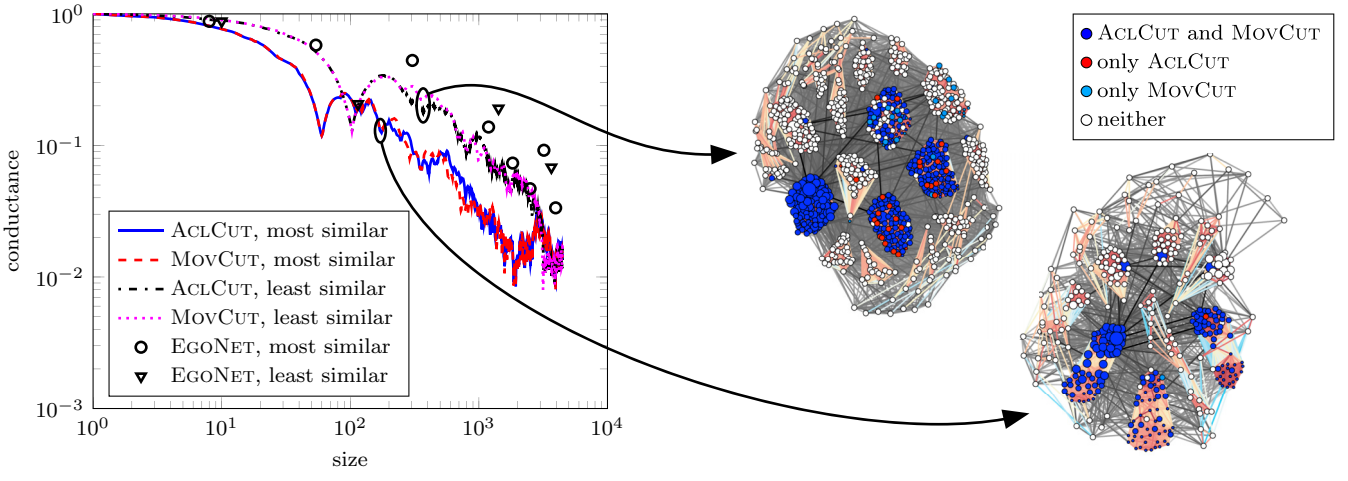


(a) Local NCP (b) Spring-embedding visualizations

FIG. 9. (Color online) Comparison of methods for FB-JOHNS55. (a) Local NCPs for the seed nodes (of the 50 nodes that we sample) with the highest and lowest mean Spearman correlation over the sampled parameter values. These NCPs highlight the difference in behavior of the ACLCUT and MOVCUT methods for large communities. (b) Kamada-Kawai-like spring-embedding visualization [87] of the 2-neighborhoods of both seed nodes. The one with the smallest difference is on the bottom right, and the one with the largest difference is on the top left. In these two visualizations, the node size decreases as a function of geodesic distance from the seed node. The smallest nodes are more than two steps away from the seed node, but they appear in at least one of the local communities. We color the nodes according to whether they belong to the local community that we obtain using the ACLCUT method, the one that we obtain using the MOVCUT method, the intersection of the two communities, or neither community.

more generally and is not just an artifact of the particular communities that we show in Figs. 8 and 9, then we should obtain higher Spearman rank correlations between ACLCUT and EGONET than between MOVCUT and EGONET. Indeed, Tables II–IV consistently show this feature for almost all choices of ϵ and α and for all three networks. Note that this feature is also present in US-SENATE, though it is less prominent in its k-neighborhood visualization than is the case for the other two networks.

Figures 8–10 also reveal that the three networks look very different from a local perspective. For FB-JOHNS55 (see Fig. 9), the two seed nodes (of the sample of 50) that we explore in detail allow us to reach a large fraction of all nodes after just two steps. This is consistent with known properties of



(a) Local NCP (b) Spring-embedding visualizations

FIG. 10. (Color online) Comparison of methods for US-SENATE. (a) Local NCP for the seed nodes (of the 50 nodes that we sample) with the highest and lowest mean Spearman correlation over the sampled parameter values. These NCPs highlight the difference in behavior of the ACLCUT and MOVCUT methods for large communities. (b) Kamada-Kawai-like spring-embedding visualization [87] of the 3-neighborhoods of both seed nodes. The one with the smallest difference is on the bottom right, and the one with the largest difference is on the top left. In these two visualizations, the node size decreases as a function of geodesic distance from the seed node. We color the nodes according to whether they belong to the local community that we obtain using the ACLCUT method, the one that we obtain using the MOVCUT method, the intersection of the two communities, or neither community.

the full Facebook graph (circa 2012) of individuals connected by reciprocal "friendships." For example, the mean geodesic distance between pairs of nodes of the Facebook graph is very small: It was recently estimated by Facebook's Data Science Team and their collaborators to be about 4.74 [92]. Additionally, one can view Facebook as a collection of ego networks that have been patched together into a network whose global structure is sparse [44] (and such structure is an important motivation for the locally biased notion of community structure that we advocate in this paper).

For CA-GrQc, we obtain very different neighborhoods starting from our two different seed nodes. The node in the sample that results in the largest difference in behavior between the ACLCUT and the MOVCUT methods appears to be better connected in the network in the sense that the k-neighborhood (for any k until saturation occurs) is much larger than that of the node that results in the smallest difference. (That is, it is more in the "core" than in the "periphery" of the nested core-periphery structure of Refs. [24,25].) We observe a similar phenomenon for FB-JOHNS55 and US-SENATE. Furthermore, its 1-ego-net and 2-ego-net are highly clustered, in the sense that they contain many closed triangles. For the seed node that results in the smallest difference between the ACLCUT and MOVCUT methods, we need to consider the 6-ego-net (which has 20 nodes) to obtain a network of similar size to the 2-ego-net (which has 15 nodes) for the seed node with the largest difference. In the case of the seed node in our sample that results in the smallest difference between the two methods, even the 6-ego-net appears rather treelike; it contains few closed triangles and no larger cliques.

For US-SENATE, the 1-neighborhood of any seed node contains only the node itself and those corresponding to the same senator in different Congresses [93]. As one begins to consider nodes that are further away, one first reaches corresponding senators in other Congresses before reaching other senators with similar voting patterns from the same Congress. This behavior of the EGONET method contrasts with the (PageRank-based) ACLCUT and MOVCUT methods, which tend to initially select all senators from one Congress before reaching senators from other Congresses.

D. Meso-scale structure

From the perspective of the locally biased community-detection methods that we use in this paper, one can view intermediate-sized (i.e., meso-scale) structures in networks as arising from collections of local features—e.g., via overlaps of local communities that one obtains algorithmically using locally biased dynamics such as those that we consider. Such local features depend not only on the network adjacency matrix but also on the dynamical process under study, the initial seed(s) from which one is viewing a network, and the locality parameters of the method (which corresponds to the dynamical process) that determine how locally one is viewing the network. Although a full discussion of the relationship between local, meso-scale, and global structures is beyond the scope of this paper, we provide a few examples of such results.

To try to visualize the meso-scale and global network structures that we obtain from the local communities that we identify, we define an $n \times n$ association matrix $\widetilde{A}$ (where n is again the number of nodes in a network), which encodes pairwise relations between nodes based on a sample of local communities. For a given sample $\mathcal{S}$ of local communities (obtained, e.g., by running a given method with many seed nodes and values of a locality parameter), the entries of the association matrix are given by the number of times that a pair of nodes appear together in a local community, normalized by the number of times that either of them appears. That is, the elements of the association matrix are

$$\widetilde{A}_{ij} = \frac{|\{S \in \mathcal{S} : i \in S \text{ and } j \in S\}|}{|\{S \in \mathcal{S} : i \in S \text{ or } j \in S\}|}. \tag{8}$$

Our procedure for extracting global network structure from a sampled set of communities is similar in spirit to computing association (or "coclassification") matrices that have been constructed from sampling a landscape of the modularity objective function [94], and one can, in principle, analyze these matrices further using the same methods. The additional normalization in our definition of association matrices is necessary to correct for the oversampling of large communities relative to small communities (which results from sampling nodes uniformly at random). At first glance, association matrices computed by sampling a modularity landscape appear to reveal much clearer community structure in these networks than we obtain by sampling local communities. However, this is largely an artifact of the well-known resolution limit of modularity optimization [35]. One can mitigate this effect by using one of the multiresolution generalizations of modularity [95,96] to sample the modularity landscape across different values of the resolution parameter. This yields association matrices that are similar in appearance to the ones that we obtain by sampling local communities.

To visualize the association matrices in a way that reveals global network structure, it is important to find a good node order. We found the sorting method suggested in Ref. [94] to be impractically slow for the networks that we study. Instead, we sort the nodes based on the optimal leaf ordering [97] for the average-linkage hierarchical clustering tree of the association matrix. (For US-SENATE, we do this procedure within a given Congress, and we then use the natural temporal ordering to define the inter-Congressional ordering.)

In addition, to see small-scale structure using samples obtained from MOVCUT, we use a community-size parameter c that limits the volume of the resulting community based on the desired correlation with the seed vector. For the association matrices in panels (d) and (e) of Fig. 11, we sample communities using $c \in \{10^i : i = 1, \ldots, 5\}$. See Ref. [37] for details. We summarize our results in Figs. 11–14.

