Internet Mathematics Vol. 6, No. 1: 29–123

Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters

Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney

Abstract. A large body of work has been devoted to defining and identifying clusters or communities in social and information networks, i.e., in graphs in which the nodes represent underlying social entities and the edges represent some sort of interaction between pairs of nodes. Most such research begins with the premise that a community or a cluster should be thought of as a set of nodes that has more and/or better connections between its members than to the remainder of the network. In this paper, we explore from a novel perspective several questions related to identifying meaningful communities in large social and information networks, and we come to several striking conclusions.

Rather than defining a procedure to extract sets of nodes from a graph and then attempting to interpret these sets as "real" communities, we employ approximation algorithms for the graph-partitioning problem to characterize as a function of size the statistical and structural properties of partitions of graphs that could plausibly be interpreted as communities. In particular, we define the *network community profile plot*, which characterizes the "best" possible community—according to the conductance measure—over a wide range of size scales. We study over one hundred large real-world networks, ranging from traditional and online social networks, to technological and information networks and web graphs, and ranging in size from thousands up to tens of millions of nodes.

Our results suggest a significantly more refined picture of community structure in large networks than has been appreciated previously. Our observations agree with previous work on small networks, but we show that large networks have a very different structure. In particular, we observe tight communities that are barely connected to the rest of the network at very small size scales (up to ≈ 100 nodes); and communities of size scale beyond ≈ 100 nodes gradually "blend into" the expander-like core of the network and thus become less "community-like," with a roughly inverse relationship between community size and optimal community quality. This observation agrees well

1542-7951/08 $0.50 per page

with the so-called Dunbar number, which gives a limit to the size of a well-functioning community.

However, this behavior is not explained, even at a qualitative level, by any of the commonly used network-generation models. Moreover, it is exactly the opposite of what one would expect based on intuition from expander graphs, low-dimensional or manifold-like graphs, and from small social networks that have served as test beds of community-detection algorithms. The relatively gradual increase of the network community profile plot as a function of increasing community size depends in a subtle manner on the way in which local clustering information is propagated from smaller to larger size scales in the network. We have found that a generative graph model, in which new edges are added via an iterative "forest fire" burning process, is able to produce graphs exhibiting a network community profile plot similar to what we observe in our network data sets.

1. Introduction

A large amount of research has been devoted to the task of defining and identifying communities in social and information networks, i.e., in graphs in which the nodes represent underlying social entities and the edges represent interactions between pairs of nodes. Most recent papers on the subject of community detection in large networks begin by noting that it is a matter of common experience that communities exist in such networks. These papers then note that although there is no agreed-upon definition of a community, a community should be thought of as a set of nodes that has more and/or better connections between its members than between its members and the remainder of the network. These papers then apply a range of algorithmic techniques and intuitions to extract subsets of nodes and then interpret these subsets as meaningful communities corresponding to some underlying "true" real-world communities. In this paper, we explore from a novel perspective several questions related to identifying meaningful communities in large sparse networks, and we come to several striking conclusions that have implications for community detection and graph partitioning in such networks. We emphasize that, in contrast to most of the previous work on this subject, we look at very large networks of up to millions of nodes, and we observe very different phenomena from what is seen in small commonly analyzed networks.

1.1. Overview of Our Approach

At the risk of oversimplifying the large and often intricate body of work on community detection in complex networks, the following five-part story describes the general methodology:

(1) Data are modeled by an "interaction graph." In particular, part of the world gets mapped to a graph in which nodes represent entities and edges represent some type of interaction between pairs of those entities. For example, in a social network, nodes may represent individual people and edges may represent friendships, interactions, or communication between pairs of those people.

(2) The hypothesis is made that the world contains groups of entities that interact more strongly among themselves than with the outside world, and hence the interaction graph should contain sets of nodes, i.e., communities, that have more and/or better-connected "internal edges" connecting members of the set than "cut edges" connecting the set to the rest of the world.

(3) An objective function or metric is chosen to formalize this idea of groups with more intragroup than intergroup connectivity.

(4) An algorithm is then selected to find sets of nodes that exactly or approximately optimize this or some other related metric. Sets of nodes that the algorithm finds are then called "clusters," "communities," "groups," "classes," or "modules."

(5) The clusters or communities or modules are evaluated in some way. For example, one may map the sets of nodes back to the real world to see whether they appear to make intuitive sense as a plausible "real" community. Alternatively, one may attempt to acquire some form of "ground truth," in which case the set of nodes output by the algorithm may be compared with it.

With respect to points (1)–(4), we follow the usual path. In particular, we adopt points (1) and (2), and we then explore the consequence of making such a choice, i.e., of making such a hypothesis and modeling assumption. For point (3), we choose a natural and widely adopted notion of community goodness (community quality score) called *conductance*, which is also known as the normalized cut metric [Chung 97, Shi and Malik 00, Kannan et al. 04]. Informally, the conductance of a set of nodes (defined and discussed in more detail in Section 2.3) is the ratio of the number of "cut" edges between that set and its complement divided by the number of "internal" edges inside that set. Thus, to be a good community, a set of nodes should have small conductance, i.e., it should have many internal edges and few edges pointing to the rest of the network. Conductance is widely used to capture the intuition of a good community; it is a fundamental combinatorial quantity; and it has a very natural interpretation in terms of random

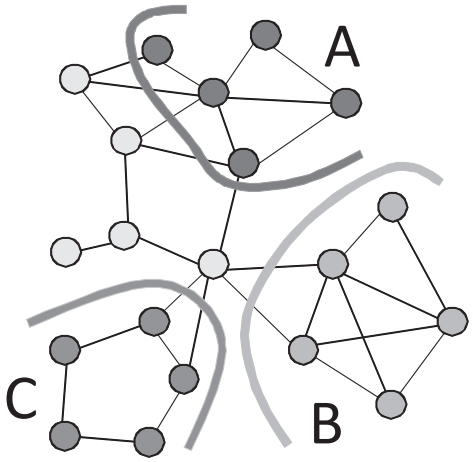


Figure 1. Network communities. Of the three five-node sets that have been marked, B has the best (i.e., the lowest) conductance, since it has the lowest ratio between the number of cut edges and the number of inside edges. So, set B is the best five-node community or the most community-like set of five nodes in this particular network.

walks on the interaction graph. Moreover, since there exists a rich suite of both theoretical and practical algorithms [Hendrickson and Leland 95, Spielman and Teng 96, Leighton and Rao 88, Leighton and Rao 99, Arora et al. 04b, Karypis and Kumar 98a, Karypis and Kumar 98b, Zhao and Karypis 04, Dhillon et al. 07], we can for point (4) compare and contrast several methods to approximately optimize it.

To illustrate conductance, note that of the three five-node sets A, B, and C illustrated in the graph in Figure 1, B has the best (the lowest) conductance and is thus the most community-like.

However, it is in point (5) that we deviate from previous work. Instead of focusing on individual groups of nodes and trying to interpret them as "real" communities, we investigate statistical properties of a large number of communities over a wide range of size scales in over one hundred large sparse real-world social and information networks.

We take a step back and ask questions such as, How well do real graphs split into communities? What is a good way to measure and characterize presence or absence of community structure in networks? What are typical community sizes and typical community qualities?

To address these and related questions, we introduce the concept of a *network community profile (NCP) plot*, which we define and describe in more detail in

Section 3.1. Intuitively, the network community profile plot measures the score of "best" community as a function of community size in a network. Formally, we define it as the conductance value of the minimum-conductance set of cardinality k in the network, as a function of k. As defined, the NCP plot will be NP-hard to compute exactly, so operationally we will use several natural approximation algorithms for solving the minimum-conductance cut problem in order to compute different approximations to it. By comparing and contrasting these plots for a large number of networks, and by computing other related structural properties, we obtain results that suggest a significantly more refined picture of the community structure in large real-world networks than has been appreciated previously.

We have gone to a great deal of effort to be confident that we are computing quantities fundamental to the networks we are considering, rather than artifacts of the approximation algorithms we employ. In particular:

- We use several classes of graph-partitioning algorithms to probe the networks for sets of nodes that could plausibly be interpreted as communities. These algorithms, including flow-based methods, spectral methods, and hierarchical methods, have complementary strengths and weaknesses that are well understood both in theory and in practice. For example, flow-based methods are known to have difficulties with expanders [Leighton and Rao 88, Leighton and Rao 99], and flow-based postprocessing of other methods is known in practice to yield cuts with extremely good conductance values [Lang 04, Lang and Rao 04]. On the other hand, spectral methods are known to have difficulties when they confuse long paths with deep cuts [Spielman and Teng 96, Guattery and Miller 98], a consequence of which is that they may be viewed as computing a "regularized" approximation to the network community profile plot. (See Section 5 for a more detailed discussion of these and related issues.)

- We compute spectral-based lower bounds and semidefinite-programming-based lower bounds as well for the conductance of our network data sets.

- We compute a wide range of other structural properties of the networks, such as sizes, degree distributions, maximum and average diameters of the purported communities, and internal versus external conductance values of the purported communities.

- We recompute statistics on versions of the networks that have been modified in well-understood ways, e.g., by removing small barely connected sets of nodes or by randomizing the edges.

- We compare our results across not only over one hundred large social and information networks, but also over numerous commonly studied small social networks, expanders, and low-dimensional manifold-like objects, and we compare our results on each network with what is known from the field from which the network is drawn. To our knowledge, this makes ours the most extensive such analysis of the community structure in large real-world social and information networks.

- We compare results with analytical and/or simulational results on a wide range of commonly and not so commonly used network-generation models [Newman 03, Bollobás and Riordan 04, Albert and Barabási 99, Kumar et al. 00, Ravasz and Barabási 03, Leskovec et al. 05b, Flaxman et al. 04, Flaxman et al. 07].

1.2. Summary of Our Results

Main Empirical Findings. Taken as a whole, the results we present in this paper suggest a rather detailed and somewhat counterintuitive picture of the community structure in large social and information networks. Several qualitative properties of community structure, as revealed by the network community profile plot, are nearly universal:

- Up to a size scale, which empirically is roughly one hundred nodes, not only do there exist cuts with relatively good conductance, i.e., good communities, but also the slope of the network community profile plot is generally sloping downward. This latter point suggests that smaller communities can be combined into meaningful larger communities, a phenomenon that we empirically observe in many cases.

- At the size scale of roughly one hundred nodes, we often observe the global minimum of the network community profile plot; these are the "best" communities, according to the conductance measure, in the entire graph. These are, however, rather interestingly connected to the rest of the network; for example, in most cases, we observe empirically that they are a small set of nodes barely connected to the remainder of the network by just a *single* edge.

- Above the size scale of roughly one hundred nodes, the network community profile plot gradually increases in size, and thus there is a nearly inverse relationship between community size and community quality. As a function of increasing size, the best possible communities become more and more "blended into" the remainder of the network. Intuitively, communities

blend in with one another and gradually disappear as they grow larger. In particular, in many cases, larger communities can be broken into smaller and smaller pieces, often recursively, each of which is more community-like than the original supposed community.

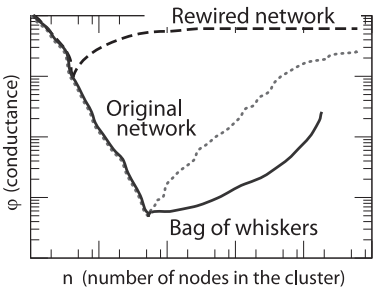
- Even up to the largest size scales, we observe significantly more structure than would be seen, for example, in an expander-like random graph on the same degree sequence.

A schematic picture of a typical network community profile plot is illustrated in Figure 2(a). With a dotted line (labeled as "original network"), we plot community size against community quality score for the sets of nodes extracted from the original network. With a dashed line (rewired network), we plot the scores of communities extracted from a random network conditioned on the same degree distribution as the original network.

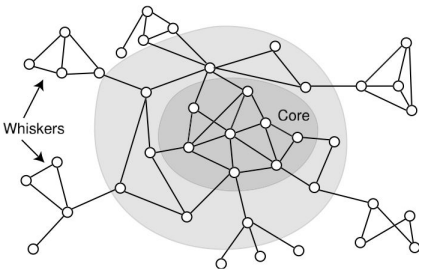
This illustrates not only tight communities at very small scales, but also that at larger and larger size scales (the precise cutoff point for which is difficult to specify precisely), the best possible communities gradually "blend in" more and more with the rest of the network and thus gradually become less and less community-like. Eventually, even the existence of large well-defined communities is quite questionable if one models the world with an interaction graph, as in point (1) above, and if one also defines good communities as densely linked clusters that are weakly connected to the outside, as in hypothesis (2) above.

Finally, with a solid line (bag of whiskers), we also plot the scores of communities that are composed of disconnected pieces (found according to a procedure we describe in Section 4). This curve shows, perhaps somewhat surprisingly, that one can often obtain better community quality scores by combining unrelated disconnected pieces.

To understand the properties of generative models sufficient to reproduce the phenomena we have observed, we have examined in detail the structure of our social and information networks. Although nearly every network is an exception to any simple rule, we have observed that an "octopus" or "jellyfish" model [Chung and Lu 06a, Tauro et al. 01, Siganos et al. 06] provides a rough first approximation to the structure of many of the networks we have examined. That is, most networks may be viewed as having a "core," with no obvious underlying geometry and containing a constant fraction of the nodes, and then there is a periphery consisting of a large number of relatively small "whiskers" that are only tenuously connected to the core. Figure 2(b) presents a caricature of this network structure. Of course, our network data sets are far from random in numerous ways. For instance, they have higher edge density in the core; the small barely connected whisker-like pieces are generally larger, denser, and more



(a) Typical NCP plot



(b) Caricature of network structure

Figure 2. (a) Typical network community profile plot for a large social or information network: networks have better and better communities up to a size scale of ≈ 100 nodes, and beyond that size scale, communities "blend in" with the rest of the network (dotted curve). However, real networks still have more structure than their randomized (conditioned on the same degree distribution) counterparts (dashed curve). Even more surprisingly, if one allows for disconnected communities (solid curve), the community quality scores often get even better (even though such communities have no intuitive meaning). (b) Network structure for a large social or information network, as suggested by our empirical evaluations. See the text for more information on the "core" and "whiskers," and note that the core in our real-world networks is actually extremely sparse.

common than in corresponding random graphs; they have higher local clustering coefficients; and this local clustering information gets propagated globally into larger clusters or communities in a subtle and location-specific manner. More interestingly, as shown in Figure 13 in Section 4.4, the core itself consists of a nested core–periphery structure.

Main Modeling Results. The behavior that we observe is not reproduced, at even a qualitative level, by any of the commonly used network-generation models we have examined, including but not limited to preferential attachment models, copying models, small-world models, and hierarchical network models. Moreover, this behavior is qualitatively different from what is observed in networks with an underlying meshlike or manifold-like geometry (which may not be surprising, but is significant insofar as these structures are often used as a scaffolding upon which to build other models), in networks that are good expanders (which may be surprising, since it is often observed that large social networks are expander-like), and in small social networks such as those used as test beds for community-detection algorithms (which may have implications for the applicability of these methods to detect large community-like structures in these networks). For the commonly used network-generation models, as well as for expander-like, low-

dimensional, and small social networks, the network community profile plots are generally downward sloping or relatively flat.

Although it is well understood at a qualitative level that nodes that are "far apart" or "less alike" (in some sense) should be less likely to be connected in a generative model, understanding this point quantitatively so as to reproduce the empirically observed relationship between small-scale and large-scale community structure turns out to be rather subtle. We can make the following observations:

- Very sparse random graph models with no underlying geometry have relatively deep cuts at small size scales, the best cuts at large size scales are very shallow, and there is a relatively abrupt transition in between. (This is shown pictorially in Figure 2(a) for a randomly rewired version of the original network.) This is a consequence of the extreme sparsity of the data: sufficiently dense random graphs do not have these small deep cuts; and the relatively deep cuts in sparse graphs are due to small treelike pieces that are connected by a single edge to a core that is an extremely good expander.

- A forest fire generative model [Leskovec et al. 05b, Leskovec et al. 07b], in which edges are added in a manner that imitates a fire-spreading process, reproduces not only the deep cuts at small size scales and the absence of deep cuts at large size scales but other properties as well: the small barely connected pieces are significantly larger and denser than random; and for appropriate parameter settings the network community profile plot increases relatively gradually as the size of the communities increases.

- The details of the "forest fire" burning mechanism are crucial for reproducing how local clustering information gets propagated to larger size scales in the network, and those details shed light on the failures of commonly used network-generation models. In the forest fire model, a new node selects a "seed" node and links to it. Then with some probability it "burns" or adds an edge to the each of the seed's neighbors, and so on, recursively. Although there is a "preferential attachment" and also a "copying" flavor to this mechanism, two factors are particularly important: first is the local (in a graph sense, for there is no underlying geometry in the model) manner in which the edges are added; and second is that the number of edges that a new node can add can vary widely, depending on the local structure around the seed node. Depending on the neighborhood structure around the seed, small fires will keep the community well separated from the network, but occasional large fires will connect the community to the rest of the network and make it blend into the network core.

Thus, intuitively, the structure of the whiskers (components connected to the rest of the graph via a single edge) is responsible for the downward part of the network community profile plot, while the core of the network and the manner in which the whiskers root themselves to the core help to determine the upward part of the network community profile plot. Due to local clustering effects, whiskers in real networks are larger and give deeper cuts than whiskers in corresponding randomized graphs, fluctuations in the core are larger and deeper than in corresponding randomized graphs, and thus the network community profile plot increases more gradually and levels off to a conductance value well below the value for a corresponding rewired network.

Main Methodological Contributions. To obtain these and other conclusions, we have employed approximation algorithms for graph partitioning to investigate structural properties of our network data sets. Briefly, we have done the following:

- We have used Metis+MQI, which consists in using the popular graph-partitioning package Metis [Karypis and Kumar 98a] followed by a flow-based MQI postprocessing [Lang and Rao 04]. With this procedure, we obtain sets of nodes that have very good conductance scores. At very small size scales, these sets of nodes could plausibly be interpreted as good communities, but at larger size scales, we often obtain tenuously connected (and in some cases unions of disconnected) pieces, which perhaps do not correspond to intuitive communities.

- Thus, we have also used the local spectral method of [Andersen et al. 06] to obtain sets of nodes with good conductance value that are "compact" or more "regularized" than those pieces returned by Metis+MQI. Since spectral methods confuse long paths with deep cuts [Spielman and Teng 96, Guattery and Miller 98], empirically we obtain sets of nodes that have worse conductance scores than sets returned by Metis+MQI, but that are "tighter" and more "community-like." For example, at small size scales the sets of nodes returned by the local spectral algorithm agrees with the output of Metis+MQI, but at larger scales this algorithm returns sets of nodes with substantially smaller diameter and average diameter, which seem plausibly more community-like.

We have also used what we call the bag-of-whiskers heuristic to identify small barely connected sets of nodes that exert a surprisingly large influence on the network community profile plot.

Both Metis+MQI and the local spectral algorithm scale well, and thus either may be used to obtain sets of nodes from very large graphs. For many of

the small to medium-sized networks, we have checked our results by applying one or more other spectral, flow-based, or heuristic algorithms, although these do not scale as well to very large graphs. Finally, for some of our smaller network data sets, we have computed spectral-based and semidefinite-programming-based lower bounds, and the results are consistent with the conclusions we have drawn.

Broader Implications. Our observation that, independently of the network size, compact communities exist only up to a size scale of around 100 nodes agrees well with the "Dunbar number" [Dunbar 98], which predicts that roughly 150 individuals is the upper limit on the size of a well-functioning human community.

Moreover, we should emphasize that our results do not disagree with the literature at small size scales. One reason for the difference in our findings is that previous studies mainly focused on small networks, which are simply not large enough for the clusters to gradually blend into one another as one looks at larger size scales. In order to make our observations, one needs to look at a large number (due to the complex noise properties of real graphs) of large networks. It is only when Dunbar's limit is exceeded by several orders of magnitude that it is relatively easy to observe large communities blurring together and eventually vanishing. A second reason for the difference is that previous work did not measure and examine the *network community profile* of cluster size versus cluster quality. Finally, we should note that our explanation also aligns well with the *common bond* versus *common identity* theory of group attachment [Ren et al. 07] from social psychology, where it has been noted that bond communities tend to be smaller and more cohesive than identity communities [Back 51], since they are based on interpersonal ties, while identity communities are focused around a common theme or interest. We discuss these implications and connections further in Section 7.

1.3. Review of Recent Literature

Since the appearance of the preliminary conference version of this paper, there has continued to be interest in trying to find communities in large networks. Here, we provide a brief summary of the most relevant recent work. Additional pointers may be found in the very recent review [Fortunato 10]. In addition to those papers mentioned in the remainder of this section, we should also mention [Lancichinetti and Fortunato 09, Wang et al. 09, Yang et al. 09, Zinoviev and Duong 09, Zinoviev 08, Boguñá et al. 09].

Several papers (of which we were not originally aware) have observed that networks can consist of a core–periphery structure [Borgattia and Everett 00, Dorogovtsev et al. 06, Holme 05], not unlike the ubiquitous "nested core–periphery

structure" that we report here. On the other hand, it has recently been noted that the use of modularity is "not well suited for dealing with networks having a core–periphery structure" [da Silva et al. 08].