In Fig. 11, we show the result of applying this procedure with communities that we sample using the ACLCUT, MOVCUT, and EGONET methods. In each case, we keep only the best conductance community for each sampled ranking vector. The most obvious feature of the visualizations in Fig. 11 is that—except for US-SENATE, for which there is a natural large-scale global structure defined by the one-dimensional temporal ordering—the visualizations are much more complicated than any of the idealized structures in Fig. 1 (which suggests that the visualizations in Fig. 1 might be revealing at least as much about the inner workings

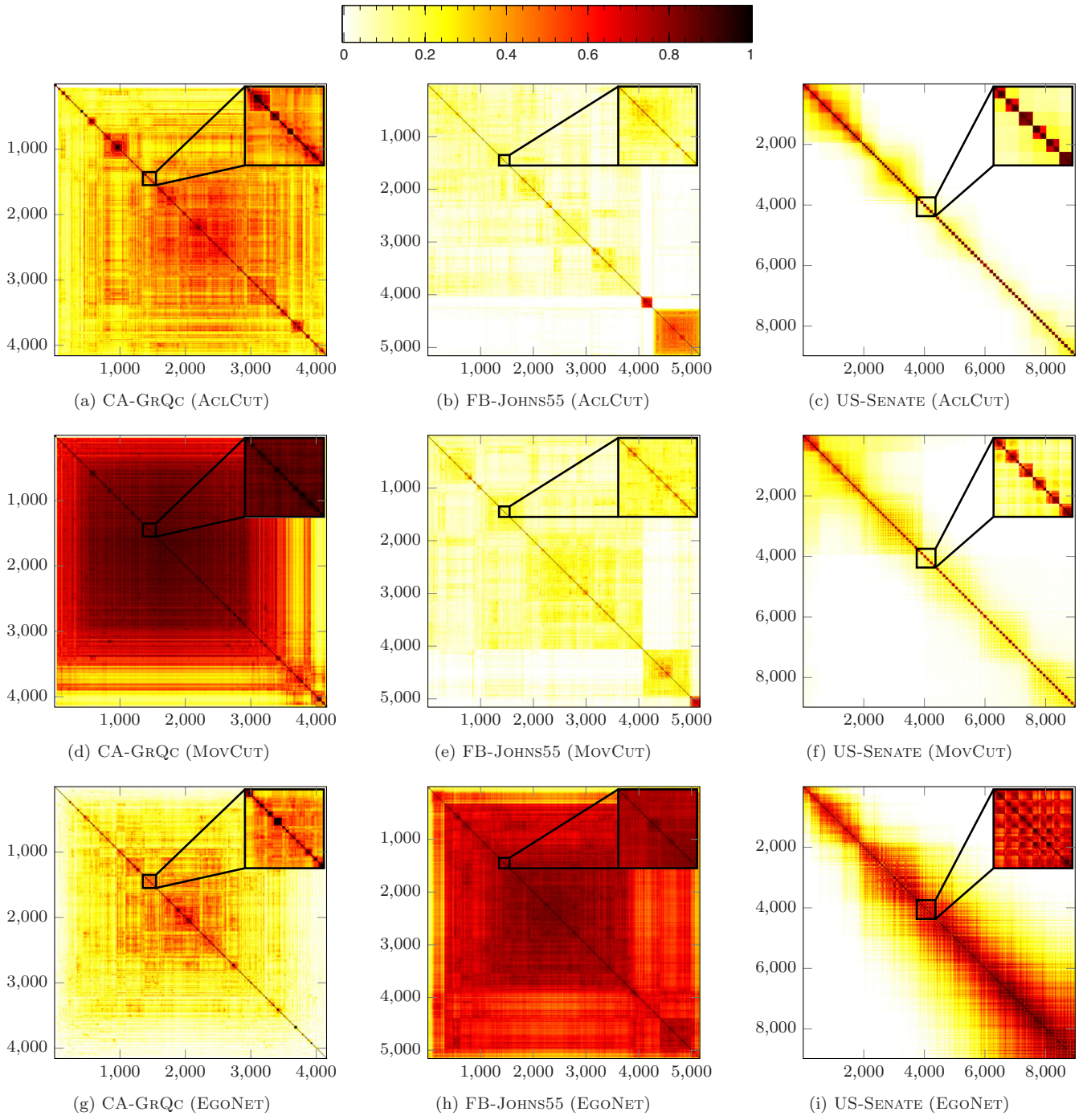
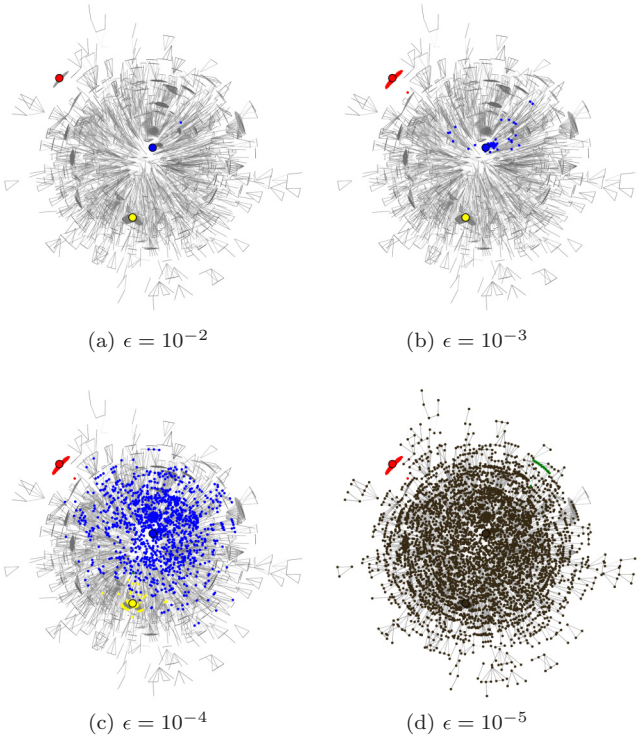


FIG. 11. (Color online) Visualizations of association matrices for CA-GrQc, FB-Johns55, and US-Senate illustrate how meso-scale and global structures emerge from the superposition (and overlap) of many local communities. See the main text for a description of how we construct the association matrices. For each of the three networks, we generate the subfigures using the same three sampling procedures that we use to generate the NCPs: We use AclCut for panels (a)–(c), we use MovCut for panels (d)–(e), and we use EgoNet for panels (g)–(i). (We color the plots according to the strength of association.)

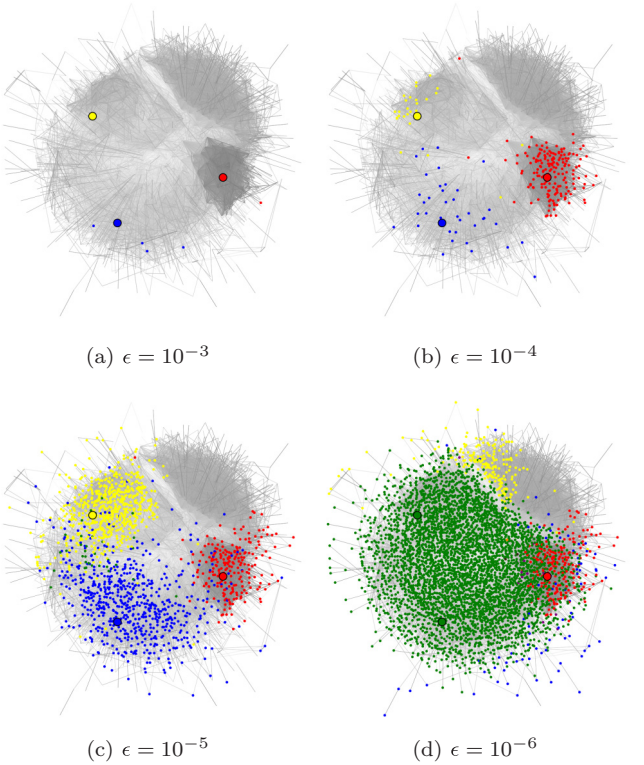
of the visualization algorithm as about the networks being visualized). The structures in Fig. 1 are trivially interpretable, whereas those in real networks (e.g., as illustrated in Fig. 11) are extremely messy and very difficult to interpret. In the paragraphs below, we discuss the structural features in Fig. 11 in more detail.

For CA-GrQc (see Figs. 11 and 12), we observe many small communities that are composed of about 10–100 nodes. These communities, which correspond to the dark red blocks along the diagonal [see the inset in Fig. 11(a)], are responsible for the dips in the NCPs [see Figs. 6(a), 16(a), and 18(a)] for this network. However, these small communities do not



(a) $\epsilon = 10^{-2}$ (b) $\epsilon = 10^{-3}$ (c) $\epsilon = 10^{-4}$ (d) $\epsilon = 10^{-5}$

FIG. 12. (Color online) Visualization of global structure in CA-GRQC. We construct the network layout by weighting each edge using the corresponding entry of the association matrix for the ACLCUT method [see Fig. 11(a)]. We then apply the Kamada-Kawai spring-embedding visualization algorithm [86,87] to the resulting weighted network. For ease of visualization, we only plot edges whose weight is larger than the mean edge weight. Colored nodes correspond to local communities for three different seed nodes. We draw nodes that are members of more than one community using a mixed color (e.g., blue and yellow become green; and blue, yellow, and red become blackish). As we decrease the truncation (i.e., locality bias) parameter ϵ, the different communities first explore local structure before merging. In panel (d), observe that each community covers most of the network.



(a) $\epsilon = 10^{-3}$ (b) $\epsilon = 10^{-4}$ (c) $\epsilon = 10^{-5}$ (d) $\epsilon = 10^{-6}$

FIG. 13. (Color online) Visualization of global structure in FB-JOHNS55. We construct the network layout by weighting each edge using the corresponding entry of the association matrix for the ACLCUT method [see Fig. 11(b)] and then applying the same procedure as in Fig. 12. Colored nodes correspond to local communities for three different seed nodes. Note the difference in behavior for the red community versus the blue and yellow communities. The blue and yellow communities gradually spread as we decrease ϵ, and they eventually merge to cover a large part of the network. However, the red community initially spreads as we decrease ϵ, and it then remains roughly the same as we decrease ϵ further.

combine to form large communities, which would result in large diagonal blocks in the association matrices. Instead, the small communities appear to amalgamate into a single large block (or "core"). In Fig. 12, we aim to make this observation more intuitive by showing how the local communities for three different seed nodes spread through the network as we change the locality bias parameter ϵ. We construct the weighted network $\widetilde{G} = (V, E, \widetilde{w})$ shown in Fig. 12 from the unweighted CA-GRQC network $G = (V, E)$ using the association matrix for the ACLCUT method [see Fig. 11(a)]. We assign each edge $\{i, j\} \in E$ a weight based on the corresponding entry of the association matrix. That is, $\widetilde{w}_{ij} = \widetilde{A}_{ij}$ if $\{i, j\} \in E$ and $\widetilde{w}_{ij} = 0$ otherwise. Based on our earlier results with the slowly increasing NCP, as well as previous results in Refs. [24–26], we interpret these features in Fig. 12 in terms of a nested core-periphery structure, in which the network periphery consists of relatively good communities and the core consists of relatively densely connected nodes.