This observation is consistent with our findings reported here, and also with our recent work [Leskovec et al. 10a, Leskovec et al. 10b] that demonstrates that the qualitative properties we report here are robust to other formalizations of the community concept that capture the interconnectivity versus intraconnectivity tension, but that formalizations such as modularity that take into account only one or the other of those phenomena behave in very different ways for rather trivial reasons. In addition, several researchers have also observed that diffusion-based methods can find small communities [Estrada and Hatano 09, Estrada 07, Rosvall and Bergstrom 08]. This observation is consistent with our findings that local spectral methods can find small to medium-sized community-like structures. Finally, several other researchers have explored the idea of viewing communities as clusters of edges, rather than clusters of nodes [Ahn et al. 09, Evans and Lambiotte 09, Gulbahce and Lehmann 08]. This, of course, is a very different notion of "community" from what most people have historically considered, and it will not capture the interconnectivity versus intraconnectivity tension. On the other hand, not only is it consistent with our findings, but it also provides a very reasonable mechanism to construct so-called overlapping communities, which have been of recent interest.

1.4. Outline of the Paper

The rest of the paper is organized as follows. In Section 2 we describe some useful background, including a brief description of the network data sets we have analyzed. Then, in Section 3 we present our main results on the properties of the network community profile plot for our network data sets. We place an emphasis on how the phenomena we observe in large social and information networks are qualitatively different from what one would expect based on intuition from and experience with expander-like graphs, low-dimensional networks, and commonly studied small social networks.

Then, in Sections 4 and 5, we summarize the results of additional empirical evaluations. In particular, in Section 4, we describe some of the observations we have made in an effort to understand what structural properties of these large networks are responsible for the phenomena we observe; and in Section 5, we describe some of the results of probing the networks with different approximation algorithms in an effort to be confident that the phenomena we observed really are properties of the networks we study, rather than artifactual properties of the algorithms we chose to use to study those networks.

We follow this in Section 6 with a discussion of complex-network-generation models. We observe that the commonly used network-generation models fail to reproduce the counterintuitive phenomena we observe. We also notice that very sparse random networks reproduce certain aspects of the phenomena, and that a generative model based on an iterative "forest fire" burning mechanism reproduces very well the qualitative properties of the phenomena we observe. Finally, in Section 7 we provide a discussion of our results in a broader context, and in Section 8 we present a brief conclusion. Color versions of several figures and corresponding text may be found in [Leskovec et al. 08a].

2. Background on Communities and Overview of Our Methods

In this section, we will provide background on our data and methods. We start in Section 2.1 with a description of the network data sets we will analyze. Then, in Section 2.2, we review related community-detection and graph-clustering ideas. Finally, in Section 2.3, we provide a brief description of approximation algorithms that we will use. There is a large number of reviews on topics related to those discussed in this paper. For example, see the reviews on community identification [Newman 04, Danon et al. 05], data clustering [Jain et al. 99], graph and spectral clustering [Gaertler 05, von Luxburg 06, Schaeffer 07], graph and heavy-tailed data analysis [Newman 05, Chakrabarti and Faloutsos 06, Clauset et al. 07], surveys on various aspects of complex networks [Albert and Barabási 02, Dorogovtsev and Mendes 02, Newman 03, Bollobás and Riordan 04, Costa et al. 07, Li et al. 06, Boccaletti et al. 06], the monographs on spectral graph theory and complex networks [Chung 97, Chung and Lu 06a], and the book on social network analysis [Wasserman and Faust 94]. See Section 7 for a more detailed discussion of the relationship of our work to some of this prior work.

2.1. Social and Information Network Data Sets That We Analyze

We have examined a large number of real-world complex networks. See Tables 1, 2, and 3 for a summary. For convenience, we have organized the networks into the following categories: social networks; information/citation networks; collaboration networks; web graphs; Internet networks; bipartite affiliation networks; biological networks; low-dimensional networks; IMDB networks; and Amazon networks. We have also examined numerous small social networks that have been used as a test bed for community-detection algorithms (e.g., Zachary's karate club [Zachary 77, Newman 09], interactions between dolphins [Lusseau et al. 03, Newman 09], interactions between monks [Sampson 68, Newman 09], Newman's science network [Newman 06a, Newman 09]), numerous simple network

models in which by design there is an underlying geometry (e.g., power grid and road networks [Watts and Strogatz 98], simple meshes, low-dimensional manifolds including graphs corresponding to the well-studied "Swiss roll" data set [Tenenbaum et al. 00], a geometric preferential attachment model [Flaxman et al. 04, Flaxman et al. 07]), several networks that are very good expanders, and many simulated networks generated by commonly used network-generation models (e.g., preferential attachment models [Newman 03], copying models [Kumar et al. 00], hierarchical models [Ravasz and Barabási 03]).

Social Networks. The class of social networks in Table 1 is particularly diverse and interesting. It includes several large online social networks: a network of professional contacts from LinkedIn (LINKEDIN); a friendship network of a LiveJournal blogging community (LIVEJOURNAL01); and a who-trusts-whom network of Epinions (EPINIONS). It also includes an email network from Enron (EMAIL-ENRON) and from a large European research organization. For the latter we generated three networks: EMAIL-INSIDE uses only the communication inside the organization; EMAIL-INOUT adds external email addresses for which email has been both sent and received; and EMAIL-ALL adds all communication both inside the organization and to the outside world. Also included in the class of social networks are networks that are not the central focus of the websites from which they come, but that instead serve as a tool for people to share information more easily. For example, we have the networks of a social bookmarking site Delicious (DELICIOUS); a Flickr photo sharing website (FLICKR); and a network from the Yahoo! Answers question-answering website (ANSWERS). In all these networks, a node refers to an individual, and an edge is used to indicate that one person has some sort of interaction with another person, e.g., one person subscribes to a neighbor's bookmarks or photos, or answers the neighbor's questions.

Information and Citation Networks. The class of information/citation networks contains several different citation networks. It contains two citation networks of physics papers on arxiv.org (CIT-HEP-TH and CIT-HEP-PH), and a network of citations of U.S. patents (CIT-PATENTS). (These paper-to-paper citation networks are to be distinguished from scientific collaboration networks and author-to-paper bipartite networks, as described below.) It also contains two types of blog citation networks. In the so-called post networks, nodes are posts and edges represent hyperlinks between blog posts (POST-NAT05-6M and POST-NAT06ALL). On the other hand, the so-called blog network is the blog-level aggregation of the same data, i.e., there is a link between two blogs if there is a post in the first blog that links the post in a second blog (BLOG-NAT05-6M and BLOG-NAT06ALL).

Network	N	E	N_b	E_b	$\bar{d}$	$\tilde{d}$	$\bar{C}$	D	$\bar{D}$	Description
Social networks										
Delicious	147,567	301,921	0.40	0.65	4.09	48.44	0.30	24	6.28	del.icio.us collaborative tagging social network
Epinions	75,877	405,739	0.48	0.90	10.69	183.88	0.26	15	4.27	Who-trusts-whom network from epinions.com [Richardson 03]
Flickr	404,733	2,110,078	0.33	0.86	10.43	442.75	0.40	18	5.42	Flickr photo sharing social network [Kumar et al. 06]
LinkedIn	6,946,668	30,507,070	0.47	0.88	8.78	351.66	0.23	23	5.43	Social network of professional contacts
LiveJournal01	3,766,521	30,629,297	0.78	0.97	16.26	111.24	0.36	23	5.55	Friendship network of a blogging community [Backstrom et al. 06]
LiveJournal11	4,145,160	34,469,135	0.77	0.97	16.63	122.44	0.36	23	5.61	Friendship network of a blogging community [Backstrom et al. 06]
LiveJournal12	4,843,953	42,845,684	0.76	0.97	17.69	170.66	0.35	20	5.53	Friendship network of a blogging community [Backstrom et al. 06]
Messenger	1,878,736	4,079,161	0.53	0.78	4.34	15.40	0.09	26	7.42	Instant messenger social network
Email-All	234,352	383,111	0.18	0.50	3.27	576.87	0.50	14	4.07	Research organization email network (all addresses) [Leskovec et al. 07b]
Email-InOut	37,803	114,199	0.47	0.82	6.04	165.73	0.58	8	3.74	(all addresses but email has to be sent both ways) [Leskovec et al. 07b]
Email-Inside	986	16,064	0.90	0.99	32.58	74.66	0.45	7	2.60	(only emails inside the research organization) [Leskovec et al. 07b]
Email-Enron	33,696	180,811	0.61	0.90	10.73	142.36	0.71	13	3.99	Enron email data set [Klimt and Yang 04]
Answers	488,484	1,240,189	0.45	0.78	5.08	251.78	0.11	22	5.72	Yahoo Answers social network
Answers-1	26,971	91,812	0.56	0.87	6.81	59.17	0.08	16	4.49	Cluster 1 from Yahoo Answers
Answers-2	25,431	65,551	0.48	0.80	5.16	56.57	0.10	15	4.76	Cluster 2 from Yahoo Answers
Answers-3	45,122	165,648	0.53	0.87	7.34	417.83	0.21	15	3.94	Cluster 3 from Yahoo Answers
Answers-4	93,971	266,199	0.49	0.82	5.67	94.48	0.08	16	4.91	Cluster 4 from Yahoo Answers
Answers-5	5,313	11,528	0.41	0.73	4.34	29.55	0.12	14	4.75	Cluster 5 from Yahoo Answers
Answers-6	290,351	613,237	0.40	0.71	4.22	57.16	0.09	22	5.92	Cluster 6 from Yahoo Answers
Information (citation) networks										
Cit-Patents	3,764,105	16,511,682	0.82	0.96	8.77	21.34	0.09	26	8.15	Citation network of all US patents [Leskovec et al. 05b]
Cit-hep-ph	34,401	420,784	0.96	1.00	24.46	63.50	0.30	14	4.33	Citations between physics (ArXiv `hep-th`) papers [Gehrke et al. 03]
Cit-hep-th	27,400	352,021	0.94	0.99	25.69	106.40	0.33	15	4.20	Citations between physics (ArXiv `hep-ph`) papers [Gehrke et al. 03]
Blog-nat05-6m	29,150	182,212	0.74	0.96	12.50	342.51	0.24	10	3.40	Blog citation network (6 months of data) [Leskovec et al. 07c]
Blog-nat06all	32,384	315,713	0.87	0.99	19.50	153.08	0.20	18	3.94	Blog citation network (1 year of data) [Leskovec et al. 07c]
Post-nat05-6m	238,305	297,338	0.21	0.34	2.50	39.51	0.13	45	10.34	Blog post citation network (6 months) [Leskovec et al. 07c]
Post-nat06all	437,305	565,072	0.22	0.38	2.58	35.54	0.11	54	10.48	Blog post citation network (1 year) [Leskovec et al. 07c]
Collaboration networks										
AtA-IMDB	883,963	27,473,042	0.87	0.99	62.16	517.40	0.79	15	3.48	IMDB actor collaboration network from Dec 2007
CA-astro-ph	17,903	196,972	0.89	0.98	22.00	65.70	0.67	14	4.21	Co-authorship in `astro-ph` of arxiv.org [Leskovec et al. 05b]
CA-cond-mat	21,363	91,286	0.81	0.93	8.55	22.47	0.70	15	5.36	Co-authorship in `cond-mat` category [Leskovec et al. 05b]
CA-gr-qc	4,158	13,422	0.64	0.78	6.46	17.98	0.66	17	6.10	Co-authorship in `gr-qc` category [Leskovec et al. 05b]
CA-hep-ph	11,204	117,619	0.81	0.97	21.00	130.88	0.69	13	4.71	Co-authorship in `hep-ph` category [Leskovec et al. 05b]
CA-hep-th	8,638	24,806	0.68	0.85	5.74	12.99	0.58	18	5.96	Co-authorship in `hep-th` category [Leskovec et al. 05b]
CA-DBLP	317,080	1,049,866	0.67	0.84	6.62	21.75	0.73	23	6.75	DBLP co-authorship network [Backstrom et al. 06]

Table 1. Network data sets we analyzed. Statistics of networks we consider: number of nodes N; number of edges E; fraction nodes not in whiskers (size of largest biconnected component) N_b/N; fraction of edges in biconnected component E_b/E; average degree $\bar{d} = 2E/N$; second-order average degree $\tilde{d}$; average clustering coefficient $\bar{C}$; diameter D; and average path length $\bar{D}$.

Network	N	E	N_b	E_b	$\bar{d}$	$\tilde{d}$	$\bar{C}$	D	$\bar{D}$	Description
Web graphs										
Web-BerkStan	319,717	1,542,940	0.57	0.88	9.65	1,067.55	0.32	35	5.66	Web graph of Stanford and UC Berkeley [Khalil and Liu 04]
Web-Google	855,802	4,291,352	0.75	0.92	10.03	170.35	0.62	24	6.27	Web graph Google released in 2002 [Google 02]
Web-NotreDame	325,729	1,090,108	0.41	0.76	6.69	280.68	0.47	46	7.22	Web graph of University of Notre Dame [Albert et al. 99]
Web-Trec	1,458,316	6,225,033	0.59	0.78	8.54	682.89	0.68	112	8.58	Web graph of TREC WT10G web corpus [Glasgow 00]
Internet networks										
As-RouteViews	6,474	12,572	0.62	0.80	3.88	164.81	0.40	9	3.72	AS from Oregon Exchange BGP Route View [Leskovec et al. 05b]
As-Caida	26,389	52,861	0.61	0.81	4.01	281.93	0.33	17	3.86	CAIDA AS Relationships Data Set
As-Skitter	1,719,037	12,814,089	0.99	1.00	14.91	9,934.01	0.17	5	3.44	AS from traceroutes run daily in 2005 by Skitter
As-Newman	22,963	48,436	0.65	0.83	4.22	261.46	0.35	11	3.83	AS graph from Newman [Newman 09]
As-Oregon	13,579	37,448	0.72	0.90	5.52	235.97	0.46	9	3.58	Autonomous systems [Oregon 97]
Gnutella-25	22,663	54,693	0.59	0.83	4.83	10.75	0.01	11	5.57	Gnutella network on March 25 2000 [Ripeanu et al. 02]
Gnutella-30	36,646	88,303	0.55	0.81	4.82	11.46	0.01	11	5.75	Gnutella P2P network on March 30 2000 [Ripeanu et al. 02]
Gnutella-31	62,561	147,878	0.54	0.81	4.73	11.60	0.01	11	5.94	Gnutella network on March 31 2000 [Ripeanu et al. 02]
eDonkey	5,792,297	147,829,887	0.93	1.00	51.04	6,139.99	0.08	5	3.66	P2P eDonkey graph for a period of 47 hours in 2004
Bi-partite networks										
IpTraffic	2,250,498	21,643,497	1.00	1.00	19.23	94,889.05	0.00	5	2.53	IP traffic graph a single router for 24 hours
AtP-astro-ph	54,498	131,123	0.70	0.87	4.81	16.67	0.00	28	7.78	Authors-to-papers network of `astro-ph` [Leskovec et al. 07c]
AtP-cond-mat	57,552	104,179	0.65	0.79	3.62	10.54	0.00	31	9.96	Authors-to-papers network of `cond-mat` [Leskovec et al. 07c]
AtP-gr-qc	14,832	22,266	0.47	0.60	3.00	9.72	0.00	35	11.08	Authors-to-papers network of `gr-qc` [Leskovec et al. 07c]
AtP-hep-ph	47,832	86,434	0.60	0.76	3.61	16.80	0.00	27	8.55	Authors-to-papers network of `hep-ph` [Leskovec et al. 07c]
AtP-hep-th	39,986	64,154	0.53	0.68	3.21	13.07	0.00	36	10.74	Authors-to-papers network of `hep-th` [Leskovec et al. 07c]
AtP-DBLP	615,678	944,456	0.49	0.64	3.07	13.61	0.00	48	12.69	DBLP authors-to-papers bipartite network
Spending	1,831,540	2,918,920	0.34	0.58	3.19	1,536.35	0.00	26	5.62	Users-to-keywords they bid
Hw7	653,260	2,278,448	0.99	0.99	6.98	346.85	0.00	24	6.26	Downsampled advertiser-query bid graph
Netflix	497,959	100,480,507	1.00	1.00	403.57	28,432.89	0.00	5	2.31	Users-to-movies they rated. From Netflix prize [Netflix 09]
QueryTerms	13,805,808	17,498,668	0.28	0.41	2.53	14.92	0.00	86	19.81	Users-to-queries they submit to a search engine
Clickstream	199,308	951,649	0.39	0.87	9.55	430.74	0.00	7	3.83	Users-to-URLs they visited [Montgomery and Faloutsos 01]
Biological networks										
Bio-Proteins	4,626	14,801	0.72	0.91	6.40	24.25	0.12	12	4.24	Yeast protein interaction network [Colizza et al. 05]
Bio-Yeast	1,458	1,948	0.37	0.51	2.67	7.13	0.14	19	6.89	Yeast protein interaction network data [Jeong et al. 01]
Bio-YeastP0.001	353	1,517	0.73	0.93	8.59	20.18	0.57	11	4.33	Yeast protein-protein interaction map [Qi et al. 06]
Bio-YeastP0.01	1,266	8,511	0.79	0.97	13.45	47.73	0.44	12	3.87	Yeast protein-protein interaction map [Qi et al. 06]

Table 2. Network data sets we analyzed. Statistics of networks we consider: number of nodes N; number of edges E; fraction nodes not in whiskers (size of largest biconnected component) N_b/N; fraction of edges in biconnected component E_b/E; average degree $\bar{d} = 2E/N$; second-order average degree $\tilde{d}$; average clustering coefficient $\bar{C}$; diameter D; and average path length $\bar{D}$.

Network	N	E	N_b	E_b	$\bar{d}$	$\tilde{d}$	$\bar{C}$	D	$\bar{D}$	Description
Nearly low-dimensional networks										
ROAD-CA	1,957,027	2,760,388	0.80	0.85	2.82	3.17	0.06	865	310.97	California road network
ROAD-USA	126,146	161,950	0.97	0.98	2.57	2.81	0.03	617	218.55	USA road network (only main roads)
ROAD-PA	1,087,562	1,541,514	0.79	0.85	2.83	3.20	0.06	794	306.89	Pennsylvania road network
ROAD-TX	1,351,137	1,879,201	0.78	0.84	2.78	3.15	0.06	1,064	418.73	Texas road network
POWERGRID	4,941	6,594	0.62	0.69	2.67	3.87	0.11	46	19.07	Power grid of Western States Power Grid [Watts and Strogatz 98]
MANI-FACES7K	696	6,979	0.98	0.99	20.05	37.99	0.56	16	5.52	Faces (64x64 gray-scale images) (connect 7k closest pairs)
MANI-FACES4K	663	3,465	0.90	0.97	10.45	20.20	0.56	29	8.96	Faces (connect 4k closest pairs)
MANI-FACES2K	551	1,981	0.84	0.94	7.19	12.77	0.54	32	11.07	Faces (connect 2k closest pairs)
MANI-FACESK10	698	6,935	1.00	1.00	19.87	25.32	0.51	6	3.25	Faces (connect every to 10 nearest neighbors)
MANI-FACESK3	698	2,091	1.00	1.00	5.99	7.98	0.45	9	4.89	Faces (connect every to 5 nearest neighbors)
MANI-FACESK5	698	3,480	1.00	1.00	9.97	12.91	0.48	7	4.03	Faces (connect every to 3 nearest neighbors)
MANI-SWISS200K	20,000	200,000	1.00	1.00	20.00	21.08	0.59	103	37.21	Swiss-roll (connect 200k nearest pairs of nodes)
MANI-SWISS100K	19,990	99,979	1.00	1.00	10.00	11.02	0.59	162	58.32	Swiss-roll (connect 100k nearest pairs of nodes)
MANI-SWISS60K	19,042	57,747	0.93	0.96	6.07	7.03	0.59	243	89.15	Swiss-roll (connect 60k nearest pairs of nodes)
MANI-SWISSK10	20,000	199,955	1.00	1.00	20.00	25.38	0.56	10	5.47	Swiss-roll (every node connects to 10 nearest neighbors)
MANI-SWISSK5	20,000	99,990	1.00	1.00	10.00	12.89	0.54	13	8.34	Swiss-roll (every node connects to 5 nearest neighbors)
MANI-SWISSK3	20,000	59,997	1.00	1.00	6.00	7.88	0.50	17	6.89	Swiss-roll (every node connects to 3 nearest neighbors)
IMDB Actor-to-Movie graphs										
ATM-IMDB	2,076,978	5,847,693	0.49	0.82	5.63	65.41	0.00	32	6.82	Actors-to-movies graph from IMDB (imdb.com)
IMDB-TOP30	198,430	566,756	0.99	1.00	5.71	18.19	0.00	26	8.32	Actors-to-movies graph heavily preprocessed
IMDB-RAW07	601,481	1,320,616	0.54	0.79	4.39	20.94	0.00	32	8.55	Country clusters were extracted from this graph
IMDB-FRANCE	35,827	74,201	0.51	0.76	4.14	14.62	0.00	20	6.57	Cluster of French movies
IMDB-GERMANY	21,258	42,197	0.56	0.78	3.97	13.69	0.00	34	7.47	German movies (to actors that played in them)
IMDB-INDIA	12,999	25,836	0.57	0.78	3.98	31.55	0.00	19	6.00	Indian movies
IMDB-ITALY	19,189	37,534	0.55	0.77	3.91	11.66	0.00	30	6.91	Italian movies
IMDB-JAPAN	15,042	34,131	0.60	0.82	4.54	16.98	0.00	19	6.81	Japanese movies
IMDB-MEXICO	13,783	36,986	0.64	0.86	5.37	24.15	0.00	19	5.43	Mexican movies
IMDB-SPAIN	15,494	31,313	0.51	0.76	4.04	14.22	0.00	28	6.44	Spanish movies
IMDB-UK	42,133	82,915	0.52	0.76	3.94	15.14	0.00	23	7.04	UK movies
IMDB-USA	241,360	530,494	0.51	0.78	4.40	25.25	0.00	30	7.63	USA movies
IMDB-WGERMANY	12,120	24,117	0.56	0.78	3.98	11.73	0.00	22	6.26	West German movies
Amazon product co-purchasing networks										
AMAZON0302	262,111	899,792	0.95	0.97	6.87	11.14	0.43	38	8.85	Amazon products from 2003 03 02 [Clauset et al. 04]
AMAZON0312	400,727	2,349,869	0.94	0.99	11.73	30.33	0.42	20	6.46	Amazon products from 2003 03 12 [Clauset et al. 04]
AMAZON0505	410,236	2,439,437	0.94	0.99	11.89	30.93	0.43	22	6.48	Amazon products from 2003 05 05 [Clauset et al. 04]
AMAZON0601	403,364	2,443,311	0.96	0.99	12.11	30.55	0.43	25	6.42	Amazon products from 2003 06 01 [Clauset et al. 04]
AMAZONALL	473,315	3,505,519	0.94	0.99	14.81	52.70	0.41	19	5.66	Amazon products (all 4 graphs merged) [Clauset et al. 04]
AMAZONALLPROD	524,371	1,491,793	0.80	0.91	5.69	11.75	0.35	42	11.18	Products (all products, source+target) [Leskovec et al. 07a]
AMAZONSRCPROD	334,863	925,872	0.84	0.91	5.53	11.53	0.43	47	12.11	Products (only source products) [Leskovec et al. 07a]

Table 3. Network data sets we analyzed. Statistics of networks we consider: number of nodes N; number of edges E; fraction nodes not in whiskers (size of largest biconnected component) N_b/N; fraction of edges in biconnected component E_b/E; average degree $\bar{d} = 2E/N$; second-order average degree $\tilde{d}$; average clustering coefficient $\bar{C}$; diameter D; and average path length $\bar{D}$.

Collaboration Networks. The class of collaboration networks contains academic collaboration (i.e., coauthorship) networks between physicists from various categories in arxiv.org (CA-ASTRO-PH, etc.) and between authors in computer science (CA-DBLP). It also contains a network of collaborations between pairs of actors in IMDB (ATA-IMDB), i.e., there is an edge connecting a pair of actors if they appeared in the same movie. (Again, this should be distinguished from actor-to-movie bipartite networks, as described below.)

Web Graphs. The class of web graph networks includes four different web graphs in which nodes represent web pages and edges represent hyperlinks between those pages. Networks were obtained from Google (WEB-GOOGLE), the University of Notre Dame (WEB-NOTREDAME), TREC (WEB-TREC), and Stanford University (WEB-BERKSTAN). The class of Internet networks consists of various autonomous systems networks obtained from different sources, as well as a Gnutella and eDonkey peer-to-peer file-sharing network.

Bipartite Networks. The class of bipartite networks is particularly diverse and includes authors-to-papers graphs from both computer science (ATP-DBLP) and physics (ATP-ASTRO-PH, etc.); a network representing users and the URLs they visited (CLICKSTREAM); a network representing users and the movies they rated (NETFLIX); and a users-to-queries network representing query terms that users typed into a search engine (QUERYTERMS). (We also have analyzed several bipartite actors-to-movies networks extracted from the IMDB database, which we have listed separately below.)

Biological Networks. The class of biological networks includes protein–protein interaction networks of yeast obtained from various sources.

Low-Dimensional Gridlike Networks. The class of low-dimensional networks consists of graphs constructed from road (ROAD-CA, etc.) or power grid (POWERGRID) connections and as such might be expected to "live" on a two-dimensional surface in a way that all the other networks do not. We also added a "Swiss roll" network, a 2-dimensional manifold embedded in three dimensions, and a "Faces" data set in which each point is a 64×64 gray-scale image of a face (embedded in 4,096-dimensional space) and where we connected the faces that were most similar (using the Euclidean distance).

IMDB, Yahoo! Answers and Amazon Networks. Finally, we have networks from IMDB, Amazon, and Yahoo! Answers, and for each of these we have separately analyzed subnetworks. The IMDB networks consist of actor-to-movie links, and we include the full network as well as subnetworks associated with individual countries based on the country of production. For the Amazon networks, recall

that Amazon sells a variety of products, and for each item A one may compile a list of up to ten other items most frequently purchased by buyers of A. This information can be presented as a directed network in which vertices represent items and there is an edge from item A to another item B if B was frequently purchased by buyers of A. We consider the network undirected. We use five networks from a study of [Clauset et al. 04], and two networks from the viral marketing study from [Leskovec et al. 07a]. Finally, for the Yahoo! Answers networks, we observe several deep cuts at large size scales, and so in addition to the full network, we analyze the top six best-connected subnetworks.

In addition to providing a brief description of the network, Tables 1, 2, and 3 show the number of nodes and edges in each network, as well as other statistics that will be described in Section 4.1. (In all cases, we consider the network to be undirected, and we extract and analyze the largest connected component.) The sizes of these networks range from about 5,000 nodes up to nearly 14 million nodes, and from about 6,000 edges up to more than 100 million edges. All of the networks are quite sparse—their densities range from an average degree of about 2.5 for the blog post network up to an average degree of about 400 in the network of movie ratings from Netflix, and most of the other networks, including the purely social networks, have average degree around 10 (median average degree of 6). In many cases, we examined several versions of a given network. For example, we considered the entire IMDB actor-to-movie network, as well as subnetworks of it corresponding to different language and country groups. Detailed statistics for all these networks are presented in Tables 1, 2, and 3 and are described in Section 4. In total, we have examined over one hundred large real-world social and information networks, making this, to our knowledge, the largest and most comprehensive study of such networks.

2.2. Clusters and Communities in Networks

Hierarchical clustering is a common approach to community identification in the social sciences [Wasserman and Faust 94], but it has also found application more generally [Girvan and Newman 02, Hopcroft et al. 04]. In this procedure, one first defines a distance metric between pairs of nodes and then produces a tree (in either a bottom-up or a top-down manner) describing how nodes group into communities and how these communities group further into supercommunities. A quite different approach that has received a great deal of attention (and that will be central to our analysis) is based on ideas from *graph partitioning* [Schaeffer 07, Brandes et al. 07]. In this case, the network is modeled as a simple undirected graph in which nodes and edges have no attributes, and a partition of the graph is determined by optimizing a merit function. The graph-partitioning

problem is find some number k of groups of nodes, generally with roughly equal size, such that the number of edges between the groups, perhaps normalized in some way, is minimized.

Let $G = (V, E)$ denote a graph. Then the *conductance* ϕ of a set of nodes $S \subset V$ (where S is assumed to contain no more than half of all the nodes) is defined as follows. Let v be the sum of the degrees of the nodes in S, and let s be the number of edges with one endpoint in S and one endpoint in $\overline{S}$, where $\overline{S}$ denotes the complement of S. Then the conductance of S is $\phi = s/v$, or equivalently $\phi = s/(s + 2e)$, where e is the number of edges with both endpoints is S. More formally, we make the following definition.

Definition 2.1. Given a graph G with adjacency matrix A, the *conductance of a set* of nodes S is defined as

$$\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}}, \tag{2.1}$$

where $A(S) = \sum_{i \in S} \sum_{j \in V} A_{ij}$, or equivalently $A(S) = \sum_{i \in S} d(i)$, where $d(i)$ is the degree of node i in G. Moreover, in this case, the *conductance of the graph* G is given by

$$\phi_G = \min_{S \subset V} \phi(S).$$

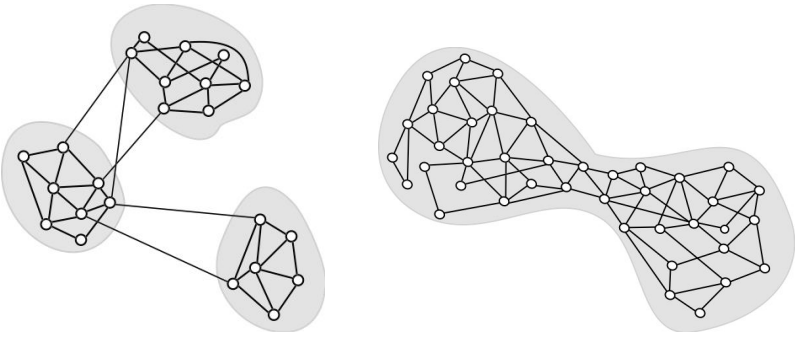
Thus, the conductance of a set provides a measure of the quality of the cut $(S, \overline{S})$, or relatedly, the goodness of a community S.[1]

Indeed, it is often noted that communities should be thought of as sets of nodes with more and/or better intraconnections than interconnections; see Figure 3 for an illustration. When we are interested in detecting communities and evaluating their quality, we prefer sets with small conductances, i.e., sets that are densely linked inside and sparsely linked to the outside.

Although numerous measures have been proposed for the extent to which a set of nodes is like a community, it is commonly noted—see, for example, [Shi and Malik 00] and [Kannan et al. 04]—that conductance captures the "gestalt" notion of clustering [Zahn 71], and as such it has been widely used for graph clustering and community detection [Gaertler 05, von Luxburg 06, Schaeffer 07].

There are many other density-based measures that have been used to partition a graph into a set of communities [Gaertler 05, von Luxburg 06, Schaeffer 07]. One that deserves particular mention is modularity [Newman and

[1] Throughout this chapter we consistently use shorthand phrases such s "this piece has good conductance" to mean "this piece is separated from the rest of the graph by a low-conductance cut."



(a) Three communities (b) Conductance bottleneck

Figure 3. (a) Caricature of the traditional view of communities as sets of nodes with more and/or better intraconnections than interconnections. (b) A graph with its minimum-conductance bottleneck illustrated.

Girvan 04, Newman 06b]. For a given partition of a network into a set of communities, modularity measures the number of within-community edges, relative to a null model that is usually taken to be a random graph with the same degree distribution. Thus, modularity was originally introduced to measure the strength or quality of a particular partition of a network, and that is how it is typically used.

We, however, are interested in a quite different question from those that motivated the introduction of modularity. Rather than seeking to partition a graph into the "best" possible partition of communities, we would like to know how good a particular element of that partition is, i.e., how community-like are the best possible communities that modularity or any other merit function can hope to find, in particular as a function of the size of that partition.

2.3. Approximation Algorithms for Finding Low-Conductance Cuts

In addition to capturing very well our intuitive notion of what it means for a set of nodes to be a good community, the use of conductance as an objective function has an added benefit: there exists an extensive theoretical and practical literature on methods for approximately optimizing it. (Finding cuts with exactly minimal conductance is NP-hard.) In particular, the theory literature contains several algorithms with provable approximation performance guarantees.

First, there is the spectral method, which uses an eigenvector of the graph's Laplacian matrix to find a cut whose conductance is no bigger than ϕ if the graph actually contains a cut with conductance $O(\phi^2)$ [Cheeger 69, Donath and Hoffman 72, Fiedler 73, Mohar 91, Chung 97]. The spectral method also produces

lower bounds that can show that the solution for a given graph is closer to optimal than promised by the worst-case guarantee.

Second, there is an algorithm that uses multicommodity flow to find a cut whose conductance is within an $O(\log n)$ factor of optimal [Leighton and Rao 88, Leighton and Rao 99]. Spectral and multicommodity-flow-based methods are complementary in that the worst-case $O(\log n)$ approximation factor is obtained for flow-based methods on expander graphs [Leighton and Rao 88, Leighton and Rao 99], a class of graphs that does not cause problems for spectral methods, whereas spectral methods can confuse long paths with deep cuts [Guattery and Miller 98, Spielman and Teng 96], a difference that does not cause problems for flow-based methods.

Third, and very recently, there exists an algorithm that uses semidefinite programming to find a solution that is within $O(\sqrt{\log n})$ of optimal [Arora et al. 04b]. The paper cited sparked a flurry of theoretical research on a family of closely related algorithms including [Arora et al. 04a, Khandekar et al. 06, Arora and Kale 07], all of which can be informally described as combinations of spectral and flow-based techniques that exploit their complementary strengths. However, none of those algorithms is currently practicable enough to use in our study.

Of the three above-mentioned theoretical algorithms, the spectral method is by far the most practical. Also very common are recursive bisection heuristics: recursively divide the graph into two groups, and then further subdivide the new groups until the desired number of clusters is achieved. This may be combined with local improvement methods such the Kernighan–Lin and Fiduccia–Mattheyses procedures [Kernighan and Lin 70, Fiduccia and Mattheyses 82], which are fast and can climb out of some local minima. The latter was combined with a multiresolution framework to create Metis [Karypis and Kumar 98a, Karypis and Kumar 98b], a very fast program intended to split meshlike graphs into equal-sized pieces. The authors of Metis later created Cluto [Zhao and Karypis 04], which is better tuned for clustering-type tasks. Finally, we mention Graclus [Dhillon et al. 07], which uses multiresolution techniques and kernel k-means to optimize a metric that is closely related to conductance.

While the preceding are all approximate algorithms for finding the lowest conductance cut in a whole graph, we now mention MQI [Gallo et al. 89, Lang and Rao 04], an *exact* algorithm for the slightly different problem of finding the lowest conductance cut in *half* of a graph. This algorithm can be combined with a good method for initially splitting the graph into two pieces (such as Metis or the spectral method) to obtain a surprisingly strong heuristic method for finding low-conductance cuts in the whole graph [Lang and Rao 04]. The exactness of the second optimization step frequently results in cuts with extremely low conductance scores, as will be visible in many of our plots.

MQI can be implemented by solving single-parametric max-flow problems, or sequences of ordinary max-flow problems. Parametric max-flow (with MQI described as one of the applications) was introduced in [Gallo et al. 89], and recent empirical work is described in [Babenko et al. 07], but currently there is no publicly available code that scales to the sizes we need. Ordinary max-flow is a very thoroughly studied problem. Currently, the best theoretical time bounds are those given in [Goldberg and Rao 98], the most practical algorithm is in [Goldberg and Tarjan 88], while the best implementation is `hi_pr`, from [Cherkassky and Goldberg 95]. Since Metis+MQI using the `hi_pr` code is very fast and scalable, while the method empirically seems usually to find the lowest or nearly lowest conductance cuts in a wide variety of graphs, we have used it extensively in this study.

We will also extensively use the local spectral algorithm of [Andersen et al. 06] to find node sets of low conductance, i.e., good communities, around a seed node. This algorithm is also very fast, and it can be successfully applied to very large graphs to obtain more "well-rounded," "compact," or "evenly connected" communities than those returned by Meits+MQI. The latter observation (described in more detail in Section 5) is important, since local spectral methods also confuse long paths (which tend to occur in our very sparse network data sets) with deep cuts.

This algorithm takes as input two parameters—the seed node and a parameter ϵ that intuitively controls the locality of the computation—and it outputs a set of nodes. Local spectral methods were introduced in [Spielman and Teng 04a]; see also [Andersen et al. 06], and they have roughly the same kind of quadratic approximation guarantees as the global spectral method, but their computational cost is proportional to the size of the obtained piece [Chung 07a, Chung 07c, Chung 07b].

3. The Network Community Profile Plot

In this section, we discuss the *network community profile plot* (NCP plot), which measures the quality of network communities at different size scales. We start in Section 3.1 by introducing the concept. Then, in Section 3.2, we present the NCP plot for several examples of networks that inform people's intuition and for which the NCP plot behaves in a characteristic manner. Then, in Sections 3.3 and 3.4 we present the NCP plot for a wide range of large real-world social and information networks. We will see that in such networks the NCP plot behaves in a qualitatively different manner.

3.1. Definitions for the Network Community Profile Plot

In order to resolve community structure in large networks more finely, we introduce the *network community profile plot* (NCP plot). Intuitively, the NCP plot measures the quality of the best possible community in a large network, as a function of the community size. Formally, we may define it as the conductance value of the best conductance set of cardinality k in the entire network, as a function of k.

Definition 3.1. Given a graph G with adjacency matrix A, the *network community profile plot (NCP plot)* plots $\Phi(k)$ as a function of k, where

$$\Phi(k) = \min_{S \subset V, |S|=k} \phi(S),$$

where $|S|$ denotes the cardinality of the set S, and where the conductance $\phi(S)$ of S is given by (2.1).

Since this quantity is intractable to compute, we will employ well-studied approximation algorithms for the minimum-conductance cut problem to approximate it. In particular, operationally we will use several natural heuristics based on approximation algorithms to do graph partitioning in order to compute different approximations to the NCP plot. Although other procedures will be described in Section 5, we will primarily employ two procedures.

First, Metis+MQI, i.e., the graph-partitioning package Metis [Karypis and Kumar 98a] followed by the flow-based postprocessing procedure MQI [Lang and Rao 04]; this procedure returns sets that have very good conductance values. Second, the local spectral algorithm of [Andersen et al. 06]; this procedure returns sets that are somewhat more "compact" or "smoothed" or "regularized," but that often have somewhat worse conductance values.

Just as the conductance of a set of nodes provides a measure of quality of that set as a community, the shape of the NCP plot provides insight into the community structure of a graph as a whole. For example, the magnitude of the conductance tells us how well clusters of different sizes are separated from the rest of the network. One might hope to obtain some sort of "smoothed" measure of the notion of the best community of size k (e.g., by considering an average of the conductance values over all sets of a given size or by considering a smoothed extremal statistic such as a 95th percentile) rather than conductance of the best set of that size. We have not defined such a measure because there is no obvious way to average over all subsets of size k and obtain a meaningful approximation to the minimum.

On the other hand, our approximation algorithm methodology implicitly incorporates such an effect. Although Metis+MQI finds sets of nodes with extremely good conductance values, empirically we observe that they often have little or no internal structure—they can even be disconnected. On the other hand, since spectral methods in general tend to confuse long paths with deep cuts [Spielman and Teng 96, Guattery and Miller 98], the local spectral algorithm finds sets that are "tighter" and more "well rounded" and thus in many ways more community-like. (See Sections 2.3 and 5 for details on these algorithmic issues and interpretations.)

3.2. Community Profile Plots for Expander, Low-Dimensional, and Small Social Networks

The NCP plot behaves in a characteristic manner for graphs that are "well embeddable" into an underlying low-dimensional geometric structure. To illustrate this, consider Figure 4. In Figure 4(a), we show the results for a 1-dimensional chain, a 2-dimensional grid, and a 3-dimensional cube. In each case, the NCP plot is steadily downward-sloping as a function of the number of nodes in the smaller cluster. Moreover, the curves are straight lines with a slope equal to $-1/d$, where d is the dimensionality of the underlying grids. In particular, as the underlying dimension increases, the slope of the NCP plot becomes less steep. Thus we make the following observation:

> If the network under consideration corresponds to a d-dimensional grid, then the NCP plot shows that
>
> $$-\frac{1}{d} = \frac{\log(\phi(k))}{\log(k)}. \tag{3.1}$$

This is simply a manifestation of the isoperimetric (i.e., surface area to volume) phenomenon: for a grid, the "best" cut is obtained by cutting out a set of adjacent nodes, in which case the surface area (number of edges cut) increases as $O(m^{d-1})$, while the volume (number of vertices/edges inside the cluster) increases as $O(m^d)$.

This qualitative phenomenon of a steadily downward-sloping NCP plot is quite robust for networks that "live" in a low-dimensional structure, e.g., on a manifold or the surface of the Earth. For example, Figure 4(b) shows the NCP plot for a power grid network of Western States Power Grid [Watts and Strogatz 98], and Figure 4(c) shows the NCP plot for a road network in California. These two networks have very different sizes—the power grid network has 4,941 nodes and 6,594 edges, while the road network has 1,957,027 nodes and 2,760,388 edges—and they arise in very different application domains. In both cases, however, we see predominantly downward-sloping NCP plots, very similar to the profile of a

simple 2-dimensional grid. Indeed, the "best-fit" line for the power grid gives a slope of ≈ -0.45, which by (3.1) suggests that $d \approx 2.2$, which is not far from the "true" dimensionality of 2.

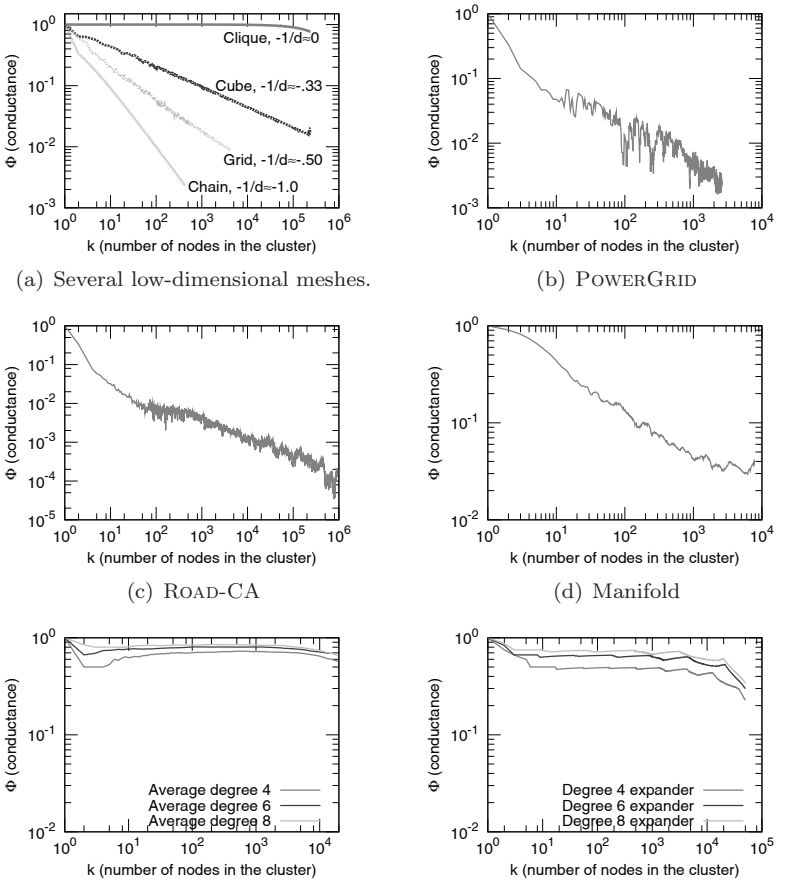
Moreover, empirically we observe that minima in the NCP plot correspond to community-like sets, which are occasionally nested. This corresponds to hierarchical community organization. For example, the nodes giving the dip at $k = 19$ are included in the nodes giving the dip at $k = 883$, while dips at $k = 94$ and $k = 105$ are both included in the dip at $k = 262$.