For FB-JOHNS55 (see Figs. 11 and 13), we observe two relatively large communities, which correspond to the two large diagonal blocks in Figs. 11(b) and 11(e) and which underlie the dips in the NCPs in Figs. 6(a) and 16(a). Note, however, from the scale of the vertical axis in Figs. 6(a) and 16(a) that the community quality of these communities is very low, so one should actually construe the visualization in Figs. 11(b) and 11(e) as highlighting a low-quality community that is only marginally better than the other low-quality communities that are present in FB-JOHNS55. Based on this visualization as well as our earlier results, the remainder of FB-JOHNS55 does not appear to have much community structure (at least based on using the conductance diagnostic to measure internal versus external connectivity). However, there do appear to be some remnants of highly overlapping communities that one could potentially identify using other methods (e.g., the one in Ref. [36]). The EGONET method [see Fig. 11(h)] is unable to resolve not only these small communities but also the larger low-quality communities. Figure 13 shows how the local communities for two seed nodes that do not belong to one of the two large communities slowly spread and eventually merge (blue and yellow nodes), whereas the red community (which

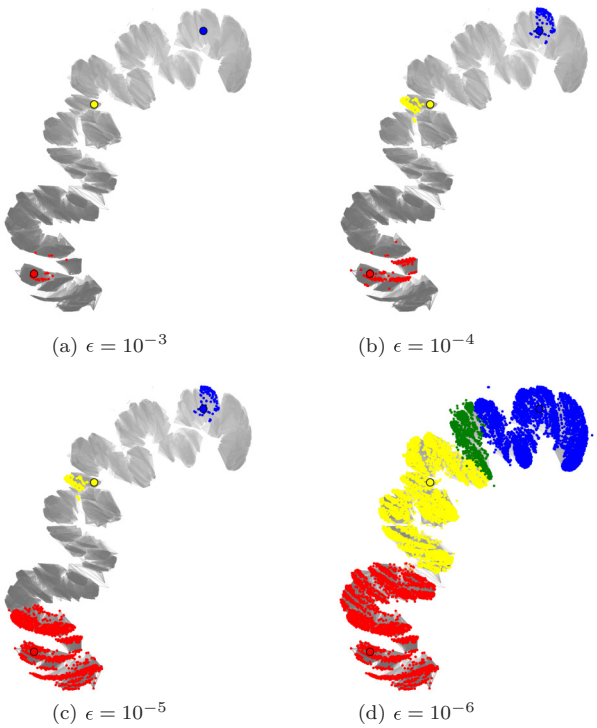


FIG. 14. (Color online) Visualization of global structure in US-SENATE. We construct the network layout by reweighting each edge using the corresponding entry of the association matrix for the ACLCUT method [see Fig. 11(c)] and then applying the same procedure as in Figs. 12 and 13. Colored nodes correspond to local communities for three different seed nodes. The spreading behavior of the different local communities largely follows the temporal structure of the network.

corresponds to the smaller of the two communities) is quickly identified and remains separate from the other communities.

For US-SENATE (see Figs. 11 and 14), we clearly observe the signature of temporal-based community structure at a large size scale. See Figs. 11(c), 11(f), and 11(i). Using ACLCUT and MOVCUT, we also obtain partitions at the scale of individual Congresses [see the insets in Figs. 11(c) and 11(f)], which sometimes split into two or occasionally three individual communities. These latter partitions have been discussed previously in terms of polarization between parties [9,90,98]. Because we fix the temporal order of Congresses for US-SENATE and only sort senators within the same Congress, this visualization reveals communities within each Senate as well as more temporally disparate communities. Figure 14 clearly shows that the temporal structure dominates the behavior of local communities for individual seed nodes.

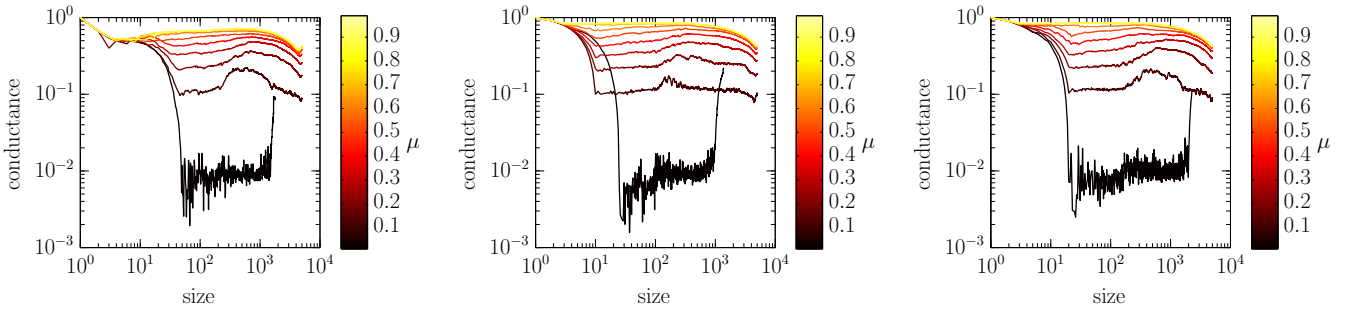
An important point from these visualizations is that, for both CA-GRQC and FB-JOHNS55, the meso-scale and large-scale structures that result from the superposition of local communities do *not* correspond particularly well to intuitive good-conductance communities. Relatedly, they also do *not* correspond particularly well to an intuitive low-dimensional structure or a nearly decomposable block-diagonal matrix of community assignments [see our illustration in Fig. 1(a)], one or both of which are often assumed (typically implicitly) by many global methods for algorithmically detecting communities in networks [5,6,37,99,100]. Of the networks that we investigate, only the temporal structure in US-SENATE (as well as in US-HOUSE, which is a related temporally dominant network) closely resembles such an idealization. This is reflected clearly in its downward-sloping NCP [see Figs. 6(a), 16(a), and 18(a)] and in the visualizations in Fig. 11.

Instead, in other (e.g., collaboration, Facebook, and many other real [24,25]) networks, community structure as a function of size is much more subtle and complicated. Our locally biased perspective provides one means to try to resolve such intricacy. By averaging over results from different seed nodes, a local approach like ours leads naturally to the presence of strongly overlapping communities. Overlapping community structure has now been studied for several years [101–103], and recent observations continue to shed new light on the ubiquity of community overlap [36]. Overlap of communities in networks is a pervasive phenomenon [36,104]; and our expectation is that most large realistic networks have communities with significant overlap, rather than merely a small amount of overlap that would amount to a small perturbation of the idealized, nearly decomposable communities in Fig. 1(a). Additionally, such overlaps imply that larger communities tend to have lower quality in terms of their internal versus external connectivity (i.e., in terms of how much they resemble the intuitive communities that many researchers know and love) than smaller communities—in agreement with our empirical results on both the collaboration networks and Facebook networks, but in strong disagreement with popular intuition. Recent work fit several real networks with upward-sloping NCPs to hierarchical Kronecker graphs and obtained parameters that are consistent with the core-periphery structure that we illustrated in Fig. 1(b) [105].

V. EMPIRICAL RESULTS ON SYNTHETIC BENCHMARKS

Synthetic benchmark networks with a known, planted community structure can be helpful for validating and gaining a better understanding of the behavior of community-detection algorithms. For such an approach to be optimally useful, it is desirable for the synthetic benchmarks to reproduce relevant features of real networks with community structure; and it is challenging to develop good benchmarks that reproduce community structure and other structural properties of medium-sized and larger realistic networks. An extremely popular and useful family of benchmark networks that aims to reproduce some features of real networks are the so-called *LFR networks* [106]. By design, LFR networks have power-law degree distributions as well as power-law community-size distributions. Additionally, they are unweighted and they have nonoverlapping planted communities. One can use a similar method to construct weighted networks and networks with overlapping planted communities [108]. Motivated by our empirical results on networks constructed from real data, we also apply our methods to LFR networks to test the extent to which they are able to reproduce the three classes of NCP behavior (upward-sloping, flat, and downward-sloping) that we have observed with real networks.

To parametrize the family of LFR networks, we specify a network's power-law degree distribution using its exponent τ_1, mean degree $\langle k \rangle$, and maximum degree $k_{\max}$. Similarly, we



(a) AclCut, $\langle k\rangle = 10$, $k_{\max} = 100$, $\tau_1 = -2$, $\tau_2 = -3$, $c_{\min} = 10$, $c_{\max} = 50$

(b) AclCut, $\langle k\rangle = 20$, $k_{\max} = 50$, $\tau_1 = -2$, $\tau_2 = -1$, $c_{\min} = 10$, $c_{\max} = 50$

(c) AclCut, $\langle k\rangle = 20$, $k_{\max} = 50$, $\tau_1 = -2$, $\tau_2 = -1$, $c_{\min} = 20$, $c_{\max} = 100$

FIG. 15. (Color online) NCPs of LFR synthetic benchmark networks [106] with $n = 10\,000$ nodes. Colors correspond to different values of the mixing parameter μ. Our choices for the mean degree $\langle k\rangle$, maximum degree $k_{\max}$, exponent τ_1 of the degree distribution, exponent τ_2 of the community-size distribution, minimum community size $c_{\min}$, and maximum community size $c_{\max}$ correspond to the ones used in Refs. [28,107] to benchmark community-detection algorithms.

specify its power-law community-size distribution using its exponent τ_2, minimum community size $c_{\min}$, and maximum community size $c_{\max}$, with the additional constraint that the sum of community sizes should equal the size n of the network. Furthermore, we specify the strength of community memberships using a mixing parameter μ, where each node shares a fraction $1-\mu$ of its edges with nodes in its own community. A simple calculation shows that this definition of the mixing parameter implies that each community in the planted partition has conductance μ (up to rounding effects).

To construct a network with these parameters, we sample n degrees from the degree distribution and sample community sizes from the community-size distribution. We then assign nodes to communities uniformly at random, with the constraint that a node cannot be assigned to a community that is too small for the node to have the correct mixing-parameter value. We then construct intercommunity and intracommunity edges separately by connecting the corresponding stubs (i.e., ends of edges) uniformly at random. We use the implementation by Lancichinetti [109] to generate LFR networks.