In a similar manner, Figure 4(d) shows the profile plot for a graph generated from a "Swiss roll" data set that is commonly examined in the manifold and machine-learning literature [Tenenbaum et al. 00]. In this case, we still observe a downward-sloping NCP plot that corresponds to internal dimensionally of the manifold (2 in this case).

Finally, Figures 4(e) and 4(f) show NCP plots for two graphs that are very good expanders. The first is a G_{nm} graph with 100,000 nodes and a number of edges such that the average degree is 4, 6, and 8. The second is a constant-degree expander: to make one with degree d, we take the union of d disjoint but otherwise random complete matchings, and we have plotted the results for $d = 4, 6, 8$. In both of these cases, the NCP plot is roughly flat, which we also observed in Figure 4(a) for a clique, which is to be expected, since the minimum-conductance cut in the entire graph cannot be too small for a good expander [Hoory et al. 06].

Somewhat surprisingly (especially when compared with large networks in Section 3.3), a steadily decreasing downward NCP plot is seen for small social networks that have been extensively studied in validating community-detection algorithms. Several examples are shown in Figure 5. For these networks, the interpretation is similar to that for low-dimensional networks: the downward slope indicates that as potential communities get larger and larger, there are relatively more intra-edges than inter-edges; and empirically we observe that local minima in the NCP plot correspond to sets of nodes that are plausible communities. Consider, for example, Zachary's karate club [Zachary 77] network (ZacharyKarate), an extensively analyzed social network [Newman 04, Newman 06b, Karrer et al. 07]. The network has 34 nodes, each of which represents a member of a karate club, and 78 edges, each of which represents a friendship tie between two members.

Figure 5(a) depicts the karate club network, and Figure 5(b) shows its NCP plot. There are two local minima in the plot: the first dip at $k = 5$ corresponds to the cut A, and the second dip at $k = 17$ corresponds to cut B. Note that cut B, which separates the graph roughly in half, has better conductance value than cut A. This corresponds to one's intuition about the NCP plot derived from



(a) Several low-dimensional meshes.

(b) PowerGrid

(c) Road-CA

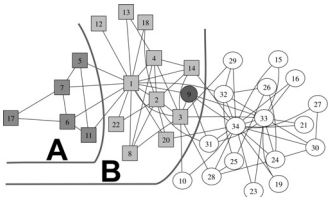
(d) Manifold

(e) Expander: dense G_{nm} graph

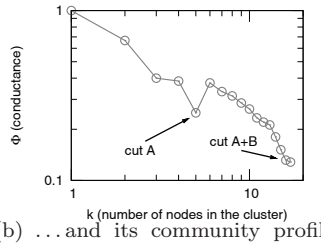
(f) Expander: union of matchings

Figure 4. (Best viewed in color; see [Leskovec et al. 08a].) Network community profile plots for expander-like graphs and several networks that "live" in low-dimensional spaces. (a) A large clique graph, a cube (3D mesh), a grid (2D mesh), and a chain (line). Note that the slope of the community profile plot directly corresponds to the dimensionality of the graph. (b) and (c) Two networks on the Earth's surface and that thus are reasonably well embeddable in two dimensions. (d) A 2D "Swiss roll" manifold embedded in three dimensions, where we have connected every point to ten nearest neighbors. (e) and (f) Two networks that are very good expanders.

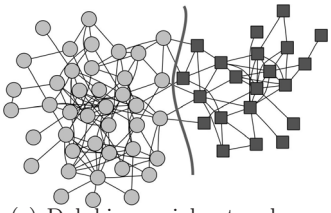
studying low-dimensional graphs. Note also that the karate network corresponds well to the intuitive notion of a community, where nodes of the community are densely linked among themselves and there are few edges between nodes of different communities.



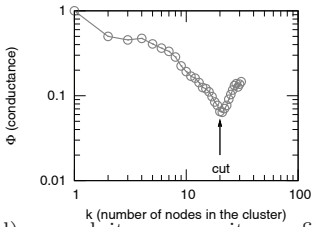
(a) Zachary's karate club network . . .



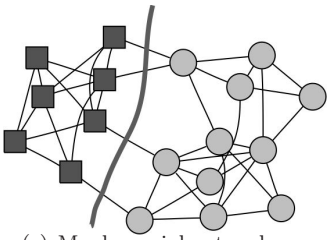
(b) . . . and its community profile plot



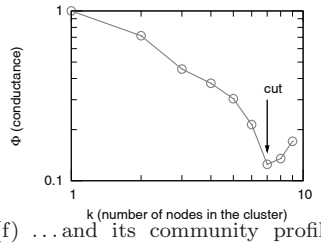
(c) Dolphins social network . . .



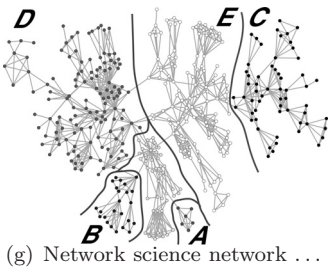
(d) . . . and its community profile plot



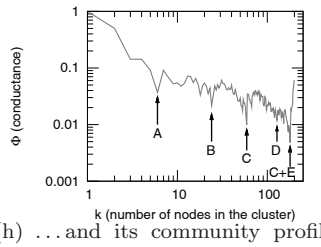
(e) Monks social network . . .



(f) . . . and its community profile plot



(g) Network science network . . .



(h) . . . and its community profile plot

Figure 5. (Best viewed in color; see [Leskovec et al. 08a].) Depiction of several small social networks that are common test sets for community-detection algorithms and their network community profile plots. (a) and (b) Zachary's karate club network. (c) and (d) A network of dolphins. (e) and (f) A network of monks. (g) and (h) A network of researchers researching networks.

In a similar manner, Figure 5(c) shows a social network (with 62 nodes and 159 edges) of interactions within a group of dolphins [Lusseau et al. 03]; Figure 5(e) shows a social network of monks (with 18 nodes representing individual monks and 41 edges representing social ties between pairs of monks) in a cloister [Sampson 68]; and Figure 5(g) depicts Newman's network (with 914 collaborations between 379 researchers) of scientists who conduct research on networks [Newman and Girvan 04]. For each network, the NCP plot exhibits a downward trend, and it has local minima at cluster sizes that correspond to good communities: the minimum for the dolphins network (Figure 5(d)) corresponds to the separation of the network into two communities, denoted by different shapes and shades of the nodes (light circles versus dark squares); the minimum of the monk network (Figure 5(f)) corresponds to the split of 7 Turks (squares) and the so-called loyal opposition (circles) [Sampson 68]; and empirically, both local minima and the global minimum in the network science network (Figure 5(h)) correspond to plausible communities. Note that in the last case, the figure also displays a hierarchical structure whereby the community defined by cut C is included in a larger community that has better conductance value.

At this point, we can observe that the following two general observations hold for networks that are well embeddable in a low-dimensional space and also for small social networks that have been extensively studied and used to validate community-detection algorithms. First, minima in the NCP plots, i.e., the best low-conductance cuts of a given size, correspond to community-like sets of nodes. Second, the NCP plots are generally relatively gradually sloping downward, meaning that smaller communities can be combined into larger sets of nodes that can also be meaningfully interpreted as communities.

3.3. Community Profile Plots for Large Social and Information Networks

We have examined NCP plots for each of the networks listed in Tables 1, 2, and 3. In Figure 6, we present NCP plots for six of these networks. (These particular networks were chosen to be representative of the wide range of networks we have examined, and for ease of comparison we will compute other properties for them in future sections. See Figures 7, 8, and 9 in Section 3.4 for the NCP plots of other networks listed in Tables 1, 2, and 3, and for a discussion of them.) The most striking feature of these plots is that the NCP plot is steadily increasing for nearly its entire range.

Consider first the NCP plot for the LiveJournal01 social network, as shown in Figure 6(a), and focus first on the medium gray curve, which presents the

results of applying the local spectral algorithm.[2] We make the following observations:

- Up to a size scale, which empirically is roughly one hundred nodes, the slope of the NCP plot is generally sloping downward.
- At that size scale, we observe the global minimum of the NCP plot. This set of nodes as well as others achieving local minima of the NCP plot in the same size range are the "best" communities, according to the conductance measure, in the entire graph.
- These best communities (the best denoted by a square) are barely connected to the rest of the graph, e.g., they are typically connected to the rest of the nodes by a *single* edge.
- Above the size scale of roughly one hundred nodes, the NCP plot gradually increases over several orders of magnitude. The "best" communities in the entire graph are quite good (in that they have size roughly 10^2 nodes and conductance scores less than 10^{-3}), whereas the "best" communities of size 10^5 or 10^6 have conductance scores of about 10^{-1}. In between these two size extremes, the conductance scores get gradually worse, although there are numerous local dips and even one relatively large dip between 10^5 and 10^6 nodes.

Note that both axes in Figure 6 are logarithmic, and thus the upward trend of the NCP plot is over a wide range of size scales. Note also that the dark gray curve plots the results of Metis+MQI (which returns disconnected clusters), and the light gray curve plots the results of applying the bag-of-whiskers heuristic, as described in Section 4.3. These procedures will be discussed in detail in Sections 4 and 5.

The black curve in Figure 6(a) plots the results of the local spectral algorithm applied to a *rewired version* of the LiveJournal01 network, i.e., to a random

[2]The algorithm takes as input two parameters—the seed node and the parameter ϵ that intuitively controls the locality of the computation—and it outputs a set of nodes. For a given seed node and resolution parameter ϵ we obtain a local community profile plot, which tells us about conductance of cuts in the vicinity of the seed node. By taking the lower envelope over community profiles of different seed nodes and ϵ values, we obtain the global network community profile plot. For our experiments, we typically considered 100 different values of ϵ. Since very local random walks discover small clusters, in this case we considered every node as a seed node. As we examine larger clusters, the random walk computation spreads farther away from the seed node, in which case the exact choice of seed node becomes less important. Thus, in this case, we sampled fewer seed nodes. Additionally, in our experiments, for each value of ϵ we randomly sampled nodes until each node in the network was visited by random walks starting from 10 different seed nodes on average.

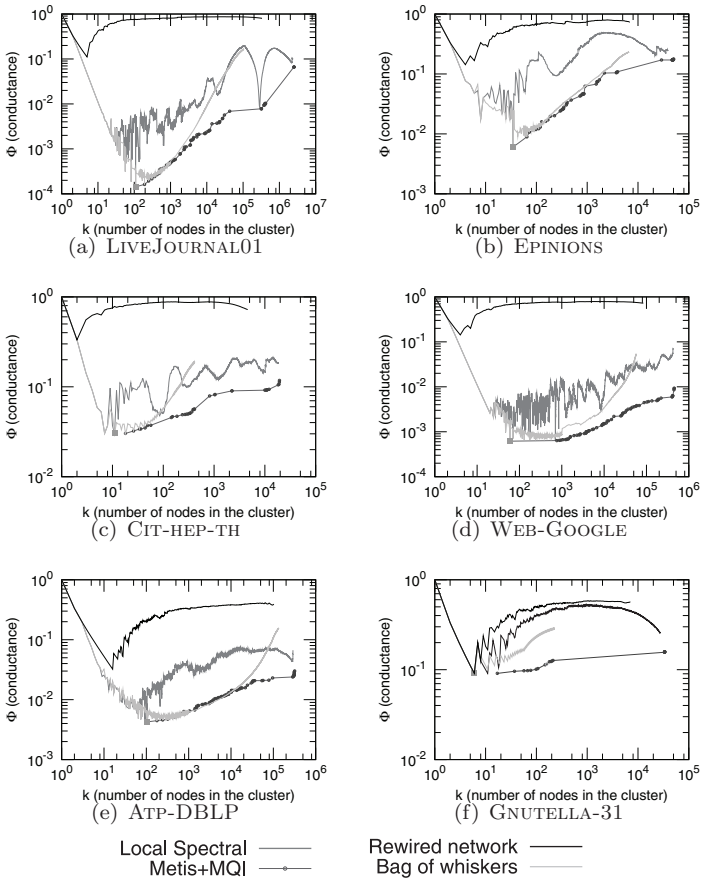


Figure 6. (Best viewed in color; see [Leskovec et al. 08a].) Network community profile plots for a representative sample of large networks listed in Tables 1, 2, and 3. The medium gray curves plot the results of the local spectral algorithm on the specified network; dark gray curves plot the results of Metis+MQI; light gray curves plot the results of the bag-of-whiskers heuristic; and black curves plot the results of the local spectral algorithm applied to a randomly rewired version of the same network. Notice that in all cases the "best" communities are quite small (typically between 10 and 100 nodes) and that the network community profile plot steadily increases for nearly its entire range. See Figures 7, 8, and 9 for the NCP plots of other networks.

graph conditioned on the same degree distribution as the original network. (We obtain such random graph by starting with the original network and then randomly selecting pairs of edges and rewiring the endpoints. By doing the rewiring

long enough, we obtain a random graph that has the same degree sequence as the original network [Milo et al. 04].)

Interestingly, the NCP of a rewired network first slightly decreases but then increases and flattens out. Several things should be noted:

- The original LiveJournal01 network has considerably more structure, i.e., deeper/better cuts, than its rewired version, even up to the largest size scales. That is, we observe significantly more structure than would be seen, for example, in a random graph on the same degree sequence.
- Relative to the original network, the "best" community in the rewired graph, i.e., the global minimum of the conductance curve, shifts upward and to the left. This means that in rewired networks the best conductance clusters get smaller and have worse conductance scores.
- Sets at and near the minimum are small trees that are connected to the core of the random graph by a single edge.
- After the small dip at a very small size scale (≈ 10 nodes), the NCP plot increases to a high level rather quickly. This is due to the absence of structure in the core.

Finally, note also that the variance in the rewired version of the NCP plot (data not shown) is not much larger than the width of the curve in the figure.

We have observed qualitatively similar results in nearly every large social and information network we have examined. For example, several additional examples are presented in Figure 6: another network from the class of social networks (Epinions, in Figure 6(b)); an information/citation network (Cit-hep-th, in Figure 6(c)); a web graph (Web-Google, in Figure 6(d)); a bipartite affiliation network (AtP-DBLP, in Figure 6(e)); and an Internet network (Gnutella-31, in Figure 6(f)).

Qualitative observations are consistent across the range of network sizes, densities, and different domains from which the networks are drawn. Of course, these six networks are very different from one another—some of these differences are hidden due to the definition of the NCP plot, whereas others are evident. Perhaps the most obvious example of the latter is that even the best cuts in Gnutella-31 are not significantly smaller or deeper than in the corresponding rewired network, whereas for Web-Google we observe cuts that are orders of magnitude deeper.

Intuitively, the upward trend in the NCP plot means that separating large clusters from the rest of the network is especially expensive. It suggests that larger and larger clusters are "blended in" more and more with the rest of the

network. The interpretation we draw, based on these data and data presented in subsequent sections is that if a density-based concept such as conductance captures our intuitive notion of community goodness and if we model large networks with interaction graphs, then the best possible communities get less and less community-like as they grow in size.

3.4. More Community Profile Plots for Large Social and Information Networks

Figures 7, 8, and 9 show additional examples of NCP plots for networks from Tables 1, 2, and 3. In the first two rows of Figure 7, we have several examples of purely social networks and two email networks; in the third row we have patent and blog information/citation networks, and in the final row we have three examples of actor and author collaboration networks. In Figure 8, we see three examples each of web graphs, Internet networks, bipartite affiliation networks, and biological networks. Finally, in the first row of Figure 9, we see low-dimensional networks, including two road networks and a manifold network; in the second row, we have an IMDB actor-to-movie graph and two subgraphs induced by restricting to individual countries; in the third row, we see three Amazon product copurchasing networks; and in the final row we see a Yahoo! Answers network and two subgraphs that are large good conductance cuts from the full network.

For most of these networks, the same four versions of the NCP plot are plotted that were presented in Figure 6. Note that as before, the scales of the vertical axes in these graphs are not all the same; the minima range from 10^{-2} to 10^{-5}. These network data sets are drawn from a wide range of areas, and these graphs contain a wealth of information, a full analysis of which is well beyond the scope of this paper. Note, however, that the general trends we discussed in Section 3.3 still manifest themselves in nearly every network.

The IMDB-RAW07 network is interesting in that its NCP plot does not increase much (at least not the version computed by the local spectral algorithm), and we clearly observe large sets with good conductance values. Upon examination, many of the large good conductance cuts seem to be associated with different language groups. Two things should be noted. First, and not surprisingly, in this network and others, we have observed that there is some sensitivity to how the data are prepared. For example, we obtain somewhat stronger communities if ambiguous nodes (and there are many ambiguous nodes in network data sets with millions of nodes) are removed than if, say, they are assigned to a country based on a voting mechanism or some other heuristic. A full analysis of these data-preparation issues is beyond the scope of this paper, but our overall conclusions seem to hold independently of the preparation details. Second, if we

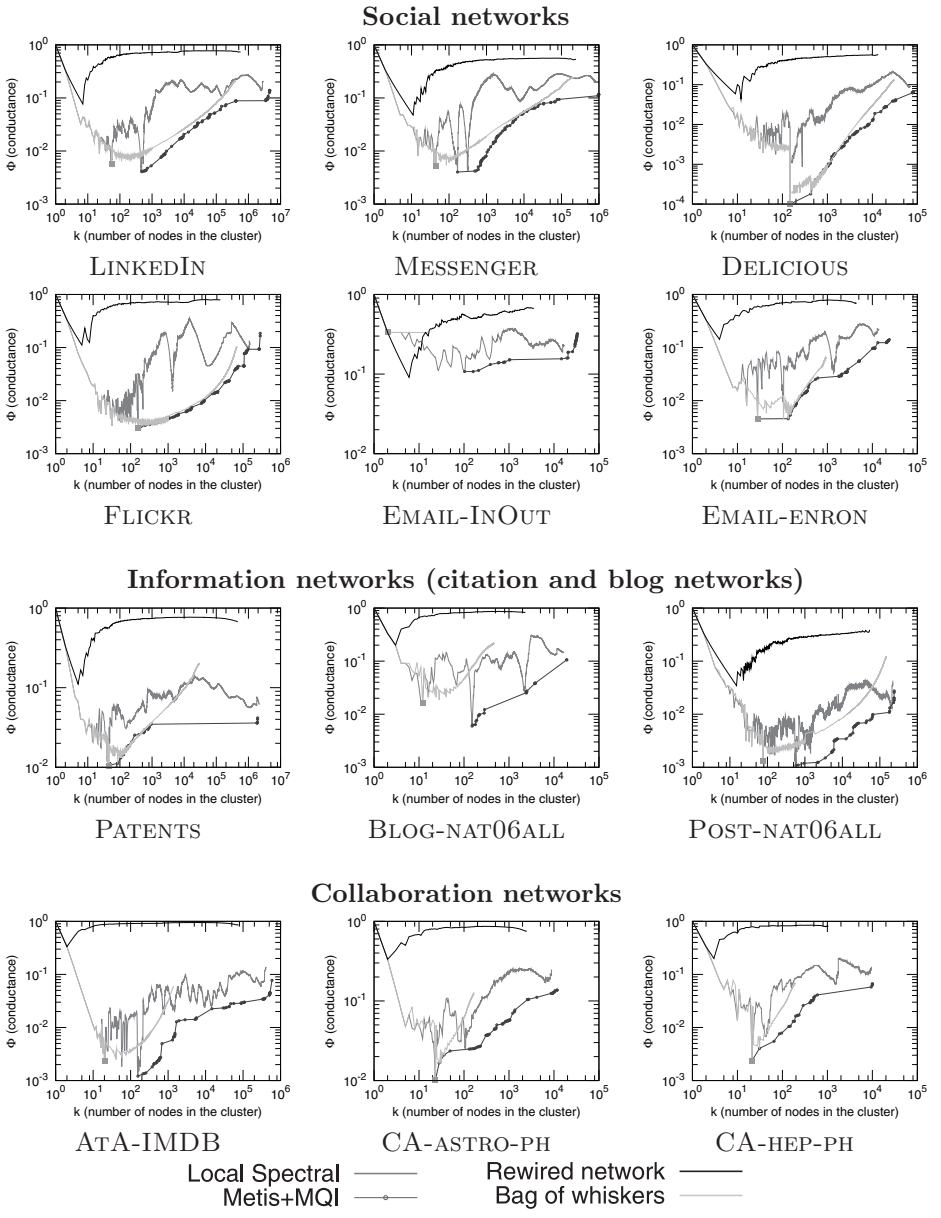


Figure 7. (Best viewed in color; see [Leskovec et al. 08a].) Community profile plots of networks from Table 1.

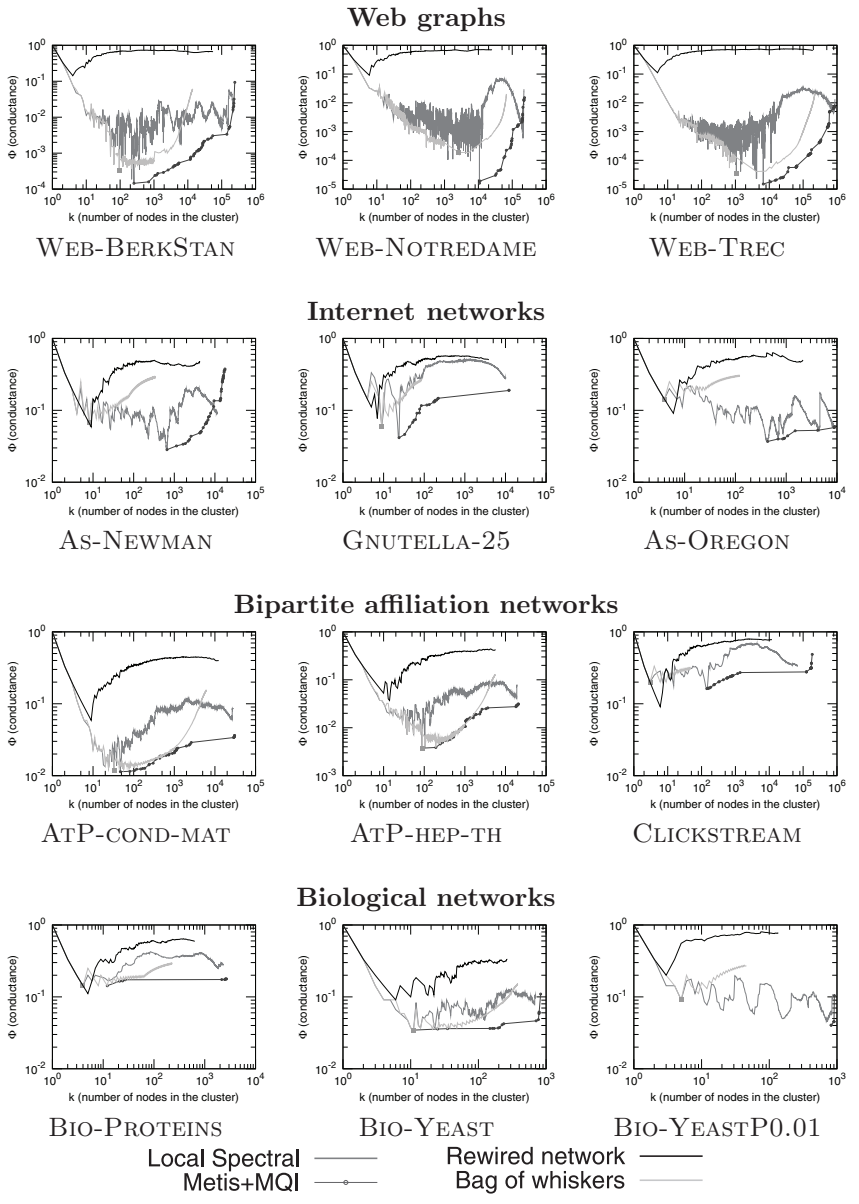


Figure 8. (Best viewed in color; see [Leskovec et al. 08a].) Community profile plots of networks from Table 2.

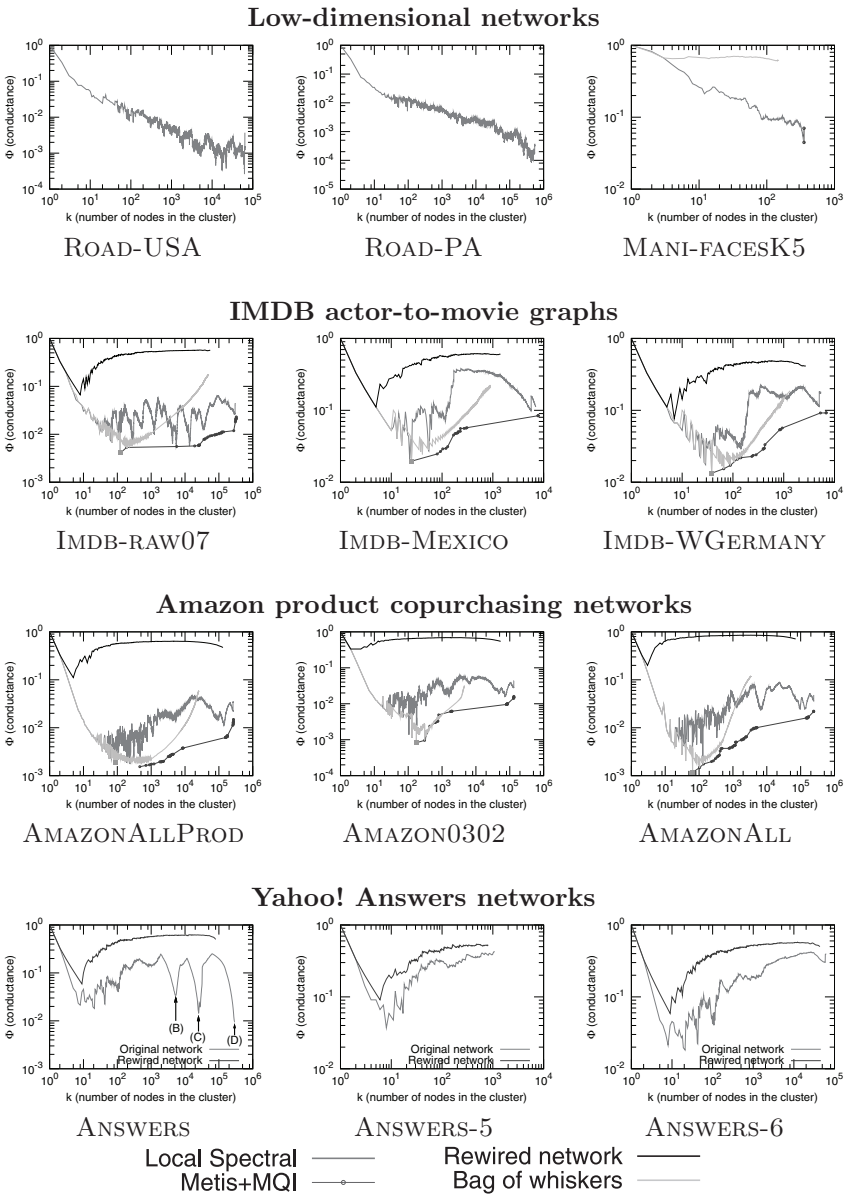


Figure 9. (Best viewed in color; see [Leskovec et al. 08a].) Community profile plots of networks from Table 3, as well as ANSWERS and two subpieces of ANSWERS.

examine individual countries—two representative examples are shown—then we see substantially less structure at large size scales.

The Yahoo! Answers social network (see ANSWERS) also has several large cuts with good conductance value—actually, the best cut in the network has more than 10^5 nodes. (It is likely that exogenous factors are responsible for these large deep cuts.) Using standard graph-partitioning procedures, we obtained four large disjoint clusters consisting of approximately 5,300, 25,400, 27,000, and 290,000 nodes, respectively, corresponding to the four dips (two of which visually overlap) in the NCP plot. We then examined the community profile plots for each of these pieces. The two representative examples that we show clearly indicate an NCP plot that is much more like other network data sets we have examined.

4. More Structural Observations of Our Network Data Sets

We have examined in greater detail our network data sets in order to understand which structural properties are responsible for the observed properties of the NCP plot. We first present statistics for our network data sets in Section 4.1. Then, in Section 4.2 we describe a heuristic to identify small sets of nodes that have strong connections among themselves but that are connected to the remainder of the network by only a single edge. In Section 4.3, we show that these "whiskers" (or disjoint unions of them) are often the "best" conductance communities in the network. Last, in Section 4.4 we examine NCP plots for networks in which these whiskers have been removed.

4.1. General Statistics on Our Network Data Sets

In Tables 1, 2, and 3, we also present the following statistics for our network data sets: the number of nodes N; the number of edges E; the fraction of nodes in the largest biconnected component N_b/N; the fraction of edges in the largest biconnected component E_b/E; the average degree $\bar{d} = 2E/N$; the empirical second-order average degree [Chung and Lu 06a] $\tilde{d}$; the average clustering coefficient [Watts and Strogatz 98] $\bar{C}$; the estimated diameter D; and the estimated average path length $\bar{D}$. (The diameter was estimated using the following algorithm: pick a random node; find the farthest node X (via shortest path); move to X and find the farthest node from X; iterate this procedure until the distance to the farthest node no longer increases. The average path length was estimated based on 10,000 randomly sampled nodes.)

In nearly every network we have examined, there is a substantial fraction of nodes that are barely connected to the main part of the network, i.e., that are

part of a small cluster of approximately 10 to 100 nodes that are attached to the remainder of the network via one or a small number of edges. In particular, a large fraction of the network consists of nodes that are not in the biconnected core.[3]

For example, the EPINIONS network has 75,877 nodes and 405,739 edges, and the core of the network has only 36,111 (47%) nodes and 365,253 (90%) edges. For DELICIOUS, the core is even smaller: it contains only 40% of the nodes, and 65% of the edges. Averaging over our network data sets, we see that the largest biconnected component contains only around 60% of the nodes and 80% of the edges of the original network. This is somewhat akin to the so-called jellyfish model [Tauro et al. 01, Siganos et al. 06] (which was proposed as a model for the graph of the Internet topology) and also to the octopus model (for random power-law graphs [Chung and Lu 06a], which is described in more detail in Section 6.2). Moreover, the global minimum of the NCP plot is nearly always one of these pieces that is connected by only a single edge. Since these small barely connected pieces seem to have a disproportionately large influence on the community structure of our network data sets, we examine them in greater detail in the next section.

4.2. Network "Whiskers" and the "Core"

We define *whiskers*, or more precisely 1-*whiskers*, to be maximal subgraphs that can be detached from the rest of the network by removing a *single* edge. (Occasionally, we use the term whiskers informally to refer to barely connected sets of nodes more generally.) To find 1-whiskers, we employ the following algorithm. Using a depth-first search algorithm, we find the largest biconnected component B of the graph G. (A graph is biconnected if the removal of any single edge does not disconnect the graph.) We then delete all the edges in G that have one of their endpoints in B. We call the connected components of this new graph G' 1-whiskers, since they correspond to maximal subgraphs that can be disconnected from G by removing just a single edge. Recall that Figure 2(b) contains a schematic picture a network, including several of its whiskers.

Not surprisingly, there is a wide range of whisker sizes and shapes. Figure 10 shows the distribution of 1-whisker sizes for a representative selection of our network data sets. Empirically, 1-whisker size distribution is heavy-tailed, with the largest whisker size ranging from around less than 10 to well above 100. The largest whiskers in coauthorship and citation networks have around 10 nodes,

[3]In this paper, we are slightly abusing standard terminology by using the term biconnectivity to mean 2-edge connectivity. We *are* running the classic DFS-based biconnectivity algorithm, which identifies both bridge edges and articulation nodes, but then we are knocking out only the bridge edges, not the articulation nodes, so we end up with 2-edge-connected pieces.

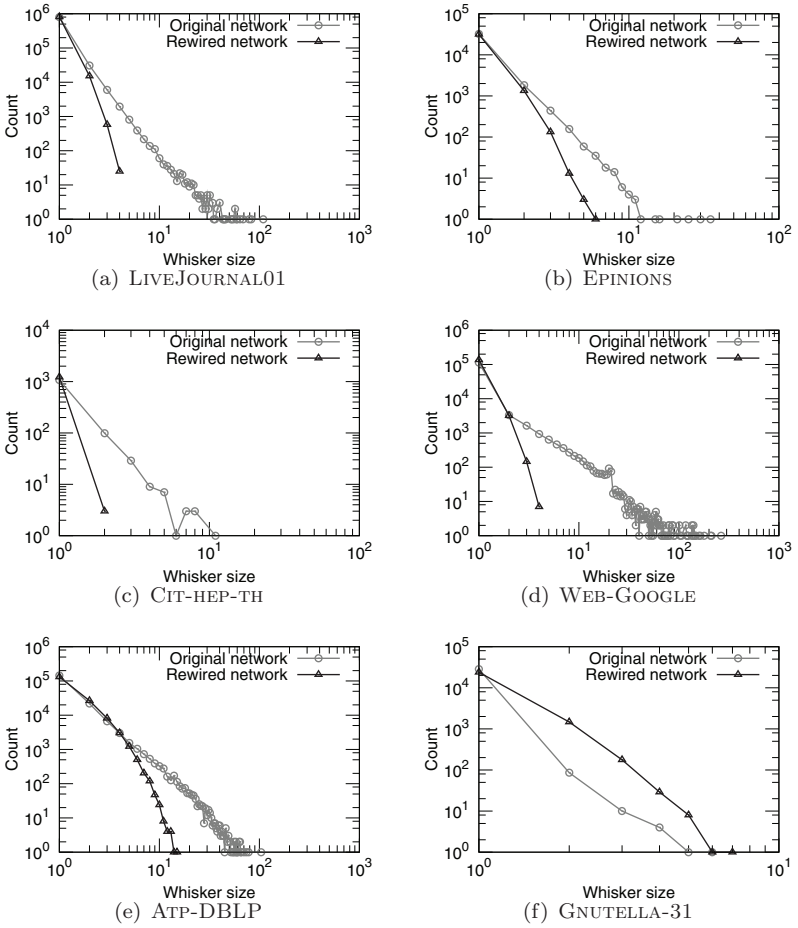


Figure 10. Distribution of whisker sizes in the true network and the rewired network (random graph with same degree distribution) for the six networks presented in Figure 6. The ten largest whiskers for the Epinions social network (the full distribution of which is presented here in panel (b)) are presented in Figure 11.

whiskers in bipartite graphs also tend to be small, and very large whiskers are found in a web graph. Figure 10 also compares the size of the whiskers with the sizes of whiskers in a rewired version of the same network. (The first thing to note is that due to the sparsity of the networks, the rewired versions all have whiskers.) In rewired networks the whiskers tend to be much smaller than in the original network. A particularly noteworthy exception is found in the

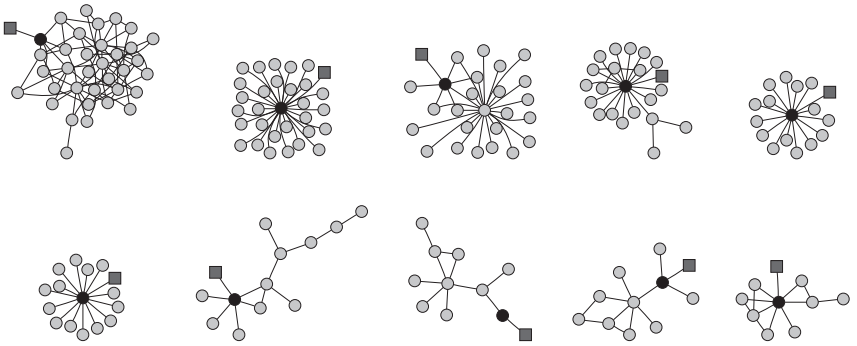


Figure 11. Ten largest whiskers of the EPINIONS social network. The dark gray square node is the node from the biconnected core of the network to which the whisker is connected. For visual clarity, the whisker node that connects to the core of the network is displayed in black, and thus it is the edge between the black circle and the dark gray square node that if cut disconnects the whisker from the core. The distribution of whisker sizes and comparison to a rewired network are plotted in Figure 10(b).

autonomous systems networks and the GNUTELLA-31 network. (See Figure 10(f) for an example of the latter.) In these cases, the whiskers are so small that even the rewired version of the network has more and larger whiskers. This makes sense, given how those networks were designed: clearly, many large whiskers would have negative effects on the Internet connectivity in case of link failures.

Figure 11 shows the ten largest whiskers of the EPINIONS social network, the full size distribution of which was plotted in Figure 10(b); and Figure 12 shows the ten largest whiskers of the CA-COND-MAT coauthorship network. In these networks, the whiskers have on the order of 10 nodes, and they are seen to have a rich internal structure. Similar but substantially more complex figures could be generated for networks with larger whiskers. In general, the results we observe are consistent with a knowledge of the fields from which the particular data sets have been drawn. For example, in WEB-GOOGLE we see very large whiskers. This probably represents a well-connected network between the main categories of a website (e.g., different projects), while the individual project websites have a main index page that then points to the rest of the documents.

The discrepancy between the sizes of the whiskers in the original and the rewired networks gives hints that real networks have much richer structure than that imposed by their heavy-tailed degree distribution. One might ask whether the conclusion from this is that real-world graphs should be thought of as being somewhat like sparse random graphs, since, for example, both have whiskers,

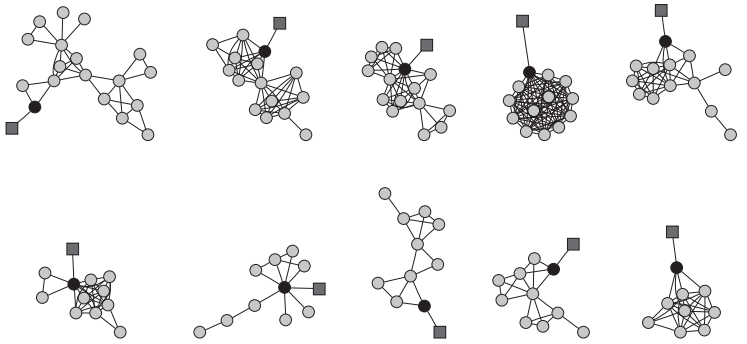


Figure 12. Ten largest whiskers of the CA-COND-MAT coauthorship network. The dark gray square node belongs to the network core, and by cutting the edge connecting it with the black circular node we separate the community of circles from the rest of the network (depicted as a dark gray square).

or should be thought of as very different from sparse random graphs, since, for example, the whiskers have much more internal structure. We will return to this issue in Section 6.

4.3. Bags of Whiskers and Communities of Composed Whiskers

Empirically, if one looks at the sets of nodes achieving the minimum in the NCP plot (dark gray Metis+MQI curve), then below the global NCP minimum communities are whiskers and above that size scale they are often unions of disjoint whiskers. To understand the extent to which these whiskers and unions of whiskers are responsible for the "best" conductance sets of different sizes, we have developed the *bag-of-whiskers heuristic*. We artificially compose "communities" from disconnected whiskers and measure conductance of such clusters. Clearly, interpreting and relating such communities to real-world communities makes little sense, since these communities are in fact disconnected.

In more detail, we performed the following experiment: Suppose we have a set $W = \{w_1, w_2, \dots\}$ of whiskers. In order to construct the optimal conductance cluster of size k, we need to solve the following problem: find a set C of whiskers such that $\sum_{i \in C} N(w_i) = k$ and $\sum_{i \in C} \frac{d(w_i)}{|C|}$ is maximized, where $N(w_i)$ is the number of nodes in w_i and $d(w_i)$ is its total internal degree. We then use dynamic programming to obtain an approximate solution to this problem.

In this way, for each size k, we find a cluster that is composed solely of (disconnected) whiskers. Figure 6, as well as Figures 7, 8, and 9, shows the results of this heuristic applied to many of our network data sets (light gray curve).

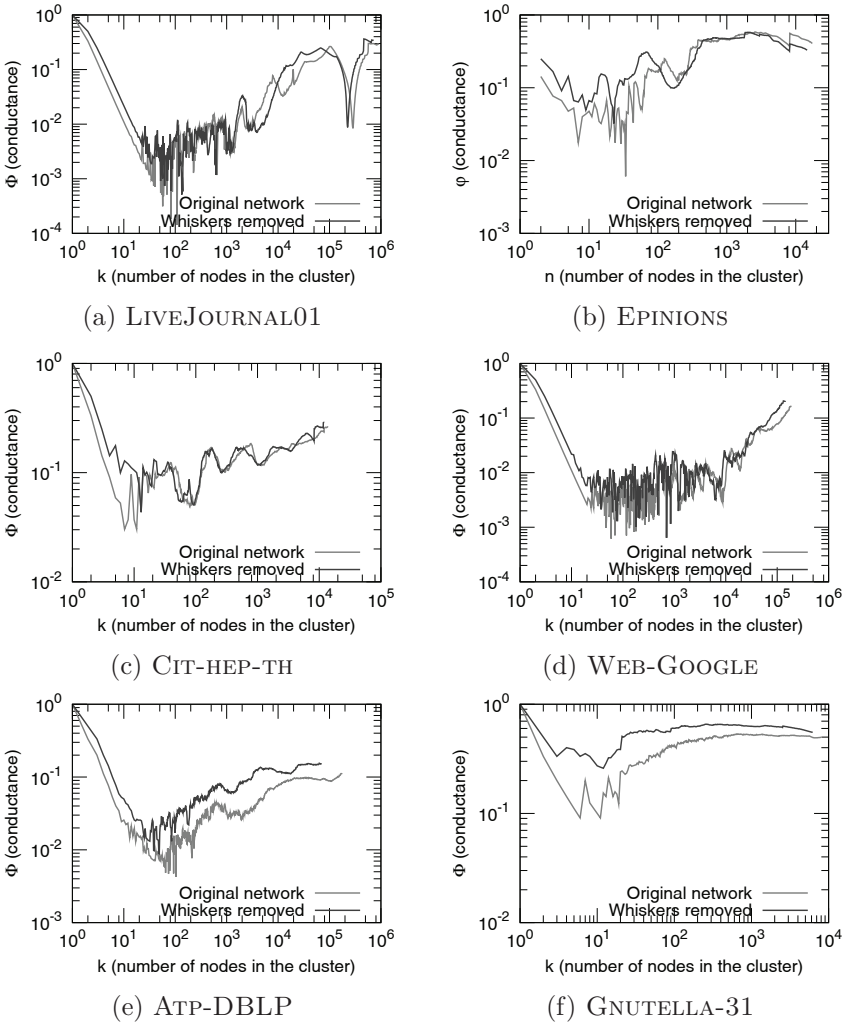
There are several observations we can make:

- The largest whisker (denoted by a square) is the lowest point in nearly all NCP plots. This means that the best conductance community is in a sense trivial, since it cuts just a single edge, and in addition that a very simple heuristic can find this set.
- For community size below the critical size of ≈ 100 nodes (i.e., of size smaller than the largest whisker), the best community in the network is actually a whisker and can be cut by a single edge (light gray and medium gray curve overlap).
- For community size larger than the critical size of ≈ 100, the bag-of-whiskers communities have better scores than the internally well connected communities extracted by the local spectral algorithm (medium gray curve). The shape of this light gray curve in that size region depends on the distribution of sizes of whiskers, but in nearly every case it is seen to yield better conductance sets than the local spectral algorithm.