In Fig. 15, we show representative NCPs for LFR networks for three choices of parameters for the degree distribution and community-size distribution that have been used previously to benchmark community-detection algorithms [28,106,107]. (We generated the results presented in Fig. 15 using the AclCut method, but we obtain nearly identical NCPs using the MovCut method.) The three subfigures demonstrate that all three parameter choices yield networks with similar NCPs. In particular, we observe that—above a certain critical size—the best communities have comparable quality as a function of size. Depending on the particular parameter values, this can be of similar quality to or somewhat better than that which would be obtained by, e.g., a vanilla (not extremely sparse) Erdős-Rényi (ER) random graph at large size scales. That is, above the critical size, the NCP is approximately flat. Increasing the topological mixing parameter μ in the LFR network generative mechanism at first shifts the entire NCP upwards because the number of intercommunity edges increases. For $\mu \approx 1$, it levels off to the characteristic flat shape for an NCP of a network generated from the configuration model of random graphs.

Importantly, the behavior for the LFR benchmark networks from Ref. [106] that we illustrate in Fig. 15 does *not* resemble the NCPs for any of the real-world networks in either the present paper or in Ref. [24,25]. In addition, we have been unable to find parameter values for which the qualitative properties of realistic NCPs—in particular, a relatively gradually upward-sloping NCP—are reproduced, which suggests that the community structure generated by the LFR benchmarks is *not* realistic in terms of its size-resolved properties.

To verify that this behavior is not an artifact of the particular choices of parameters in Fig. 15, we sample sets of parameters uniformly at random with $n \in \{1000, 10\,000, 50\,000\}$, $\tau_1, \tau_2 \in \{-1, -2, \ldots, -5\}$, $\langle k\rangle \in \{10, 11, \ldots, 100\}$, $k_{\max} \in \{\langle k\rangle, \langle k\rangle + 1, \ldots, 250\}$, $c_{\min} \in \{10, 11, \ldots, 250\}$, and $c_{\max} \in \{\max(c_{\min}, k_{\max}), \ldots, 250\}$ and compute the NCPs of the resulting LFR benchmark networks. The aggregate trends of these NCPs are similar to the results that we show in Fig. 15. Hence, although the LFR benchmark networks are useful as tests for community-detection techniques, our calculations suggest that they are unable to reproduce a fundamental feature of many real networks with respect to variation in community quality (and, in particular, worsening community quality) as a function of increasing community size.

Based on our empirical observations, our locally biased perspective on community detection suggests a natural approach to determine whether synthetic benchmarks possess small-scale, medium-scale, and large-scale community structure that resembles that of large realistic networks: Namely, a family of synthetic benchmark networks ought to include parameter values that generate networks with (robust) upward-sloping, flat, and downward-sloping NCPs [as observed in Figs. 2(a) and 6(a)].

VI. CONCLUSIONS AND DISCUSSION

In this paper, we conducted a thorough investigation of community quality as a function of community size in a suite of realistic networks, and we reached several conclusions with important implications for the investigation of realistic medium-sized to large networks. Our results build on previous work on using NCPs to study large networks [24–26]. In this

paper, we employed a wider class of community-identification procedures, and we discovered a wider class of communitylike behaviors (as a function of community size), in realistic networks than what had been reported previously in the literature [110]. In addition, we discovered using NCPs that the popular LFR synthetic benchmark networks, which are often used to validate community-detection algorithms—and which are advocated as realistic synthetic benchmark networks for testing methods for community detection [111]—exhibit behavior that is markedly different from many real networks. Our result thus underscores the importance of developing realistic benchmark graphs whose NCPs are qualitatively similar to those of real networks. Taken together, our empirical results yield a much better understanding of realistic community structure in large realistic networks than was previously available, and they provide promising directions for future work. More generally, because our approach for comparing community structures in networks (using NCPs and CRPs) is very general—e.g., one can follow an analogous procedure with other community-quality diagnostics, other procedures for community generation, etc.—our locally biased and size-resolved methodology is an effective way to investigate meso-scale network structures much more generally.

The main conclusion of our work is that community structure in real networks is much more intricate than is suggested by the block-diagonal assumption that is (either implicitly or explicitly) made by most community-detection methods (including ones that allow overlapping communities [101]) and when using the synthetic benchmark networks that have been developed to test those methods. Community structure interplays with other meso-scale features, such as core-periphery structure [36,62,64], and investigating only community structure without consideration of other structures can lead to misleading results. A local perspective on community detection, like the one that we have advocated in the present paper, allows pervasive community overlap in a natural way. This is an important feature to capture when considering real social networks. Additionally, the large-scale consensus community structure that one obtains subsequently by "pasting together" local communities is not constrained to resemble a global block-diagonal structure. This is a key consideration in the study of meso-scale structures in real networks.

Although most algorithmic methods for community detection take a different approach from ours, the observation that network community structure depends not only on network structure per se but also on the dynamical processes that take place on a network and the initial conditions (i.e., seed node or nodes) for those processes is rather traditional in many ways. Recall, for example, Granovetter's observation that a node with many weak ties is ideally suited to initialize a successful social contagion process [112]. Our perspective also meshes better than global ones with real-life experience in our own networks. Both of these observations underscore our point that whether particular network structures form bottlenecks for a dynamical process depends not only on the process itself but also on the initial conditions of that process.

More generally, one might hope that our size-resolved and locally biased perspective on community detection can be used to help develop new diagnostics that complement widely used and intuitive concepts such as closeness centrality, betweenness centrality, and many other existing global notions. These will be of particular interest for investigating large networks—or even medium-sized networks such as those that we have considered—where traditional algorithmic and visualization methods have serious difficulties. Because the study of meso-scale structure in networks is important for understanding how local and small-scale properties of a network interact with global and large-scale properties, we expect that taking a locally biased perspective on community detection and related problems will yield numerous interesting and novel insights.

ACKNOWLEDGMENTS

L.G.S.J. acknowledges support through a CASE studentship award from the EPSRC (Award No. BK/10/039). M.A.P. was supported by a research award (Award No. 220020177) from the James S. McDonnell Foundation, the EPSRC (Award No. EP/J001759/1), and the FET-Proactive project PLEXMATH (Award No. FP7-ICT-2011-8; Grant No. 317614) funded by the European Commission; M.A.P. also thanks SAMSI for supporting several visits and M.W.M. for his hospitality during his sabbatical at Stanford. P.J.M. was funded by the NSF (Grant No. DMS-0645369) and by Award No. R21GM099493 from the National Institute of General Medical Sciences. M.W.M. acknowledges funding from the Army Research Office and from the Defense Advanced Research Projects Agency. In addition, we thank Adam D'Angelo and Facebook for providing the Facebook data, Keith Poole for providing the Congressional voting data (which is available from Ref. [130]), and Jure Leskovec for making many large network data sets publicly available as part of SNAP [83].

APPENDIX A: EXPANDER GRAPHS

In this section, we provide a brief introduction to the concept of an *expander graph* (or, more simply, an *expander*) [113]. Essentially, expanders are graphs that are very well connected and thus do not have any good communities (when measured with respect to diagnostics such as conductance). Because our empirical results indicate that many large social and information networks are expanders—at least when viewed at large size scales—it is useful to review basic properties about expander graphs. Most of the technical aspects of expander graphs are beyond the scope of this paper, but Ref. [114] provides an excellent overview of this topic.

Let $G = (V, E)$ be a graph, which we assume for simplicity is undirected and unweighted. For the moment, we assume that all nodes have the same degree d (i.e., G is d-regular). For $S_1, S_2 \subset V$, the set of edges between S_1 and S_2 is

$$E(S_1, S_2) = \{\{i, j\} : i \in S_1, j \in S_2, \{i, j\} \in E\}. \tag{A1}$$

In this case, the number $|S|$ of nodes in S is a natural measure of the size of S. Additionally, the quantity $|E(S, \overline{S})|$, which indicates the number of edges that cross between S and $\overline{S}$, is a natural measure of the size of the boundary between S and $\overline{S}$.

We also define the *edge expansion of a set of nodes* $S \subset V$ as

$$h(S) = \frac{|E(S,\overline{S})|}{|S|}. \tag{A2}$$

The *edge expansion of a graph* G is the minimum edge expansion of any subset (of size no greater than $n/2$) of nodes:

$$h(G) = \min_{S \subset V : |S| \leqslant \frac{n}{2}} h(S). \tag{A3}$$

A sequence of d-regular graphs $\{G_t\}_{t\in\mathbb{N}}$ is a *family of expander graphs* if there exists an $\epsilon > 0$ such that $h(G_t) \geqslant \epsilon$ for all $t \in \mathbb{N}$. Informally, a given graph G is an expander if its edge expansion is large.

As reviewed in Ref. [114], one can view expanders from several complementary viewpoints. From a combinatorial perspective, expanders are graphs that are highly connected in the sense that one has to sever many edges to disconnect a large part of an expander graph. From a geometric perspective, this disconnection difficulty implies that every set of nodes has a very large boundary relative to its size. From a probabilistic perspective, expanders are graphs for which the natural random-walk process converges to its limiting distribution as rapidly as possible. Finally, from an algebraic perspective, expanders are graphs in which the first nontrivial eigenvalue of the Laplacian operator is bounded away from 0. (Because we are discussing d-regular graphs, note that this statement holds for both the combinatorial Laplacian and the normalized Laplacian.) In addition, constant-degree (i.e., d-regular, for some fixed value of d) expanders are the metric spaces that (in a very precise and strong sense [114]) embed least well in low-dimensional spaces (such as those discussed informally in Sec. II B). All of these interpretations imply that smaller values of expansion correspond more closely to the intuitive notion of better communities (whereas larger values of expansion correspond, by definition, to better expanders).

Note the similarities between Eq. (A2) and Eq. (A3), which define expansion, and Eq. (3) and Eq. (4), which define conductance. These equations make it clear that the difference between expansion and conductance simply amounts to a different notion of the size (or volume) of sets of nodes and the size of the boundary (or surface area) between a set of nodes and its complement. This difference is inconsequential for d-regular graphs. However, because of the deep connections between expansion and rapidly mixing random walks, the latter notion (i.e., conductance) is much more natural for graphs with a substantial degree of heterogeneity. The interpretation of failing to embed well in low-dimensional spaces (like lines or planes) is not extremal in the case of conductance and degree-heterogeneous graphs (as it is in the case of expansion and degree-homogeneous graphs). However, the interpretations of many other properties, such as being well-connected and failing to provide bottlenecks to random walks, also hold for conductance and degree-heterogeneous graphs (such as those that we consider in the main text of the present paper). Accordingly, it is insightful to interpret our empirical results on small-scale versus large-scale structures in networks in the light of known facts about expander (and expanderlike) graphs.