Moreover, the bag-of-whiskers heuristic often agrees, exactly or approximately, with results from Metis+MQI (dark gray curve). In particular, the best conductance sets of a given size are often disconnected, and when they are connected they are often only tenuously connected. Thus, if one cares only about finding good cuts, then the best cuts in these large sparse graphs are obtained by composing unrelated disconnected pieces. Intuitively, a compact cluster is internally well and evenly connected. Possible measures for cluster compactness include cluster connectedness, diameter, conductance of the cut inside the cluster, and ratio of conductance of the cut outside versus the cut inside. We discuss this in more detail in Section 5.

4.4. Community Profile of Networks with No 1-Whiskers

Given the surprisingly significant effect on the community structure of real-world networks that whiskers and unions of disjoint whiskers have, one might wonder whether we see something qualitatively different if we consider a real-world network in which these barely connected pieces have been removed. To study this, we found all 1-whiskers and removed them from our networks, using the procedure we described in Section 4.2, i.e., we selected the largest biconnected component for each of our network data sets. In this way, we kept only the network core, and we then computed the NCP plots for these modified networks. Figure 13 shows the NCP plots of networks constructed when we remove whiskers (i.e., keeping only the network core) for the six networks we studied in detail before.



(a) LiveJournal01 (b) Epinions

(c) Cit-hep-th (d) Web-Google

(e) Atp-DBLP (f) Gnutella-31

Figure 13. (Best viewed in color; see [Leskovec et al. 08a].) Network community profile plots with (in medium gray) and without (in dark gray) 1-whiskers, for each of the six networks shown in Figure 6. Whiskers were removed as described in the text. In the former case, we plot results for the full network, and in the latter case, we plot results for the largest biconnected component.

Notice that whisker removal does not change the NCP plot much: the plot shifts slightly upward, but the general trends remain the same. On examination, we see that the global minimum occurs with a "2-whisker" that is connected by

two edges to the remainder of the graph. Intuitively, the largest biconnected core has a large number of barely connected pieces—connected now by two edges rather than by one edge—and thus the "core" itself has a core–periphery structure. Since the "volume" of these pieces is similar to that of the original whiskers, whereas the "surface area" is a factor of two larger, the conductance value is roughly a factor of two worse. Thus, although we have been discussing 1-whiskers in this section, one should really view them as the simplest example of weakly connected pieces that exert a significant effect on the community structure in large real-world networks.

5. Comparison to Other Algorithms

So far, we have been primarily relying on two graph-partitioning algorithms: a local spectral algorithm and Metis+MQI. Next, we want to demonstrate that what we are observing is a true structural property of our network data sets, rather than properties of our algorithms; and we want to use the differences between different approximation algorithms to further highlight structural properties of our network data sets. In this section we discuss several meta-issues related to this, including whether our algorithms are sufficiently powerful to recover the true shape of the minimal conductance curves, and whether we should actually be trying to optimize a slightly different measure that combines conductance of the separating cut with the piece compactness.

Recall that we defined the NCP plot to be a curve showing the minimum conductance ϕ as a function of piece size k. Finding the points on this curve is NP-hard. Any cut that we find will provide only an upper bound on the true minimum at the resulting piece's size. Given that fact, how confident can we be that the curve of upper bounds that we have computed has the same rising or falling shape as the true curve?

One method for finding out whether any given algorithm is doing a good job of pushing down the upper bounding curve in a non-size-biased way is to compare its curves for numerous graphs with those produced by other algorithms. In such experiments, it is good if the algorithms are very powerful and also independent of each other. We have done extensive experiments along these lines, and our choice of local spectral and Metis+MQI as the two algorithms for the main body of this paper was based on the results. In Section 5.1 we mention a few interesting points related to this.

A different method for reducing our uncertainty about the shape of the true curve would be also to compute lower bounds on the curve. Ideally, one would compute a complete curve of tight lower bounds, leaving a thin band between the

upper- and lower-bounding curves, which would make the rising or falling shape of the true curve obvious. In Section 5.2 we discuss some experiments with lower bounds. Although we obtained only a few lower bounds rather than a full curve, the results are consistent with our main results obtained from upper-bounding curves.

Finally, in Section 5.3 we will discuss our decision to use the local spectral algorithm in addition to Metis+MQI in the main body of the paper, despite the fact that Metis+MQI clearly dominates local spectral at the nominal task of finding the lowest possible upper-bounding curve for the minimal conductance curve. The reason for this decision is that local spectral often returns "nicer" and more "compact" pieces, because rather than minimizing conductance alone, it optimizes a slightly different measure that produces a compromise between the conductance of the bounding cut and the "compactness" of the resulting piece.

5.1. Cross-Checking between Algorithms

As just mentioned, one way to gain some confidence in the upper-bounding curves produced by a given algorithm is to compare them with the curves produced by other algorithms that are as strong as possible, and as independent as possible.

We have extensively experimented with several variants of the global spectral method, both the usual eigenvector-based embedding on a line, and an SDP-based embedding on a hypersphere, both with the usual hyperplane-sweep rounding method and a fancier flow-based rounding method that includes MQI as the last step. In addition, special postprocessing can be done to obtain either connected or disconnected sets. After examining the output of those eight comparatively expensive algorithms on more than one hundred graphs, we found that our two cheaper main algorithms did miss an occasional cut on an occasional graph, but nothing at all serious enough to change our main conclusions. All of those detailed results are suppressed in this paper.

We have also done experiments with a practical version of the Leighton–Rao algorithm [Leighton and Rao 88, Leighton and Rao 99], similar to the implementation described in [Lang and Rao 93] and [Lang and Rao 04]. These results are especially interesting because the Leighton–Rao algorithm, which is based on multicommodity flow, provides a completely independent check on Metis, and on spectral methods generally, and therefore on our two main algorithms, namely Metis+MQI and local spectral.

The Leighton–Rao algorithm has two phases. In the first phase, edge congestions are produced by routing a large number of commodities through the network. We adapted our program to optimize conductance (rather than ordinary ratio cut score) by letting the expected demand between a pair of nodes be

proportional to the product of their degrees. In the second phase, a rounding algorithm is used to convert edge congestions into actual cuts. Our method was to sweep over node orderings produced by running Prim's MST algorithm on the congestion graph, starting from a large number of different initial nodes, using a range of different scales to avoid quadratic run time. We used two variations of this method, one that produces only connected sets, and another one that can also produce disconnected sets.

In the second row of Figure 14, we show Leighton–Rao curves for three example graphs. Our standard local spectral and Metis+MQI curves are drawn in black, while the Leighton–Rao curves for connected and possibly disconnected sets are drawn in light gray and medium gray respectively. We note that for small to medium scales, the Leighton–Rao curves for connected sets resemble the local spectral curves, while the Leighton–Rao curves for possibly disconnected sets resemble the Metis+MQI curves. This is a big hint about the structure of the sets produced by local spectral and Metis+MQI, which we will discuss further in Section 5.3.

At large scales, the Leighton–Rao curves for these example graphs shoot up and become much worse than our standard curves. This is not surprising, because expander graphs are known to be the worst-case input for the Leighton–Rao approximation guarantee, and we believe that these graphs contain an expander-like core that is necessarily encountered at large scales.

We remark that Leighton–Rao does not work poorly at large scales on every kind of graph. (In fact, for large low-dimensional meshlike graphs, Leighton–Rao is a very cheap and effective method for finding cuts at all scales, while our local spectral method becomes impractically slow at medium to large scales. We will not discuss this point further, except to note that in the main body of the paper we have silently substituted Leighton–Rao curves for local spectral curves for the large road networks and similar graphs.)

We have now covered the main theoretical algorithms that are practical enough to actually run, which are based on spectral embeddings and on multicommodity flow. Starting with [Arora et al. 04b], there has been a recent burst of theoretical activity showing that spectral and flow-based ideas, which were already known to have complementary strengths and weaknesses, can in fact be combined to obtain the best ever approximations. At present none of the resulting algorithms are sufficiently practical at the sizes that we require, so they were not included in this study.

Finally, we mention that in addition to the above theoretically based practical methods for finding low-conductance cuts, there is a very large number of heuristic graph-clustering methods. We have tried a number of them, including Graclus [Dhillon et al. 07] and Newman's modularity optimizing program (we refer to it

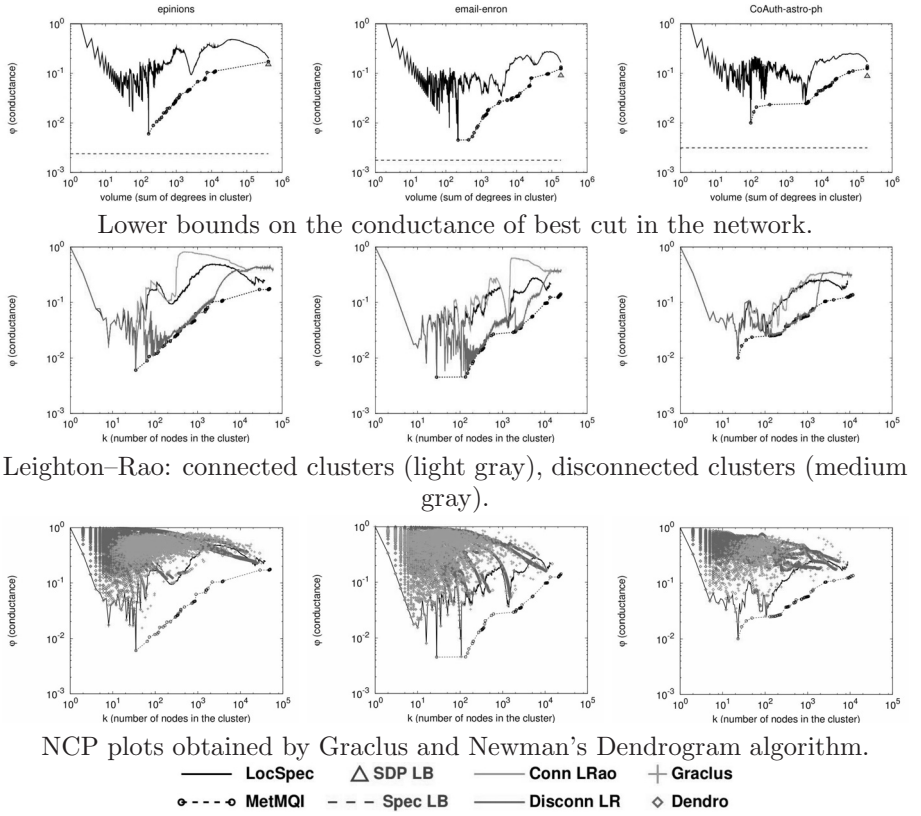


Figure 14. (Best viewed in color; see [Leskovec et al. 08a].) Result of other algorithms for three networks: EPINIONS, EMAIL-ENRON, and CA-ASTRO-PH. Top row plots (in black) conductance curves as obtained by local spectral and Metis+MQI. Top row also shows lower bounds on conductance of any cut (spectral lower bound, dashed line) and the cut separating the graph in half (SDP lower bound, triangle). Middle row shows NCP plots for connected (light gray) and disconnected (medium gray) pieces from our implementation of the Leighton–Rao algorithm. Bottom row shows the conductance of some cuts found by Graclus and by Newman's dendrogram algorithm. The overall conclusion is that the qualitative shape of the NCP plots is a structural property of large networks, and the plot remains practically unchanged regardless of what particular community-detection algorithm we use.

as dendrogram) [Girvan and Newman 02]. Graclus attempts to find a partitioning of a graph into pieces bounded by low-conductance cuts using a kernel k-means algorithm. We ran Graclus repeatedly, asking for $2, 3, \ldots, i, \ldots, i*\sqrt{2}, \ldots$ pieces. Then we measured the size and conductance of all of the resulting pieces.

Newman's dendrogram program constructs a recursive partitioning of a graph (that is, a dendrogram) from the bottom up by repeatedly deleting the surviving edge with the highest betweenness centrality. A flat partitioning could then be obtained by cutting at the level that gives the highest modularity score, but instead of doing that, we measured the size of conductance of every piece defined by a subtree in the dendrogram.

In the bottom row of Figure 14, we present these results as scatter plots. Again our two standard curves are drawn in black. No Graclus or dendrogram point lies below the Metis+MQI curve. The lower envelopes of the points are roughly similar to those produced by local spectral.

Our main point with these experiments is that the lowest points produced by either Graclus or dendrogram gradually rise as one moves from small scales to larger scales, so in principle we could have made the same observations about the structure of large social and information networks by running one of those easily downloadable programs instead of the algorithms that we did run. We chose the algorithms we did due to their speed and power, although they may not be as familiar to many readers.

5.2. Lower Bounds on Cut Conductance

As mentioned above, our main arguments are all based on curves that are actually upper bounds on the true minimum-conductance curve. To get a better idea of how good those upper bounds are, we also compute some lower bounds. Here we will discuss the spectral lower bound [Chung 97] on the conductance of cuts of arbitrary balance, and we will also discuss a related SDP-based lower bound [Burer and Monteiro 03] on the conductance of any cut that divides the graph into two pieces of equal volume.

First, we introduce the following notation: $\vec{d}$ is a column vector of the graph's node degrees; D is a square matrix whose only nonzero entries are the graph's node degrees on the diagonal; A is the adjacency matrix of G; $L = D - A$ is then the unnormalized Laplacian matrix of G; $\mathbf{1}$ is a vector of 1's; and $A \bullet B = \text{trace}(A^T B)$ is the matrix dot-product operator.

Now consider the following optimization problem (which is well known to be equivalent to an eigenproblem):

$$\lambda_G = \min \left\{ \frac{x^T L x}{x^T D x} : x \perp \vec{d}, x \neq 0 \right\}.$$

Let $\hat{x}$ be a vector achieving the minimum value λ_G. Then $\lambda_G/2$ is the spectral lower bound on the conductance of any cut in the graph, regardless of balance, while $\hat{x}$ defines a spectral embedding of the graph on a line, to which rounding

algorithms can be applied to obtain actual cuts that can serve as upper bounds at various sizes.

Next, we discuss an SDP-based lower bound on cuts that partition the graph into two sets of exactly equal volume. Consider

$$\mathcal{C}_G = \min \left\{ \frac{1}{4} L \bullet Y : \mathrm{diag}(Y) = \mathbf{1}, Y \bullet (\vec{d}\,\vec{d}^T) = 0, Y \succeq 0 \right\},$$

and let $\hat{Y}$ be a matrix achieving the minimum value $\mathcal{C}_G$. Then $\mathcal{C}_G$ is a lower bound on the weight of any cut with perfect volume balance, and $2\mathcal{C}_G/\mathrm{Vol}(G)$ is a lower bound on the conductance of any cut with perfect volume balance. We briefly mention that since $Y \succeq 0$, we can view Y as a Gram matrix that can be factored as RR^T. Then the rows of R are the coordinates of an embedding of the graph on a hypersphere. Again, rounding algorithms can be applied to the embedding to obtain actual cuts that can serve as upper bounds.

The spectral and SDP embeddings defined here were the basis for the extensive experiments with global spectral partitioning methods that were alluded to in Section 5.1. However, in this section, it is the lower bounds that concern us.

In the top row of Figure 14, we present the spectral and SDP lower bounds for three example graphs. The spectral lower bound, which applies to cuts of any balance, is drawn as a horizontal line that appears near the bottom of each plot. The SDP lower bound, which applies only to cuts separating a specific volume, namely $\mathrm{Vol}(G)/2$, appears as an upward-pointing triangle near the right side of each plot. (Note that plotting this point required us to use volume rather than number of nodes for the x-axis of these three plots.)

Clearly, for these graphs, the lower bound at $\mathrm{Vol}(G)/2$ is higher than the spectral lower bound that applies at smaller scales. More importantly, the lower bound at $\mathrm{Vol}(G)/2$ is higher than our *upper* bounds at many smaller scales, so the true curve must go up, at least at the very end, as one moves from small to large scales.

Take, for example, the top left plot of Figure 14, where in black we plot the conductance curves obtained by our (local spectral and Metis+MQI) algorithms. With a dashed line we also plot the lower bound of the best possible cut in the network, and with a triangle we plot the lower bound for the cut that separates the graph into two equal-volume parts. Thus, the true conductance curve (which is intractable to compute) lies below the black but above the dashed line and triangle.

This also demonstrates that the conductance curve that starts at the upper left corner of the NCP plot first goes down and reaches the minimum close to the horizontal dashed line (spectral lower bound) and then sharply rises and ends up above the triangle (SDP lower bound). This verifies that our conductance

curves and obtained NCP plots are not the artifacts of the community-detection algorithms we employed.

Finally, in Table 4 we list for about 40 graphs the spectral and SDP lower bounds on overall conductance and on volume-bisecting conductance, and also the ratio between the two. It is interesting to see that for these graphs this ratio of lower bounds does a fairly good job of discriminating between falling-NCP-plot graphs, which have a small ratio, and rising-NCP-plot graphs, which have a large ratio. Small networks (such as CollegeFootball, ZacharyKarate, and MonksNetwork) have a downward NCP plot and a small ratio of the SDP and spectral lower bounds. On the other hand, large networks (e.g., Epinions and Answers-3) that have downward and then upward NCP plots (as in Figure 2(a)) have a large ratio of the two lower bounds. This is further evidence that small networks have fundamentally different community structure from large networks and that one has to examine very large networks to observe the gradual absence of communities of size above ≈ 100 nodes.

5.3. Local Spectral and Metis+MQI

In this section we discuss our rationale for using local spectral in addition to Metis+MQI as one of our two main algorithms for finding sets bounded by low conductance cuts. This choice requires some justification because the NCP plots are intended to show the tightest possible upper bound on the lowest conductance cut for each piece size, while the curve for local spectral is generally above that for Metis+MQI.

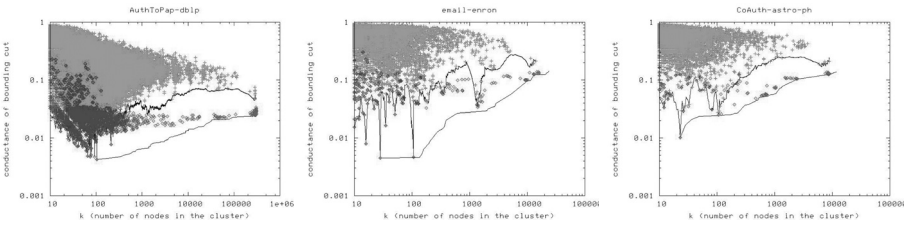
Our reason for using local spectral in addition to Metis+MQI is that local spectral returns pieces that are internally "nicer." For graphs with a rising NCP plot, we have found that many of the low-conductance sets returned by Metis+MQI (or Leighton–Rao, or the bag-of-whiskers heuristic) are actually *disconnected*. Since internally disconnected sets are not very satisfying "communities," it is natural to wonder about NCP plot-style curves with the additional requirement that pieces be internally well connected. In Section 5.1, we generated such a curve using Leighton–Rao, and found that the curve corresponding to connected pieces was higher than a curve allowing disconnected sets.