APPENDIX B: COMMUNITY QUALITY, DYNAMICS ON GRAPHS, AND BOTTLENECKS TO DYNAMICS

In this section, we describe in more detail how we algorithmically identify possible communities in graphs. Because we are interested in local properties and how they relate to meso-scale and global properties, we take an operational approach and view communities as the output of various dynamical processes (e.g., diffusions or geodesic hops), and we discuss the relationship between the output of such procedures and well-defined optimization problems. The idea of using dynamics on a network has been exploited successfully by many methods for finding "traditional" communities (of densely connected nodes) [9,32,53,115–118] as well as for finding sets of nodes that are related to each other in other ways [48,54,115,119,120].

In this paper, we build on the idea that random walks and related diffusion-based dynamics, as well as other types of local dynamics (e.g., ones, like geodesic hops, that depend on ideas based on egocentric networks), should get "trapped" in good communities. We examine three dynamical methods for community identification.

1. Dynamics type 1: Local diffusions (the "ACLCUT" method)

In this procedure, we consider a random walk that starts at a given seed node s and runs for some small number of steps. We take advantage of the idea that if a random walk starts inside a good community and takes only a small number of steps, then it should become trapped inside that community. To do this, we use the locally biased PPR procedure of Refs. [121,122]. Recall that a PPR vector is defined implicitly as the solution $\vec{\mathrm{pr}}(\alpha,\vec{s})$ of the equation

$$\vec{\mathrm{pr}}(\alpha,\vec{s}) = \alpha D^{-1} A\, \vec{\mathrm{pr}}(\alpha,\vec{s}) + (1-\alpha)\vec{s}, \tag{B1}$$

where $1-\alpha$ is a "teleportation" probability and $\vec{s}$ is a seed vector. From the perspective of random walks, evolution occurs either by the walker moving to a neighbor of the current node or by the walker "teleporting" to a random node (e.g., determined uniformly at random as in the usual PageRank procedure, or to a random node that is biased towards $\vec{s}$ in the PPR procedure). The PPR vector $\vec{\mathrm{pr}}(\alpha,\vec{s})$ represents the stationary distribution of this random walk. In general, teleportation results in a bias to the random walk, and one usually tries to minimize such a bias when detecting communities. (See Ref. [123] for clever ways to choose $\vec{s}$ with this goal in mind.)

The algorithm of Refs. [121,122] deliberately exploits the bias from teleportation to achieve localized results. It computes an approximation to the solution of Eq. (B1) (i.e., it computes an *approximate PPR vector*) by strategically "pushing" mass between the iteratively updated approximate solution vector and a residual vector in such a way that most of the nodes in the original network are *not* reached. Consequently, this algorithm is typically *much* faster for moderately large to very large graphs than is the naive algorithm to compute a solution to Eq. (B1). The algorithm is parametrized in terms of a "truncation" parameter ϵ, where larger values of ϵ correspond to more locally biased solutions. We refer to this procedure as the ACLCUT method.

2. Dynamics type 2: Local spectral partitioning (the "MovCut" method)

In this procedure, we formalize the idea of a locally biased version of the leading nontrivial eigenvector of the normalized Laplacian $\mathcal{L}$. One can use such an eigenvector for a locally biased version of traditional spectral graph partitioning.

Following Ref. [37], consider the optimization problem

$$\begin{aligned} \underset{\vec{x}}{\text{minimize}} \quad & \vec{x}^T \mathcal{L} \vec{x} \\ \text{subject to} \quad & \vec{x}^T \vec{x} = 1, \\ & \vec{x}^T D^{1/2} \vec{1} = 0, \\ & (\vec{x}^T D^{1/2} \vec{s})^2 \geqslant \kappa, \end{aligned} \tag{B2}$$

where κ is a locality parameter and $\vec{s}$ is a vector, which satisfies the constraints $\vec{s}^T D \vec{s} = 1$ and $\vec{s}^T D \vec{1} = 0$ and which represents a seed set of nodes. That is, in the norm associated with the diagonal matrix D, the seed vector $\vec{s}$ has unit length and is exactly orthogonal to the 1-vector $\vec{1}$. This *locally biased* version of standard spectral graph partitioning [which becomes the usual global spectral-partitioning problem if the locality constraint $(\vec{x}^T D^{1/2} \vec{s})^2 \geqslant \kappa$ is removed] was introduced in [37], where it was shown that the solution vector $\vec{x}^*$ inherits many of the nice properties of the solution to the usual global spectral-partitioning problem. The solution $\vec{x}^*$ is of the form

$$\vec{x}^* = a(L_G - \gamma D_G)^+ D_G \vec{s}, \tag{B3}$$

where the parameter $\gamma \in (-\infty, \lambda_2(G))$ is related to the teleportation parameter α via the relation $\gamma = \frac{\alpha - 1}{\alpha}$ (see [37]) and $a \in [0, \infty]$ is a normalization constant.

As one can see from Eq. (B3), the solution $\vec{x}^*$ of Eq. (B2) is an *exact PPR vector* with personalized teleportation vector $\vec{s}$. Consequently, it can be computed as the solution to a system of linear equations. In addition, if one performs a sweep cut (see the discussion below) of this solution vector to obtain a locally biased network partition, then one obtains Cheeger-like guarantees on approximation quality for the associated network community. Moreover, if the seed vector $\vec{s}$ corresponds to the indicator vector of a single node i, then this is a relaxation of the following *locally biased graph-partitioning problem*: Given as input a graph $G = (V, E, w)$, a node i, and a positive integer k, find a set of nodes $S \subseteq V$ that achieves the minimum conductance among all sets of nodes that contain the input node i and have volume no larger than k [37]. We refer to this procedure (with a seed vector corresponding to a single seed node) as the MovCut method.

3. Dynamics type 3: Local geodesic spreading (the "EgoNet" method)

In this procedure, we perform a geodesic-based (i.e., ego-network-based) dynamics that is analogous to the local random walks that we described above. This method is similar to the technique for finding local communities that was introduced in Ref. [38] and which was generalized to weighted networks in Ref. [124]. Starting with a seed node s and a distance parameter k, this method construes all nodes i whose geodesic distance from s is at most k—i.e., all nodes i such that $\Delta_{si} \leqslant k$—as a local community. The local communities that we obtain using this method are simply the k-ego-nets of the seed node. For consistency with the other two methods, it is useful to think of this method as inducing a ranking of the nodes,

$$\text{EgoRank}_i(s) = \frac{1}{1 + \Delta_{is}}, \tag{B4}$$

where i is some node in a network. Given the ranking interpretation in Eq. (B4), we recover local geodesic-based communities from the EgoRank vector by using the sweep cut procedure that we describe below. The underlying dynamics for this method is analogous to the extreme case of a susceptible-infected (SI) spreading process [43,126], in which an infected node infects all of its neighbors with probability 1 at the time step following the one in which it is infected. One can then interpret the EgoRank of node i for a seed node s as the inverse of the time that it takes for node i to first become infected when only the seed node s is infected initially. We refer to this procedure as the EgoNet method.

4. Sampling procedures and parameter choices

To obtain an accurate picture of local community structure at different size scales throughout a network, we run each of the above community-identification procedures many times, starting at different seed nodes and running for different numbers of steps, and we then examine which nodes get visited as the dynamical processes unfold. For each seed node and value of the parameters, each of the AclCut, MovCut, and EgoNet methods returns a vector that can be used to "rank" the nodes of a network (in a locally biased and size-resolved manner): AclCut and MovCut return a variant of the PPR vector, and EgoNet returns the EgoRank vector in Eq. (B4). Given a ranking vector $\vec{p}$, the so-called "sweep sets" are given by $S_t = \{i \in V : \vec{p}_i \geqslant t\}$. Thus, there are at most $n + 1$ distinct sweep sets (where we recall that n is the number of nodes in a graph). A corresponding "sweep cut" is the network partition obtained from a sweep set that has minimal conductance among all of the $n + 1$ possible sweep sets. By computing the conductance for each of the sweep sets, one obtains a locally biased estimate (which is centered around a seed node) for an NCP. One can then estimate a global NCP by taking the lower envelope over local NCPs for different seed nodes and parameter values. Our MATLAB code that implements these methods is available at [127].

To obtain a good estimate of a global NCP, we need to sample good communities for all possible size scales of a network. In the next several paragraphs, we discuss how we choose the parameter values and how to sample seed nodes in our three community-detection methods. Recall that AclCut has two parameters (the teleportation parameter α and the truncation parameter ϵ), but that MovCut only has a single parameter (a teleportation parameter).

For AclCut, theoretical results [121] suggest that the method should find good communities of volume roughly ϵ^{-1}, where we have ignored constants and logarithmic factors. Furthermore, for a seed node i with strength k_i, AclCut returns empty communities for $\epsilon < k_i^{-1}$. This suggests that sampling using $\epsilon \in [k_{\max}^{-1}, \text{vol}(G)^{-1}]$ gives good coverage of different size scales in practice. In this paper, we use 20 logarithmically spaced points in $[k_{\max}^{-1}, \text{vol}(G)^{-1}]$ (including the end points) to generate Figs. 6, 7, and 15. In addition,

we use $\tilde{\alpha} = 0.001$, where $\tilde{\alpha}$ is the teleportation parameter of the "lazy random walk" defined in [121]. The (conventional) teleportation parameter that we use satisfies $\alpha = 1 - \frac{2\tilde{\alpha}}{1+\tilde{\alpha}}$, so $\alpha \approx 0.998$ in Eq. (B1). In our computations, we observe that increasing α leads to more accurate NCPs at the cost of longer computation times.

For MOVCUT, we use 20 equally spaced values of α in the interval $[0.7,(1-\lambda_2)^{-1} - 10^{-10}]$ (including the end points), where $(1-\lambda_2)^{-1}$ is the theoretical maximum for α (see [37]). To sample seed nodes, we modify the strategy described in Ref. [25] to be applicable to the MOVCUT method as well as the ACLCUT method. For each choice of parameter values, we sample nodes uniformly at random without replacement and stop the sampling process either when all nodes are sampled or when the sampled local communities sufficiently cover the entire network. To determine sufficient coverage, we track how many times each node is included in the best local community that we obtain from the sweep sets and stop the procedure once each node is included at least ten times. This procedure ensures that good communities are sampled consistently.