In the top row of Figure 15, we show scatter plots illustrating a similar comparison between the conductance of the cuts bounding connected pieces generated by local spectral and by Metis+MQI. Our method for getting connected pieces from Metis+MQI here is simply to measure each of the pieces in a disconnected set separately. The light gray points in the figures show the conductance of some cuts found by local spectral. The dark gray points show the conductance of some cuts found by Metis+MQI. Apparently, local spectral and Metis+MQI

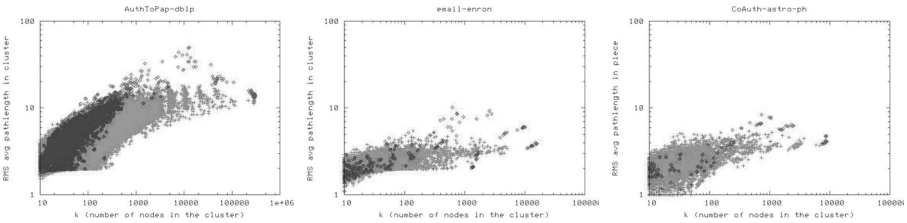
Network	Spectral lower bound on ϕ, any siz	SDP lower bound on ϕ, at Vol(G)/2	ratio of lower bounds
CollegeFootball*	0.068402	0.091017	1.330624
MonksNetwork*	0.069660	0.117117	1.681269
ZacharyKarate*	0.066136	0.127625	1.929736
PowerGrid	0.000136	0.000268	1.978484
PoliticalBooks*	0.018902	0.038031	2.011991
PoliticalBlogs*	0.040720	0.084052	2.064157
RB-Hierarchical†	0.011930	0.030335	2.542792
Email-InOut	0.038669	0.113367	2.931752
NetworkScience*	0.001513	0.004502	2.974695
As-Oregon	0.012543	0.042976	3.426417
Blog-nat05-6m	0.031604	0.108979	3.448250
Imdb-India	0.009104	0.033318	3.659573
Cit-hep-ph	0.007858	0.029243	3.721553
Bio-Proteins	0.033714	0.126137	3.741358
As-RouteViews	0.018681	0.070462	3.771821
Gnutella-31	0.029946	0.118711	3.964127
Imdb-Japan	0.003327	0.013396	4.026721
Gnutella-30	0.030621	0.124929	4.079853
DolphinsNetwork*	0.019762	0.103676	5.246171
As-Newman	0.009681	0.058952	6.089191
AtP-gr-qc	0.000846	0.006040	7.141270
Cit-hep-th	0.009193	0.068880	7.492522
AtP-cond-mat	0.001703	0.013452	7.897650
Gnutella-25	0.014185	0.131032	9.237332
Answers-2	0.009660	0.107422	11.120081
CA-cond-mat	0.003593	0.047064	13.098027
Answers-1	0.011896	0.159251	13.386528
Imdb-France	0.003462	0.048010	13.867591
Answers-5	0.008714	0.124703	14.311255
Imdb-Mexico	0.003893	0.070345	18.067513
CA-gr-qc	0.000934	0.017421	18.659710
AtP-hep-th	0.000514	0.009714	18.899660
AtP-hep-ph	0.000723	0.013770	19.040287
Imdb-WGermany	0.003025	0.065158	21.538867
AtP-astro-ph	0.001183	0.027256	23.036835
CA-hep-th	0.001561	0.041125	26.350412
CA-astro-ph	0.003143	0.086890	27.648094
Imdb-UK	0.001283	0.036572	28.514376
Imdb-Germany	0.000661	0.021017	31.810460
Blog-nat06all	0.002361	0.092908	39.350874
Imdb-Italy	0.000679	0.031954	47.077242
Email-Enron	0.001763	0.089876	50.965424
CA-hep-ph	0.000889	0.052249	58.755927
Epinions	0.002395	0.150242	62.739252
Answers-3	0.002636	0.185340	70.306807
Imdb-Spain	0.000562	0.046327	82.397702

* [Newman 09] † [Ravasz and Barabási 03]

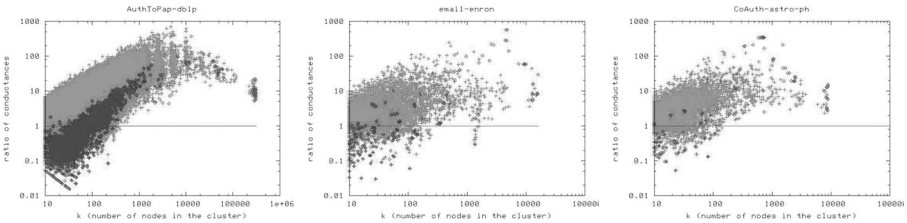
Table 4. Lower bounds on the conductance for our network data sets. Recall that the spectral lower bound applies to any cut, while the SDP lower bound applies to cuts at a specified volume fraction, taken here to be half. See the top row of Figure 14 for plots for three of these networks.



Conductance of connected clusters found by local spectral (light gray) and Metis+MQI (dark gray)



Cluster compactness: average shortest path length



Cluster compactness: external vs. internal conductance

Figure 15. (Best viewed in color; see [Leskovec et al. 08a].) Result of comparing local spectral (light gray) and Metis+MQI (dark gray) on connected clusters for three networks: AtP-DBLP, Email-Enron, and CA-astro-ph. In the top row, we plot the conductance of the bounding cut; in the middle row, the average shortest path length in the cluster; and in the bottom row, the ratio of the external conductance to the internal conductance. Observe that generally Metis+MQI yields better (lower conductance) cuts, while local spectral yields pieces that are more compact: they have shorter path lengths and internal connectivity.

find similar pieces at very small scales, but at slightly larger scales a gap opens up between the dark gray cloud and the light gray cloud. In other words, at those scales Metis+MQI is finding lower conductance cuts than local spectral, even when the pieces must be internally connected.

However, there is still a measurable sense in which the local spectral pieces are "nicer" and more "compact," as shown in the second row of scatter plots in Figure 15. For each of the same pieces for which we plotted a conductance in the

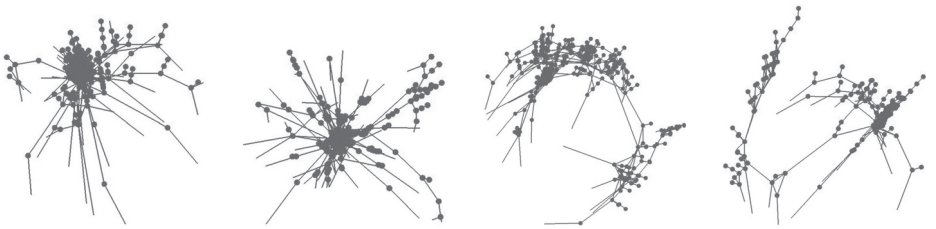


Figure 16. Two examples of "communities" found by the local spectral algorithm (on the left) and two from the Metis+MQI algorithm (on the right). Note that the local spectral "communities" are more compact—they are tighter and have smaller diameter, since the algorithm has difficulty pushing probability mass down long extended paths—while the Metis+MQI "communities" are more sprawling—they have larger diameter and more diverse internal structure, but better conductance scores. In both cases, we have shown communities with approximately 500 nodes (many of which overlap at the resolution of this figure), i.e., just above the "whisker" size scale.

top row, we are now plotting the average shortest path length between random node pairs in that piece. In these plots, we see that in the same size range where Metis+MQI is generating clearly lower conductance connected sets, we now see that local spectral is generating pieces with clearly shorter internal paths. In other words, the local spectral pieces are more "compact."

Last, in Figure 16, we further illustrate this point with drawings of some example subgraphs. The two subgraphs shown on the left of Figure 16 were found by local spectral, while the two subgraphs shown on the right of Figure 16 were found by Metis+MQI. Clearly, these two pairs of subgraphs have a qualitatively different appearance, with the Metis+MQI pieces looking longer and stringier than the local spectral pieces. All of these subgraphs contain roughly 500 nodes, which is a bit more than the natural cluster size for that graph, and thus the differences between the algorithms start to show up. In these cases, local spectral has grown a cluster out a bit past its natural boundaries (hence the spokes), while Metis+MQI has strung together a couple of different sparsely connected clusters. (We remark that the tendency of local spectral to trade off cut quality in favor of piece compactness isn't just an empirical observation, it is a well understood consequence of the theoretical analysis of spectral partitioning methods.)

Finally, in the bottom row of Figure 15 we briefly introduce the topic of internal vs. external cuts, which is something that none of our algorithms are explicitly trying to optimize. These are again scatter plots showing the same set of local spectral and Metis+MQI pieces as before, but now the y-axis is external conductance divided by internal conductance. External conductance

is the quantity that we usually plot, namely the conductance of the cut that separates the piece from the graph. Internal conductance is the score of a low-conductance cut *inside* the piece (that is, in the induced subgraph on the piece's nodes). Intuitively, good communities should have small ratios, ideally below 1.0, which would mean that they are well separated from the rest of the network but that they are internally well connected. However, the three bottom-row plots show that for these three sample graphs, there are mostly no ratios well below 1.0 except at small sizes. (Of course, any given graph could happen to contain a very distinct piece of any size, and the roughly thousand-node piece in the EMAIL-ENRON network is a good example.)

This demonstrates another aspect of our findings: small communities of size below ≈ 100 nodes are internally compact and well separated from the remainder of the network, whereas larger pieces are so hard to separate that separating them from the network is more expensive than separating them internally.

6. Models for Network Community Structure

In this section, we use results from previous sections to devise a model that explains the shape of NCP plots. In Section 6.1, we examine the NCP plot for a wide range of existing commonly used network-generation models, and we see that none of them reproduces the observed properties, at even a qualitative level. Then, in Section 6.2, we analytically demonstrate that certain aspects of the NCP plot, e.g., the existence of deep cuts at small size scales, can be explained by very sparse random graph models. Then, in Section 6.3, we present a simple toy model to develop intuition about the effect we must reproduce with a realistic generative model. Finally, in Section 6.4, we will combine these and other ideas to describe a forest fire graph generation model that reproduces quite well our main observations.

6.1. Community Profile Plots for Commonly Used Network-Generation Models

We have studied a wide range of commonly used network-generation models in an effort to reproduce the upward-sloping NCP plots and to understand the structural properties of the real-world networks that are responsible for this phenomenon. In each case, we have experimented with a range of parameters, and in no case have we been able to reproduce our empirical observations, at even a qualitative level. In Figure 17, we summarize these results.

There has been a large body of work subsequent to that [Albert and Barabási 99] on models in which edges are added via a preferential-attachment or rich-

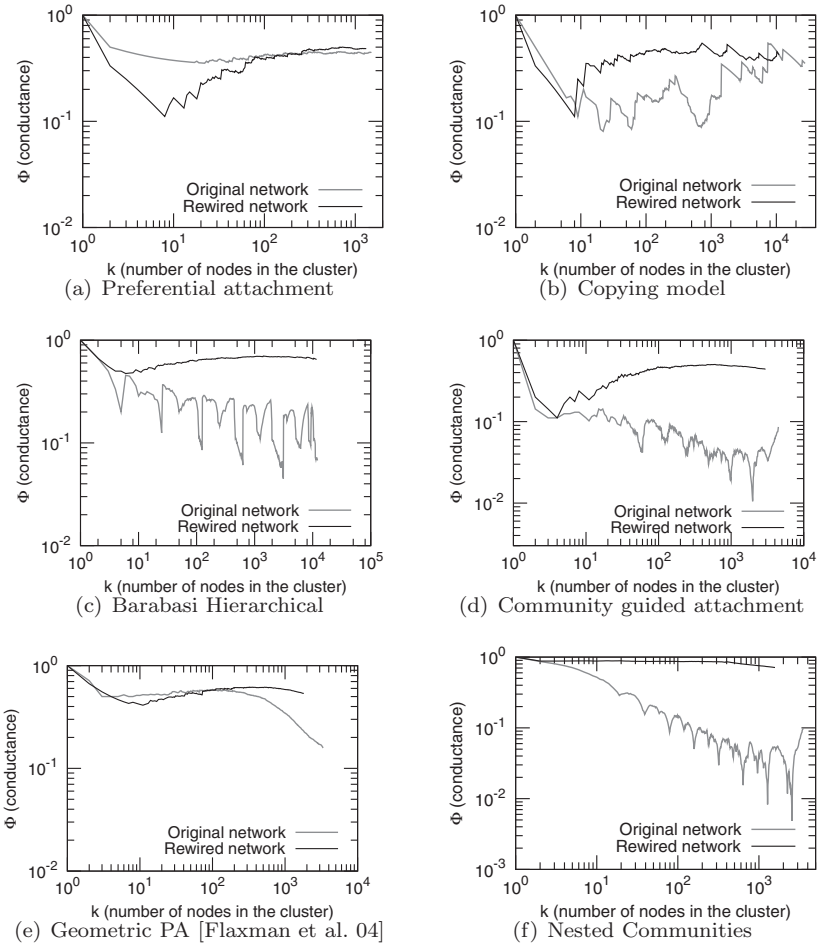


Figure 17. Network community profile for networks generated from commonly used procedures to generate graphs with heavy-tailed degree distributions: (a) preferential attachment; (b) copying model; (c) hierarchical model; (d) community guided attachment; (e) geometric preferential attachment; and (f) nested community model. See the text for details. Gray curves plot the results of the local spectral algorithm on the specified network, and black curves plot the results of the local spectral algorithm applied to a randomly rewired version of the same network.

gets-richer mechanism [Newman 03, Bollobás and Riordan 04]. Much of this work aims at reproducing properties of real-world graphs such as heavy-tailed degree distributions [Albert et al. 99, Broder et al. 00, Faloutsos et al. 99]. In these

preferential attachment models, one typically connects each new node to the existing network by adding exactly m edges to existing nodes with a probability that depends on the current degree of that existing node. Figure 17(a) shows the NCP plot for a 10,000-node network generated according to the original preferential attachment model [Albert and Barabási 99], where at each time step a node joins the graph and connects to $m = 2$ existing nodes. Note that the NCP plot is very shallow and flat (more so even than the corresponding rewired graph), and thus the network that is generated is expander-like at all size scales.

A different type of generative model is one in which edges are added via a copying mechanism [Kumar et al. 00]. In this copying model, a new node joins the network by attaching exactly m edges to existing nodes as follows: the new node first selects uniformly at random a "seed" or "ambassador" node u; then, for each of its m edges, with probability β the new node links to an existing node chosen randomly, and with probability $1 - \beta$ it links to a random neighbor of node u. In Figure 17(b), we show the results for a network with 50,000 nodes, generated with $m = 2$ and $\beta = 0.05$. Although intuitively the copying model aims to produce communities by linking a new node to neighbors of an existing node, this does not seem to be the right mechanism to reproduce the NCP plot, since potential ambassador nodes are all treated similarly and since new nodes always create the same number of edges.

Next, in Figure 17(c), we consider an example of a network that was designed to have a recursively hierarchical community structure [Ravasz et al. 02, Ravasz and Barabási 03]. In this model, we start with a five-node square-like structure with a central node, and then recursively expand the square and link it to the middle node of the network. This network has a power-law degree distribution and a clustering coefficient that decays in a characteristic manner [Ravasz and Barabási 03]. In this case, however, the NCP plot is sloping downward. The local dips in the plot correspond to multiples of the size of the basic module of the graph. Although the model generates links such that nodes that are farther apart in the hierarchy link less frequently, the NCP plot clearly indicates that in aggregate, larger communities are more easily separated than smaller communities.

A different way to generate power-law degree distributions is the community-guided attachment model [Leskovec et al. 05b]. Here we decompose the nodes of a graph into nested groups of nodes, such that the difficulty of forming links between nodes in different groups increases exponentially with the distance in the community hierarchy. Graphs generated by this principle have power-law degree distributions, and they also obey the densification power law [Leskovec et al. 05b, Leskovec et al. 07b]. As Figure 17(d) shows, though, the NCP plot is sloping downward. Qualitatively this plot from CGA is very similar to the plot

of the recursive hierarchical construction in Figure 17(c), which is not surprising given the similarities of the models.

Figure 17(e) shows the NCP plot for a geometric preferential attachment model [Flaxman et al. 04, Flaxman et al. 07]. This model aims to achieve a heavy-tailed degree distribution as well as edge locality, and it does so by making the connection probabilities depend both on the two-dimensional geometry and on the preferential attachment scheme. As we see, the effect of the underlying geometry eventually dominates the NCP plot, since the best bipartitions are fairly well balanced [Flaxman et al. 04]. Intuitively, geometric preferential attachment graphs look locally expander-like, but at larger size scales the union of such small expander graphs behaves like a geometric mesh. We also experimented with the small-world model of [Watts and Strogatz 98], in which the NCP plot in some sense behaves exactly the opposite (plot not shown): first the NCP plot decreases, and then it flattens out. Intuitively, a small-world network looks locally like a mesh, but when one reaches larger size scales, the randomly rewired edges start to appear and the graph looks like an expander.

Finally, we explored in more detail networks with explicitly planted community structure. For example, we started with ten isolated communities generated using the $G_{n,p}$ model, and then we generated a random binary tree. For each internal node at height h we link the nodes in both sides of the tree with probability p^h, for a probability parameter p. This and other related networks give a graph of nested communities resembling the hierarchical clustering algorithm of [Newman and Girvan 04]. We see, however, from Figure 17(f) that the NCP plot slopes steadily downward, and furthermore we observe that dips correspond to the cuts that separate the communities.

These experiments demonstrate that hierarchically nested networks and networks with underlying geometric or expander-like structure exhibit very different NCP plots from those observed in real networks. So the question still remains as to what causes an NCP plot to decrease and then start to increase.

6.2. Very Sparse Random Graphs Have Very Unbalanced Deep Cuts

In this section, we will analyze a very simple random graph model that reproduces relatively deep cuts at small size scales and that has an NCP plot that then flattens out. Understanding why this happens will be instructive as a baseline for understanding the community properties we have observed in our real-world networks.

Here we work with the random graph model with given expected degrees, as described in [Chung and Lu 06a, Chung and Lu 02b, Chung et al. 03a, Chung and Lu 02a, Chung and Lu 03, Chung et al. 03b, Chung et al. 04, Chung and

Lu 06b]. Let there be given n, the number of nodes in the graph, and a vector $\mathbf{w} = (w_1, \ldots, w_n)$, which will be the expected degree sequence vector (where we will assume that $\max_i w_i^2 < \sum_k w_k$). Then, in this random graph model, an edge e_{ij} between nodes i and j is added, independently, with probability $p_{ij} = w_i w_j / \sum_k w_k$. Thus, $P(e_{ij} = 1) = p_{ij}$ and $P(e_{ij} = 0) = 1 - p_{ij}$. We use $G(\mathbf{w})$ to denote a random graph generated in this manner.

Note that this model is different from the so-called configuration model, in which the degree distribution is exactly specified and which was studied in [Molloy and Reed 95, Molloy and Reed 98] and also [Aiello et al. 00, Aiello et al. 01]. This model is also different from generative models such as preferential attachment models [Albert and Barabási 99, Newman 03, Bollobás and Riordan 04] and models based on optimization [Doyle and Carlson 00, Doyle and Carlson 02, Fabrikant et al. 02], although common to all of these generative models is that they attempt to reproduce empirically observed power-law behavior [Albert et al. 99, Faloutsos et al. 99, Broder et al. 00, Newman 05, Clauset et al. 07].

In this random graph model, the expected average degree is $w_{\text{av}} = \frac{1}{n} \sum_{i=1}^{n} w_i$, and the expected second-order average degree is $\tilde{w} = \sum_{i=1}^{n} w_i^2 / \sum_k w_k$. Let $w_G = \sum_i w_i$ denote the expected total degree. Given a subset S of nodes, we define the volume of S to be $w_S = \sum_{v \in S} w_v$, and we say that S is c-giant if its volume is at least cw_G, for some constant $c > 0$. We will denote the actual degrees of the graph G by $\{d_1, d_2, \ldots, d_n\}$, and will define $d(S)$ to be the sum of the actual degrees of the vertices in S. Clearly, by linearity of expectation, for any subset S, $E(d(S)) = w_S$.

The special case of the $G(\mathbf{w})$ model in which $\mathbf{w}$ has a power-law distribution is of interest to us here. (The other interesting special case, in which all the expected degrees w_i are equal to np, for some $p \in [0, 1]$, corresponds to the classical Erdős–Rényi G_{np} random graph model [Bollobás 85].) Given the number of nodes n, the power-law exponent β, and the parameters w and $w_{\max}$, [Chung and Lu 06a] gives the degree sequence for a power-law graph:

$$w_i = ci^{-1/(\beta-1)} \quad \text{for } i \text{ such that } i_0 \leq i < n + i_0, \tag{6.1}$$

where, for the sake of consistency with their notation, we index the nodes from i_0 to $n+i_0-1$, and where $c = c(\beta, w, n)$ and $i_0 = i_0(\beta, w, n, w_{\max})$ are as follows:

$$c = \alpha w n^{1/(\beta-1)} \quad \text{and} \quad i_0 = n \left(\alpha \frac{w}{w_{\max}} \right)^{\beta-1},$$

where we have defined $\alpha = \frac{\beta-2}{\beta-1}$. It is easy to verify that $w_{\max} = \max_i w_i$ is the maximum expected degree; that the average expected degree is given by $w_{\text{av}} = \frac{1}{n} \sum_{i=1}^{n} w_i = w(1 + o(1))$; that the minimum expected degree is given

by $w_{\min} = \min_i w_i = w\alpha(1 - o(1))$; and that the number of vertices that have expected degree in the range $(k-1, k]$ is proportional to $k^{-\beta}$.

The following theorem characterizes the shape of the NCP plot for this $G(\mathbf{w})$ model when the degree distribution follows (6.1), with $\beta \in (2, 3)$. The theorem makes two complementary claims. First, there exists at least one (small but moderately deep) cut in the graph of size $\Theta(\log n)$ and conductance $\Theta(\frac{1}{\log n})$. Second, for some constants c' and ϵ, there are no cuts in the graph of size greater than $c' \log n$ having conductance less than ϵ. That is, this model has clusters of logarithmic size with logarithmically deep cuts, and once we get beyond this size scale there do not exist any such deep cuts.