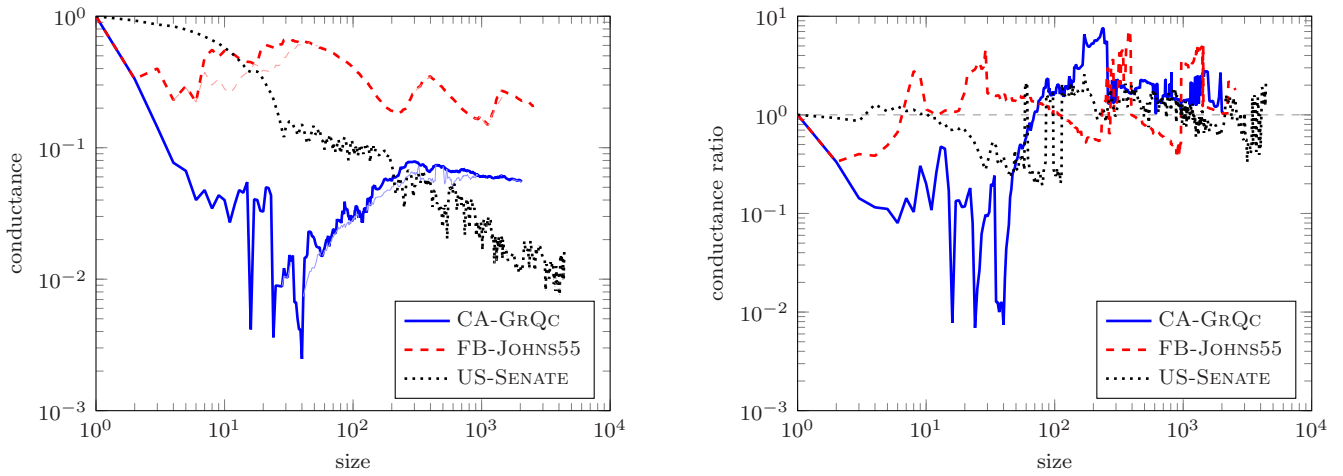
The EGONET method does not have any size-scale parameters. For the network sizes that we consider, it is feasible to use all nodes rather than sampling them. We use this approach for our calculations using this method.

Finally, for readability, we only plot the NCPs for communities that contain at most half of the nodes in a network. The symmetry in the definition of conductance [see Eq. (3)] implies that the complement of a good small community is necessarily a good large community and vice versa. Hence, a sampled NCP is roughly symmetric, though this is hard to see on a logarithmic scale, and an NCP without sampling is necessarily symmetric.

APPENDIX C: DETAILED RESULTS FOR THE MOVCUT METHOD

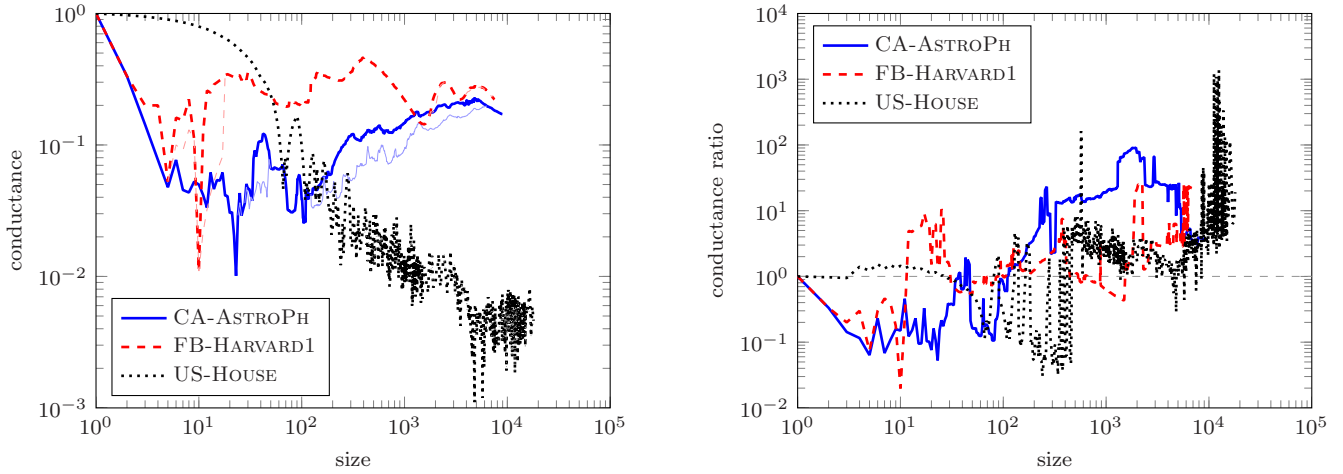
The MOVCUT method provides an alternative to ACLCUT for sampling local community profiles to construct an NCP. Unlike ACLCUT, which uses *only* local information to obtain good communities, MOVCUT also incorporates some global information about a network to construct local communities around a seed node. In particular, this implies that there can be sweep sets and thus communities that consist of disconnected components of a network. Such communities have infinitely large conductance ratios. We observe this phenomenon often for the coauthorship and Facebook networks, but it almost never occurs for the Congressional voting networks. These sweep sets consist of several small sets of peripheral nodes, each of which has moderate to very low conductance, but which are otherwise unrelated. Although one would not usually think of such a set of nodes as a single good community, optimization-based algorithms often clump several unrelated communities into a single community for networks with a global core-periphery structure. For completeness and comparison, we include our results both when we keep the disconnected sweep sets and when we restrict our attention to connected communities. As we discuss below, the NCP does not change substantially, although there are some small differences.

The resulting NCPs for the MOVCUT method [see Figs. 16(a) and 17(a)] are similar to those that we obtain for the ACLCUT method [see Figs. 6(a) and 7(a)], although there are a few differences worth discussing. The CRP plots are also very similar [compare Figs. 16(b) and 17(b) to Figs. 6(b) and 7(b)]. For the coauthorship networks (CA-GRQC and CA-ASTROPH), as well as FB-HARVARD1, both MOVCUT and ACLCUT identify the same good small communities that are responsible for the spikes in the NCP plots. In addition, the communities that yield the dips in the NCPs for FB-JOHNS55 near 220 and 1100 nodes and for FB-HARVARD1 near 1500 nodes all share more than 98% of their nodes. This indicates that both methods are able to find roughly the same communitylike structures. However, the result from the MOVCUT NCP for CA-GRQC is higher and less choppy than the one that we computed using ACLCUT because the



(a) NCP (b) CRP

FIG. 16. (Color online) NCP plots [in panel (a)] and CRP plots [in panel (b)] for FB-JOHNS55, CA-GRQC, and US-SENATE (i.e., the smaller network from each of the three pairs of networks from Table I) generated using the MOVCUT method. The thin curves are the NCPs that we obtain when we also consider disconnected sweep sets.



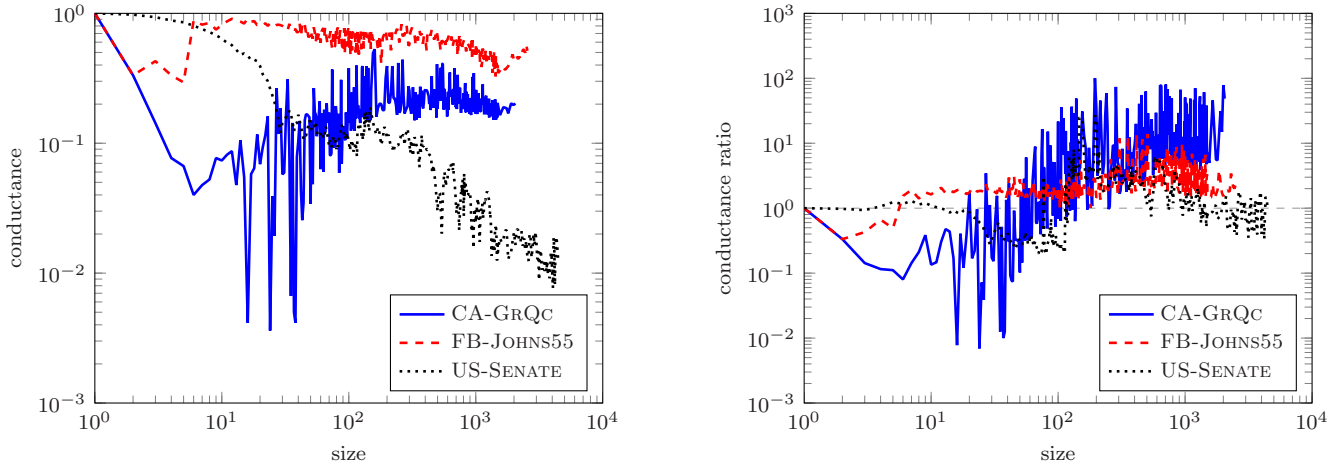
(a) NCP (b) CRP

FIG. 17. (Color online) NCP plots [in panel (a)] and CRP plots [in panel (b)] for CA-AstroPh, FB-Harvard1, and US-House (i.e., the larger network from each of the three pairs of networks from Table I) generated using the MovCut method. The thin curves are the NCPs that we obtain when we also consider disconnected sweep sets.

truncation employed by AclCut performs a form of implicit sparsity-based regularization that is absent from MovCut. See Refs. [99,100,128] for a discussion and precise characterization of this regularization. For the coauthorship and Facebook networks, we also note that there are regions of the computed NCPs, when using the MovCut method, in which one finds disconnected sweep sets (see the thin curves) with lower conductance than those for the best connected sets of the same size. At other sizes, we see some differences between the NCPs from MovCut and AclCut. This illustrates that the two methods can have somewhat different local behavior, although both methods produce similar insights about the large-scale structure in these networks.

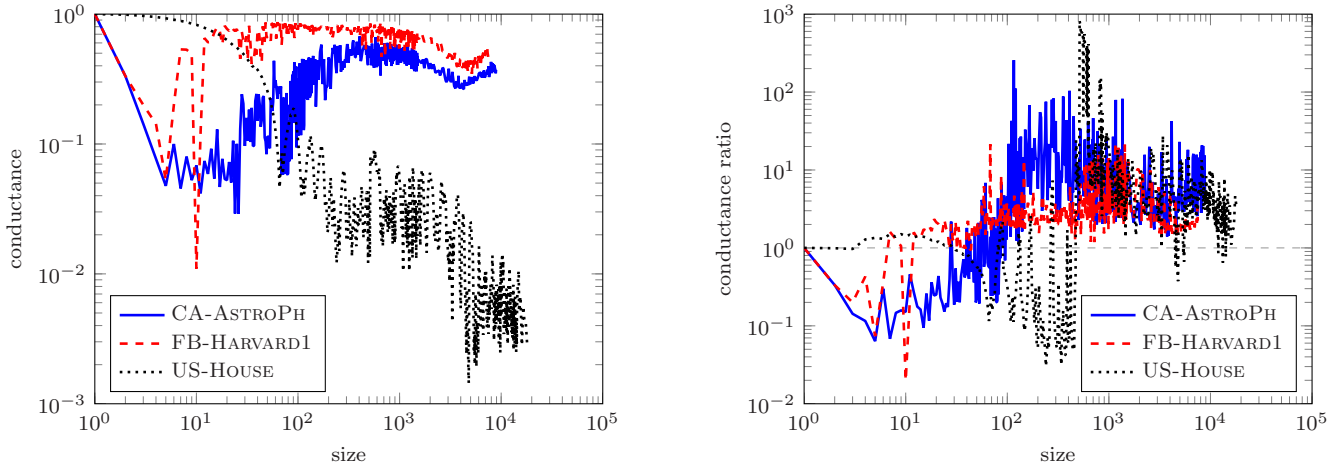
APPENDIX D: DETAILED RESULTS FOR THE EgoNet METHOD

The EgoNet method was not originally developed to optimize conductance, although there is some recent evidence that k-neighborhoods can be good conductance communities [129]. The assumption that underlies the EgoNet method is that nodes in the same community should be connected by short paths. However, unlike the spectral-based methods (AclCut and MovCut), the EgoNet method does not take into account the number of paths between nodes. In contrast to Ref. [129], which considered only 1-neighborhoods, we also examine k-neighborhoods with $k > 1$. We can then use the EgoNet method to approximate an NCP for a network.



(a) NCP (b) CRP

FIG. 18. (Color online) NCP plots [in panel (a)] and CRP plots [in panel (b)] for CA-GrQc, FB-Johns55, and US-Senate (i.e., the smaller network from each of the three pairs of networks from Table I) using the EgoNet method. We find qualitatively similar behavior as with the other two methods, although the NCPs are shifted upwards and some of the large-scale structure is no longer present (especially in the Facebook network).



(a) NCP (b) CRP

FIG. 19. (Color online) NCP plots [in panel (a)] and CRP plots [in panel (b)] for CA-AstroPh, FB-Harvard1, and US-House (i.e., the larger networks from each of the three pairs of networks from Table I) using the EgoNet method. We find qualitatively similar behavior as with the other two methods, although the NCPs are shifted upwards and some of the large-scale structure is no longer present (especially in the Facebook network).

Despite its simplicity, and in agreement with Ref. [129], the EgoNet method produces NCPs that are qualitatively similar to those from both the AclCut and MovCut methods for all of the networks that we consider (see Figs. 18 and 19). The NCPs for the EgoNet method are shifted upwards compared to those for the AclCut and MovCut methods, and this is particularly noticeable at larger community sizes. This is unsurprising, because the latter two methods more aggressively optimize the conductance objective. However, for all six of our networks, EgoNet conveys an NCP's small-scale structure as well as the global tendency to be upward-sloping, flat, or downward-sloping. This provides further evidence that the qualitative features of an NCP provide a signature of community structure in a network and are not just an artifact of a particular way to sample communities.

[1] M. O. Jackson, *Social and Economic Networks* (Princeton University Press, Princeton, NJ, 2008).

[2] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, Cambridge, UK, 2010).

[3] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, UK, 2010).

[4] A.-L. Barabási, Nat. Phys. **1**, 68 (2005).

[5] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Not. Am. Math. Soc. **56**, 1082 (2009).

[6] S. Fortunato, Phys. Rep. **486**, 75 (2010).

[7] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).

[8] M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand, Proc. Natl. Acad. Sci. USA **102**, 7057 (2005).

[9] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, Science **328**, 876 (2010).

[10] K. T. Macon, P. J. Mucha, and M. A. Porter, Physica A **391**, 343 (2012).

[11] A. L. Traud, P. J. Mucha, and M. A. Porter, Physica A **391**, 4165 (2012).

[12] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, SIAM Rev. **53**, 526 (2011).

[13] M. C. González, H. J. Herrmann, J. Kertész, and T. Vicsek, Physica A **379**, 307 (2007).

[14] A. C. F. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, BMC Syst. Biol. **4**, 100 (2010).

[15] D. S. Bassett, E. T. Owens, K. E. Daniels, and M. A. Porter, Phys. Rev. E **86**, 041306 (2012).

[16] P. Ronhovde, S. Chakrabarty, D. Hu, M. Sahu, K. K. Sahu, K. F. Kelton, N. A. Mauro, and Z. Nussinov, Eur. Phys. J. E **34**, 105 (2011).

[17] D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton, Proc. Natl. Acad. Sci. USA **108**, 7641 (2011).

[18] N. F. Wymbs, D. S. Bassett, P. J. Mucha, M. A. Porter, and S. T. Grafton, Neuron **74**, 936 (2012).

[19] D. S. Bassett, N. F. Wymbs, M. P. Rombach, M. A. Porter, P. J. Mucha, and S. T. Grafton, PLoS Comput. Biol. **9**, e1003171 (2013).

[20] T. S. Evans, R. Lambiotte, and P. Panzarasa, Scientometrics **89**, 381 (2011).

[21] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Proc. Natl. Acad. Sci. USA **104**, 7332 (2007).

[22] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Proc. Natl. Acad. Sci. USA **108**, 7663 (2011).

[23] D. Grady, R. Brune, C. Thiemann, F. Theis, and D. Brockmann, in *Handbook of Optimization in Complex Networks*, Springer Optimization and its Applications, edited by M. T. Thai and P. M. Pardalos (Springer, New York, 2012), pp. 169–208.

[24] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, in *Proceedings of the 17th International Conference on World Wide Web, WWW '08* (ACM, New York, 2008), pp. 695–704.

[25] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Math. **6**, 29 (2009).

[26] J. Leskovec, K. J. Lang, and M. W. Mahoney, in *Proceedings of the 19th International Conference on World Wide Web, WWW '10* (ACM, New York, 2010), pp. 631–640.

[27] We do not wish to tie ourselves too closely to particular numbers, because such numbers can vary dramatically in different classes of networks. As a rule of thumb, however, we think of "small," "medium-sized," and "large" communities as those that contain, respectively, tens to about a hundred, hundreds to thousands, and thousands or more nodes. Similarly, we think of "small" networks as having tens to hundreds of nodes and "large" networks as having thousands, tens of thousands, or even more nodes. Depending on one's perspective, one can construe the networks that we study in the present paper as either medium-sized or large.

[28] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 056117 (2009).

[29] B. H. Good, Y.-A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).

[30] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. Theor. Exp. (2008) P10008.

[31] J. Yang and J. Leskovec, in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, MDS '12* (ACM, New York, 2012), pp. 3:1–3:8.

[32] R. Lambiotte, J.-C. Delvenne, and M. Barahona, arXiv:0812.1770v3.

[33] M. T. Schaub, R. Lambiotte, and M. Barahona, Phys. Rev. E **86**, 026112 (2012).

[34] M. T. Schaub, J. C. Delvenne, S. N. Yaliraki, and M. Barahona, PLoS One **7**, e32210 (2012).

[35] S. Fortunato and M. Barthelemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).

[36] J. Yang and J. Leskovec, ACM Trans. Intell. Syst. Technol. **5**, 26 (2014).

[37] M. W. Mahoney, L. Orecchia, and N. K. Vishnoi, J. Mach. Learn. Res. **13**, 2339 (2012).

[38] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).

[39] N. A. Christakis and J. H. Fowler, N. Engl. J. Med. **357**, 370 (2007).

[40] N. A. Christakis and J. H. Fowler, Stat. Med. **32**, 556 (2013).

[41] F. Brauer and C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*, 2nd ed., Texts in Applied Mathematics Vol. 40 (Springer, New York, 2012).

[42] M. N. Kuperman, Pap. Phys. **5**, 050003 (2013).

[43] J. P. Gleeson, Phys. Rev. X **3**, 021004 (2013).

[44] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, arXiv:1111.4503.

[45] M. Gjoka, E. Smith, and C. Butts, *IEEE INFOCOM 2014–IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (IEEE, New York, NY, 2014), pp. 837–842.

[46] T. J. Hansen and M. W. Mahoney, in *Advances in Neural Information Processing Systems*, edited by P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (NIPS, La Jolla, CA, 2012), Vol. 25, pp. 2537–2545.

[47] M. Jerrum and A. Sinclair, in *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, STOC '88* (ACM, New York, 1988), pp. 235–244.

[48] K. Lerman and R. Ghosh, Phys. Rev. E **86**, 026108 (2012).

[49] R. Ghosh and K. Lerman, Discrete Contin. Dyn. Syst. Ser. B **19**, 1355 (2014).

[50] R. Ghosh, S.-H. Teng, K. Lerman, and X. Yan, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2014), pp. 1406–1415.

[51] J. P. Bagrow, J. Stat. Mech. Theor. Exp. (2008) P05001.

[52] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, PLoS One **6**, e18961 (2011).

[53] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008).

[54] R. Lambiotte, R. Sinatra, J.-C. Delvenne, T. S. Evans, M. Barahona, and V. Latora, Phys. Rev. E **84**, 017102 (2011).

[55] E. Ziv, M. Middendorf, and C. H. Wiggins, Phys. Rev. E **71**, 046117 (2005).

[56] C. Borgs, M. Brautbar, J. Chayes, S. Khanna, and B. Lucier, in *Internet and Network Economics*, Lecture Notes in Computer Science Vol. 7695 (Springer, Berlin, Heidelberg, 2012), pp. 406–419.

[57] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, Phys. Rev. E **83**, 036112 (2011).

[58] M. R. Bridson and A. Häfliger, *Metric Spaces of Non-positive Curvature* (Springer, Berlin, 1999).

[59] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).

[60] W. Chen, W. Fang, G. Hu, and M. W. Mahoney, Internet Math. **9**, 434 (2013).

[61] The parameter choices for the block models have subtle effects that are not explained by the relative sizes of the probabilities α_{11},α_{12}, and α_{22}. For a network with pronounced core-periphery structure [such as the one we show in Fig. 1(b)], where $\alpha_{11} \gg \alpha_{12} \gg \alpha_{22}$, one additionally needs $n_1\alpha_{11} \gg 1$ (for the core to be sufficiently connected), $n_1\alpha_{12} \ll 1$ (otherwise, most peripheral nodes are connected directly to the core), and $n_2\alpha_{22} \approx 1$ (such that the periphery is reasonably well connected but not too dense). Other parameter choices that respect the basic constraint on the probabilities still produce networks with well-defined density-based core-periphery structure [62,63,64], but they do not "look like" Fig. 1(b). (They tend to resemble Fig. 1(c), provided they are reasonably well connected.) For an expanderlike network [such as the one we show in Fig. 1(c)], where $\alpha_{11} \approx \alpha_{12} \approx \alpha_{22} \approx \alpha$, one needs $(n_1 + n_2)\alpha \gg 1$ for the network to be sufficiently well connected.

[62] P. Csermely, A. London, L.-Y. Wu, and B. Uzzi, J. Complex Networks **1**, 93 (2013).

[63] S. P. Borgatti and M. G. Everett, Soc. Networks **21**, 375 (2000).

[64] M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha, SIAM J. Appl. Math. **74**, 167 (2014).

[65] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).

[66] H. A. Simon, Proc. Am. Philos. Soc. **106**, 467 (1962).

[67] F. R. K. Chung, *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics Vol. 92 (AMS, Providence, RI, 1997).

[68] F. R. K. Chung and L. Lu, *Complex Graphs and Networks*, CBMS Regional Conference Series in Mathematics Vol. 107 (AMS, Providence, RI, 2006).

[69] S. Arora, S. Rao, and U. Vazirani, J. Assoc. Comput. Mach. **56**, 5 (2009).

[70] D. J. Watts, Am. J. Sociol. **105**, 493 (1999).
[71] A. Condon and R. M. Karp, Random Struct. Alg. **18**, 116 (2001).
[72] M. Mihail, in *Proceedings of the 30th Annual Symposium on Foundations of Computer Science* (IEEE, New York, 1989), pp. 526–531.
[73] Z. A. Zhu, S. Lattanzi, and V. Mirrokni, in *Proceedings of the 30th International Conference on Machine Learning*, edited by S. Dasgupta and D. McAllester (JMLR W&CP, 2013), pp. 396–404.
[74] P. J. Mucha and M. A. Porter, Chaos **20**, 041108 (2010).
[75] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, J. Complex Networks **2**, 203 (2014).
[76] D. S. Bassett, M. A. Porter, N. F. Wymbs, S. T. Grafton, J. M. Carlson, and P. J. Mucha, Chaos **23**, 013142 (2013).
[77] M. De Domenico, A. Solé-Ribalta, S. Gómez, and A. Arenas, Proc. Natl. Acad. Sci. USA **111**, 8351 (2014).
[78] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, Phys. Rev. X **3**, 041022 (2013).
[79] S. Gómez, A. Diaz-Guilera, J. Gómez-Gardenes, C. J. Pérez-Vicente, Y. Moreno, and A. Arenas, Phys. Rev. Lett. **110**, 028701 (2013).
[80] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, Phys. Rev. E **71**, 065103 (2005).
[81] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, Phys. Rev. E **75**, 027105 (2007).
[82] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[83] J. Leskovec, Stanford Network Analysis Project, available at http://snap.stanford.edu
[84] I. S. Jutla, L. G. S. Jeub, and P. J. Mucha, A Generalized Louvain Method for Community Detection Implemented in MATLAB (2011–2014), available at http://netwiki.amath.unc.edu/GenLouvain.
[85] Banded matrices arise, for example, in the study of finite-difference equations. In banded matrices, the nonzero entries are confined to diagonal bands near the main diagonal. Using our informal terminology from Sec. II B, this is an extreme example of a "hot dog."
[86] T. Kamada and S. Kawai, Inform. Process. Lett. **31**, 7 (1989).
[87] L. G. S. Jeub, Spring-based Visualization for Networks with Communities (2014), available at https://github.com/LJeub/SpringVisCom
[88] The two small "communities" in FB-HARVARD1 are responsible for the low value of λ_2 for this network compared with all of the other Facebook networks (recall Fig. 4). Removing these two communities from the network increases the value of λ_2 by an order of magnitude (from 0.0094 to 0.075), yielding a value that is fairly typical for networks in the FACEBOOK100 data set. In other words, aside from the 15 nodes from these two "communities," FB-HARVARD1 looks like FB-JOHNS55 and the rest of the FACEBOOK100 networks. This is an example of one of the few types of situations in which a feature of an NCP is *not* robust.
[89] For FB-JOHNS55, the community of size about 220 corresponds closely to a set of students with the same major, and the community associated with the dip near 1100 corresponds to first-year students. Similarly, the community associated with the dip near 1500 for FB-HARVARD1 also corresponds to first-year students. Consistent with the results of Ref. [11], we find similar poorly connected and moderately large communities that correlate reasonably well with first-year students in most of the other networks in the FACEBOOK100 data set. As usual, the most notable exception among the FACEBOOK100 networks is CALTECH36, which is known to be influenced much more by dormitory ("House") residence than by class year [11].
[90] M. Cucuringu and M. W. Mahoney, arXiv:1109.1355.
[91] C. Spearman, Am. J. Psychol. **15**, 72 (1904).
[92] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, in *Proceedings of the 3rd Annual ACM Web Science Conference, WebSci '12* (ACM, New York, 2012), pp. 33–42.
[93] This holds for all seed nodes, because the maximum voting similarity between a pair of senators is 1 (i.e., they voted the same way on every bill) and all interlayer edges (i.e., edges between the same senator in different Congresses) have a weight of 1. As one reduces the weight of interlayer edges, one needs to choose an increasingly large value of k for nodes from interlayer edges to appear in the k-neighborhood. Hence, as we reduce the weight of interlayer edges, an increasing number of senators from the same Congress can appear in a k-neighborhood that does not contain any nodes from interlayer edges.
[94] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and L. A. N. Amaral, Proc. Natl. Acad. Sci. USA **104**, 15224 (2007).
[95] J. Reichardt and S. Bornholdt, Phys. Rev. E **74**, 016110 (2006).
[96] A. Arenas, A. Fernández, and S. Gomez, New J. Phys. **10**, 053039 (2008).
[97] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, Bioinformatics **17**, S22 (2001).
[98] A. S. Waugh, L. Pei, J. H. Fowler, P. J. Mucha, and M. A. Porter, arXiv:0907.3509v3.
[99] M. W. Mahoney and L. Orecchia, in *Proceedings of the 28th International Conference on Machine Learning, ICML'11*, edited by L. Getoor and T. Scheffer (ACM, New York, 2011), pp. 121–128.
[100] D. Gleich and M. W. Mahoney, in *Proceedings of the 31st International Conference on Machine Learning*, edited by T. Jebara and E. P. Xing (JMLR W&CP, 2014), pp. 1018–1025.
[101] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, Nature (London) **466**, 761 (2010).
[102] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, Phys. Rev. E **83**, 066114 (2011).
[103] B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **84**, 036103 (2011).
[104] S. Lehmann, Intermittent Updates: Pervasive Overlap (2010), available at http://sunelehmann.com/2010/06/29/pervasive-overlap/
[105] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, J. Mach. Learn. Res. **11**, 985 (2010).
[106] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[107] A. Lancichinetti and S. Fortunato, Sci. Rep. **2**, 336 (2012).
[108] A. Lancichinetti and S. Fortunato, Phys. Rev. E **80**, 016118 (2009).
[109] A. Lancichinetti, Benchmarks (binary networks v.9) (2010), available at http://sites.google.com/site/andrealancichinetti/software

[110] To give one example, the coauthorship networks that we examine possess a familiar upward-sloping NCP and a nested core-periphery structure with small, well-separated communities in the periphery. However, the small and large communities that we observe in the Facebook networks do not lead to bottlenecks for diffusion processes. This NCP behavior in the Facebook graphs is consistent both with the larger edge density of the Facebook networks relative to the SNAP graphs [24–26] and with previous work on simple dynamical processes on these Facebook networks [57].

[111] G. K. Orman, V. Labatut, and H. Cherifi, Int. J. Web Based Communities **9**, 349 (2013).

[112] M. S. Granovetter, Am. J. Sociol. **78**, 1360 (1973).

[113] P. Sarnak, Not. Am. Math. Soc. **51**, 762 (2004).

[114] S. Hoory, N. Linial, and A. Wigderson, Bull. Am. Math. Soc. **43**, 439 (2006).

[115] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, Phys. Rev. Lett. **96**, 114102 (2006).

[116] R. Lambiotte, in *Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, WiOpt* (IEEE, New York, 2010), pp. 546–553.

[117] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, Proc. Natl. Acad. Sci. USA **107**, 12755 (2010).

[118] E. Estrada, Chaos **21**, 016103 (2011).

[119] D. Anderson, A. Tenzer, G. Barlev, M. Girvan, T. M. Antonsen, and E. Ott, Chaos **22**, 013102 (2012).

[120] M. Beguerisse-Díaz, G. Garduño-Hernández, B. Vangelov, S. N. Yaliraki, and M. Barahona, J. R. Soc. Interface **11**, 20140940 (2014).

[121] R. Andersen, F. Chung, and K. J. Lang, in *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, FOCS '06* (IEEE, New York, 2006), pp. 475–486.

[122] R. Andersen and K. J. Lang, in *Proceedings of the 15th International Conference on World Wide Web, WWW '06* (ACM, New York, 2006), pp. 223–232.

[123] R. Lambiotte and M. Rosvall, Phys. Rev. E **85**, 056107 (2012).

[124] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, Physica A **386**, 414 (2007).

[125] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, UK, 1994).

[126] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, Phys. Rev. E **83**, 025102 (2011).

[127] L. G. S. Jeub, LocalCommunities (2014), available at https://github.com/LJeub/LocalCommunities

[128] M. W. Mahoney, in *Proceedings of the 31st ACM Symposium on Principles of Database Systems, PODS '12* (ACM, New York, 2012), pp. 143–154.

[129] D. F. Gleich and C. Seshadhri, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12* (ACM, New York, 2012), pp. 597–605.

[130] K. T. Poole, Voteview (2014), available at http://voteview.com .