Theorem 6.1. *Consider the random power-law graph model $G(\mathbf{w})$, where $\mathbf{w}$ is given by (6.1), where $w > 5.88$, and the power-law exponent β satisfies $2 < \beta < 3$. Then with probability $1 - o(1)$:*

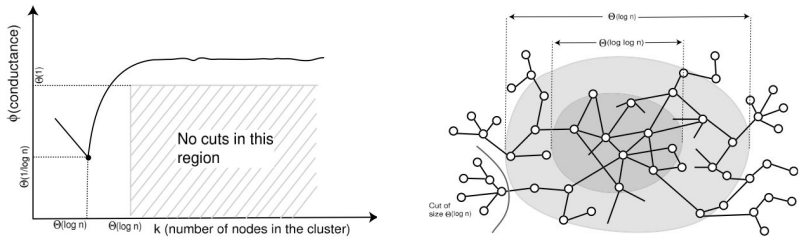
1. *There exists a cut of size $\Theta(\log n)$ whose conductance is $\Theta\left(\frac{1}{\log n}\right)$.*
2. *There exist $c', \epsilon > 0$ such that there are no sets of size larger than $c' \log n$ having conductance smaller than ϵ.*

Proof. Combine the results of Lemma 6.2 and Lemma 6.4. □

The two claims of Theorem 6.1 are illustrated in Figure 18(a). Note that when $w \geq \frac{4}{e}$ and $\beta \in (2, 3)$, then a typical graph in this model is not fully connected but does have a giant component [Chung and Lu 06a]. (The well-studied $G_{n,p}$ random graph model [Bollobás 85] has a similar regime when $p \in (1/n, \log n/n)$, as will be discussed in Section 7.4.)

In addition, under certain conditions on the average degree and second-order average degree, the average distance between nodes is $O(\log \log n)$ and yet the diameter of the graph is $\Theta(\log n)$. Thus, in this case, the graph has an "octopus" structure, with a subgraph containing $n^{c/(\log \log n)}$ nodes constituting a deep core of the graph. The diameter of this core is $O(\log \log n)$ and almost all vertices are at a distance of $O(\log \log n)$ from this core. However, the pairwise average distance of nodes in the entire graph is $O(\log n / \log \tilde{w})$. A schematic picture of the $G(\mathbf{w})$ model when $\beta \in (2, 3)$ is presented in Figure 18(b).

Our first lemma claims that for the $G(\mathbf{w})$ model, if the degree distribution $\mathbf{w}$ follows the above power law, then there exists a moderately large cut with small conductance. In order to prove the existence of a cut of size $\Theta(\log n)$ and conductance $\Theta(\frac{1}{\log n})$, it is sufficient to concentrate on the existence of whiskers that are large enough. In particular, to prove the following lemma, we compute



(a) NCP plot suggested by our main theorem

(b) Structure of $G(\mathbf{w})$ model

Figure 18. The $G(\mathbf{w})$ model in the sparse $\beta \in (2,3)$ parameter regime. (a) Network community profile plot, as suggested by our main theorem. (b) Caricature of network structure.

the probability that there exists a cut of both volume and size $\Theta(\log n)$ and cut size 1. (Note that although we formally state the lemma in terms of the power-law random graph model, the proof will show that the main claim holds for a more general representation of the heavy-tailed degree distribution.)

Lemma 6.2. *For the $G(\mathbf{w})$ model, where w follows a power-law degree distribution with $2 < \beta < 3$, then with probability $1 - o(1)$ there exists a set of size $\Theta(\log n)$ with conductance $\Theta(\frac{1}{\log n})$.*

Proof. Let S be a subset with the following description: $S = \{v_0, v_1, \ldots, v_k\}$, where $k = c_1 \log n$. Let w_i denote the degree of v_i. We have that $w_0 \in [c_2 \log n, 2c_2 \log n]$ and $w_i \leq w$ for all $i > 0$. Thus the expected volume of S is $w_S \in [(2\alpha w c_2 + c_1) \log n, (2\alpha w c_2 + c_1) \log n]$, and the size of S is $c_1 \log n + 1$. Note that the expected volume of the graph can be computed as $w_G = wn$, and hence $\rho = \frac{1}{w_G} = \frac{1}{wn}$.

Now let n_1 denote the number of vertices of expected degree at most $2\alpha w$. By simple calculation, $n_1 \geq n/2$. The number of possible choices for the vertex v_0 can be computed as follows. Let B be the set of vertices having degree greater than $2\alpha w c_2 \log n$ and let A be the set of vertices with degree at most $2\alpha w c_2 \log n$. Then the number of nodes with degree in $[c_2 \log n, 2c_2 \log n]$ is given by the size of $V \setminus (A \cup B)$, which is

$$\alpha w \left(\frac{n}{c_2 \log n}\right)^{\beta-1} - \alpha w \left(\frac{n}{2c_2 \log n}\right)^{\beta-1} \geq \alpha w \left(\frac{n}{c_2 \log n}\right)^{\beta-1},$$

since $\beta > 2$. Thus the number of possible such subsets S is given by the number of choices for v_0 times the number of possible choices for the nodes $v - 1, \ldots, v_k$.

Thus, the number N of possible such subsets S is at least

$$N = \binom{n_1}{c_1 \log n} \times \alpha \left(\frac{n}{2c_2 \log n} \right)^{\beta-1}.$$

We say that S is good if after instantiating all the edges, S has a star of size $c_1 \log n$ centered at v_0, and v_0 is connected to $\bar{S}$ by exactly one edge, and no other vertex in S has any edge to $\bar{S}$. The probability that a particular set S is good is the product of the following terms: the probability p_1 that there is a star of size $c_1 \log n$ with v_0 at the center, the probability p_2 that none of the nodes $v_1, \ldots, v_k$ link to any nodes in $\bar{S}$, and the probability p_3 that v_0 connects to $\bar{S}$ using exactly one edge. We now calculate the three probabilities as follows. First,

$$p_1 = \prod_{i=1}^{k} w_0 w_i \rho \geq (w_0 \alpha w \rho)^{c_1 \log n},$$

since for each w_i we have $w_i \geq w_{\min} \geq \alpha w$. Next,

$$p_2 = \prod_{i=1}^{k} \prod_{j \notin S} (1 - w_j \rho) \geq \prod_{i=1}^{k} \prod_{j \notin S} e^{-w_i \rho / 2} = e^{-(c_1 \rho 2 \alpha w w_{\bar{S}} \log n)/2},$$

obtained by using $1 - x \geq e^{-x/2}$ for $0 < x < 1$, and $w_i \leq 2\alpha w$ for $i \in S$, $i > 1$. Finally, we get p_3 as follows. First note that

$$\begin{aligned} p_3 &= \sum_{j \in \bar{S}} w_0 w_j \rho \prod_{k \neq j, k \in \bar{S}} (1 - w_k w_0 \rho) \\ &\geq \sum_{j \in \bar{S}} w_0 w_j \rho e^{-(w_{\bar{S}} - w_j) w_0 \rho / 2} \\ &= w_0 \rho e^{-w_{\bar{S}} w_0 \rho / 2} \Big(\sum_{j \in \bar{S}} w_j e^{w_j w_0 \rho / 2} \Big). \end{aligned}$$

Then, since $w_j w_0 \rho \ll 1$ and since $e^x \geq 1 + x$, we have that

$$\begin{aligned} p_3 &\geq w_0 \rho e^{-w_{\bar{S}} w_0 \rho / 2} \Big(\sum_{j \in \bar{S}} w_j \Big(1 + \frac{w_j w_0 \rho}{2} \Big) \Big) \\ &\geq w_0 \rho e^{-w_{\bar{S}} w_0 \rho / 2} (w_{\bar{S}} + w_0 \rho \tilde{w}_{\bar{S}} / 2), \end{aligned}$$

where $\tilde{w}_{\bar{S}} = \sum_{j \in \bar{S}} w_j^2$. So the final probability of goodness of S is

$$\begin{aligned} p &= p_1 \times p_2 \times p_3 \\ &\geq (w_0 \alpha w \rho)^{c_1 \log n} \times e^{-(c_1 \rho 2 \alpha w w_{\bar{S}} \log n)/2} \times w_0 \rho e^{-w_{\bar{S}} w_0 \rho / 2} (w_{\bar{S}} + w_0 \rho \tilde{w}_{\bar{S}} / 2) \\ &= (w_0 \alpha w \rho)^{c_1 \log n} \times e^{-(c_1 \gamma 2 \alpha w \log n)} \times w_0 \rho e^{-\gamma w_0} (w_{\bar{S}} + w_0 \rho \tilde{w}_{\bar{S}} / 2), \end{aligned}$$

using $\gamma = \rho w_{\bar{S}}/2$. So the expected number of such good subsets S is

$$\begin{aligned} Np &\geq \binom{n_1}{c_1 \log n} \times \alpha w \left(\frac{n}{2c_2 \log n}\right)^{\beta-1} \times (w_0 \alpha w \rho)^{c_1 \log n} \times e^{-(c_1 \gamma 2 \alpha w \log n)} \\ &\quad \times w_0 \rho e^{-\gamma w_0} (w_{\bar{S}} + w_0 \rho \tilde{w}_{\bar{S}}/2) \\ &\geq \left(\frac{n_1}{c_1 \log n}\right)^{c_1 \log n} \times \frac{\alpha w n^{\beta-1}}{(2c_2 \log n)^{\beta-1}} \times (w_0 \alpha w \rho)^{c_1 \log n} \times e^{-(c_1 \gamma 2 \alpha w \log n)} \\ &\quad \times w_0 \rho e^{-\gamma w_0} \times nw/2, \end{aligned}$$

using Stirling's formula and the fact that $w_{\bar{S}} \geq nw/2$.

Using the value of n_1 and since $nw\rho = 1$, we have

$$\begin{aligned} Np &\geq \left(\frac{n}{2c_1 \log n}\right)^{c_1 \log n} \times \frac{\alpha w n^{\beta-1}}{(2c_2 \log n)^{\beta-1}} \times (w_0 \alpha w \rho)^{c_1 \log n} \times e^{-(c_1 \gamma 2 \alpha w \log n)} \\ &\quad \times e^{-\gamma w_0} \times w_0/2 \\ &\geq \left(\frac{w_0 \alpha}{2c_1 \log n}\right)^{c_1 \log n} \times \frac{\alpha w n^{\beta-1}}{(2c_2 \log n)^{\beta-1}} \times (w_0 \alpha w \rho)^{c_1 \log n} \times e^{-(c_1 \gamma 2 \alpha w \log n)} \\ &\quad \times e^{-\gamma w_0} \times w_0/2. \end{aligned}$$

Using $w_0 \geq c_2 \log n$, we have that

$$\begin{aligned} Np &\geq \left(\frac{c_2 \alpha}{2c_1}\right)^{c_1 \log n} \times \frac{\alpha w n^{\beta-1}}{2(2c_2 \log n)^{\beta-2}} \times e^{-(c_1 \gamma 2 \alpha w \log n)} \times e^{-\gamma w_0} \\ &\geq e^{\Theta \log n} \times \frac{\alpha w}{2(2c_2 \log n)^{\beta-2}}, \end{aligned}$$

where $\Theta = c_1 \log(\frac{c_2 \alpha}{2c_1}) + (\beta - 1) - \gamma \alpha w c_1 - 2\gamma c_2$. Note that for $2 < \beta < 3$, we have that $0 < \alpha < \frac{1}{2}$. Also, $\gamma = \frac{1}{2} - o(1)$. Thus, choosing $c_2 = 2ec_1/\alpha$ and $c_1 = \frac{\beta-2}{2\gamma\alpha w + 4\gamma e/\alpha - 1}$, we get $\Theta = 1$. So,

$$Np \geq e^{\log n} \times \frac{\alpha w}{2(2c_2 \log n)^{\beta-2}} = \Omega(\log n).$$

Then the probability that a particular set S is good is $p \geq \Omega\left(\frac{(\log n)}{N}\right)$. Hence the probability of getting a good set is

$$1 - (1-p)^N \geq 1 - \left(1 - \Omega\left(\frac{(\log n)^{\beta-2}}{N}\right)\right)^N \geq 1 - o(1). \qquad \square$$

We next state the well-known Chernoff bound [Chung and Lu 06a], which we will use below.

Lemma 6.3. *Let $X = \sum_i X_i$, where the X_i are independent random variables with $X_i \geq -M$. Define $\|X\|^2 = \sum_i \mathbf{E}(X_i^2)$. Then,*

$$\mathbf{Pr}(X \geq \mathbf{E}(X) - \lambda) \leq \exp\left(-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}\right).$$

Finally, we show that there are no deep cuts with size greater than $\Theta(\log n)$. To state this lemma, define a connected set S to be ϵ-deficient if it has actual volume $d(S) \leq \frac{1}{2}d(G)$ and if the conductance of the cut $(S, \bar{S})$ is at most ϵ, i.e., if the number of edges leaving S is at most $\epsilon d(S)$.

Lemma 6.4. *For the $G(\mathbf{w})$ model, where w follows a power-law degree distribution with $2 < \beta < 3$, if the average degree w satisfies $w \geq 5.88$, then with probability $1 - o(1)$ there exist constants c', ϵ such that there is no ϵ-deficient set of size more than $c' \log n$.*

Proof. Let $e(S, \bar{S})$ denote the actual number of edges between S and $\bar{S}$. First we compute the probability that a given set S is ϵ-deficient, that is, S satisfies $e(S, \bar{S}) < \epsilon d(S)$. Let $\delta = \frac{2\epsilon}{1-\epsilon}$. For our case, define the variables $X_{(i,j)} = e_{ij}$ for $(i, j) \in (S, \bar{S})$ and $X_{(i,j)} = -\delta e_{ij}$ for $(i, j) \in (S, S)$. Then the sum $X = \sum X_{(i,j)}$ equals $\sum_{(i,j)\in(S,\bar{S})} e_{ij} - \delta \sum_{(i,j)\in(S,\bar{S})} e_{ij}$. Note that $e(S, \bar{S}) < \epsilon d(S) \iff X \leq 0$. Using the fact that $\mathbf{E}(e_{ij}) = w_i w_j \rho$, we have $\|X\|^2 = \sum \mathbf{E}(X_{ij}^2) = w_S w_{\bar{S}} \rho + \delta^2 w_S^2 \rho$. Furthermore, exploiting the fact that each X_i is greater than or equal to $-\delta$, we get that

$$\begin{aligned}
\mathbf{Pr}(X \leq 0) &= \mathbf{Pr}(X \leq \mathbf{E}(X) - \mathbf{E}(X)) \\
&\leq \exp\left(-\frac{\mathbf{E}(X)^2}{2(\|X\|^2 + \delta\mathbf{E}(X)/3)}\right) \\
&= \exp\left(-\frac{\rho^2 w_S^2 (w_{\bar{S}} - \delta w_S)^2}{2(w_S\rho(w_{\bar{S}} + \delta^2 w_S) + \delta w_S \rho(w_{\bar{S}} - \delta w_S)/3)}\right).
\end{aligned}$$

Canceling ρw_S from both numerator and denominator, we obtain

$$\begin{aligned}
\mathbf{Pr}(X \leq 0) &\leq \exp\left(-\frac{\rho w_S (w_{\bar{S}} - \delta w_S)^2}{2(w_{\bar{S}} + \delta^2 w_S + \delta w_{\bar{S}}/3 - \delta^2 w_S/3)}\right) \\
&\leq \exp\left(-\frac{\rho w_S (w_{\bar{S}} - \delta w_S)^2}{2(1 + \delta/3 + 2\delta^2/3) w_{\bar{S}}}\right) \\
&\leq \exp\left(-\frac{\rho w_S w_{\bar{S}} (1 - 2\delta w_S/w_{\bar{S}})}{2(1 + \delta/3 + 2\delta^2/3)}\right) \\
&\leq \exp\left(-\frac{\rho w_S w_{\bar{S}} (1 - 2\delta)}{2(1 + \delta/3 + 2\delta^2/3)}\right) \leq \exp\left(-\rho w_S w_{\bar{S}} A_\delta/2\right),
\end{aligned}$$

where $A_\delta = (1-2\delta)/((1+2\delta/3+2\delta^2/3))$. So this bounds the probability that a particular set S of size k is ϵ-deficient. We will bound the expected number of such ϵ-deficient subsets of size k.

First, let $N_{k,\epsilon,\gamma}$ denote the expected number of ϵ-deficient sets of size k that have expected volume $w_S \le \gamma w_G$. By linearity of expectation,

$$\begin{aligned} N_{k,\epsilon,\gamma} &\le \sum_{\substack{S:|S|=k \\ w_S \le \gamma w_G}} w_{i_1} \cdots w_{i_k} w_S^{k-2} \rho^{k-1} \exp\left(-\rho w_S w_{\bar{S}} A_\delta/2\right) \\ &\le \sum_{\substack{S:|S|=k \\ w_S \le \gamma w_G}} \frac{w_S^{2k-2}}{k^k} \rho^{k-1} \exp((-w_S(1-\gamma)A_\delta)), \end{aligned}$$

where we used the fact that $\gamma = \rho w_{\bar{S}}/2$ and also the AM-GM inequality to say that

$$\prod_{i \in S} w_i \le \left(\frac{\sum_{i \in S} w_i}{k}\right)^k.$$

Now, $F(x) = x^{2k-2} e^{-xA_\delta(1-\gamma)}$ is maximized at $x = (2k-2)/(A_\delta(1-\gamma))$. Thus, the above sum is maximized when $w_S = (2k-2)/(A_\delta(1-\gamma))$. Hence,

$$\begin{aligned} N_{k,\epsilon,\gamma} &\le \frac{n^k}{k!} \frac{\rho^{k-1}}{k^k} \frac{2^{(2k-2)} \cdot (k-1)^{(2k-2)}}{(A_\delta(1-\gamma))^{(2k-2)}} \exp(-2k+2) \\ &\le \frac{(n\rho)^k}{\rho\sqrt{k}(k/e)^k} \frac{1}{k^k} \frac{2^{(2k-2)} \cdot (k-1)^{(2k-2)}}{(A_\delta(1-\gamma))^{(2k-2)}} \exp(-2k+2). \end{aligned}$$

Using $(1-\frac{1}{k})^{2k} \le e^{-2}$, it follows that

$$N_{k,\epsilon,\gamma} \le \frac{1}{4e\sqrt{k}(k-1)^2} \left(\frac{4}{ewA_\delta^2(1-\gamma)^2}\right)^k.$$

We would like $\sum_{k=c\log n}^{cn} N_{k,\epsilon,\gamma}$ to be $o(1)$, for which we need

$$\frac{4}{ewA_\delta^2(1-\gamma)^2} < 1,$$

which gives a bound on the average degree:

$$w \ge \frac{4}{A_\delta^2(1-\gamma)^2 e}.$$

For sets of volume $w_S \ge \gamma w_G$, we have the following. From the double-sided Chernoff bound, for any fixed set S,

$$|w_S - d(S)| \le \lambda \quad \text{with probability } 1 - 2\exp\left(-\frac{\lambda^2}{2(w_S + \lambda/3)}\right).$$

So if $\lambda = \sqrt{w_S} \log n$, we have the above statement with probability

$$1 - 2\exp(-3\log^2 n/8).$$

Similarly,

$$|e(S,\bar{S}) - \mathbf{E}(e(S,\bar{S}))| \leq \lambda \quad \text{with probability } 1 - 2\exp\left(-\frac{\lambda^2}{2(\rho w_S w_{\bar{S}} + \lambda/3)}\right).$$

With $\lambda = \sqrt{\rho w_S w_{\bar{S}}} \log n$, the above probability becomes $1 - 2\exp(-3\log^2 n/8)$. Now, if both these events occur, then the conductance of the set S is at least $1/3$. So the only way we can get an ϵ-deficient set is by having one of these conditions invalid. The total number of sets of expected volume γw_G is bounded by $\binom{w_G}{\gamma w_G}$. So, the expected number of ϵ-deficient sets of volume at least γw_G is bounded by

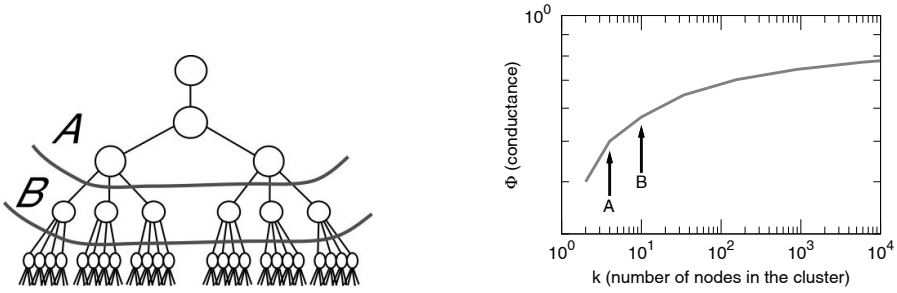
$$\begin{aligned}
&\sum_{\gamma \leq \theta \leq 1/2} \binom{w_G}{\theta w_G} 4\exp(-3\log^2 n/8) \\
&\qquad \leq \int_{\gamma \leq \theta \leq 1/2} \frac{1}{\sqrt{\theta w_G}} \left(\frac{1}{\theta}\right)^{\theta w_G} 4\exp(-3\log^2 n/8) \\
&\qquad \leq o(1).
\end{aligned}$$

Thus, putting the two bounds together, the expected number of ϵ-deficient sets of size greater that $c\log n$ is at most $o(1)$. Thus with probability $1 - o(1)$ there does not exist an ϵ-deficient set of size greater than $c\log n$. □

6.3. An Intuitive Toy Model for Generating an Upward-Sloping NCP Plot

We have seen that commonly studied models, including preferential attachment models, copying models, simple hierarchical models, and models in which there is an underlying meshlike or manifold-like geometry are not the right way to think out the community structure of large social and information networks. We have also seen that the extreme sparsity of the networks, coupled with randomness, can be responsible for the deep cuts at small scales.

To build intuition as to what the gradually increasing NCP plot might mean, consider Figure 19. This is a toy example of a network construction in which the NCP plot has a deep dip at a small size scale and then steadily increases. The network shown in Figure 19(a) is an infinite tree that has two parts. The top part, a subtree (with one node in this example, but more generally consisting of n_T nodes) is indicative of the whiskers, or the "small scale" structure of the graph. The remaining tree has the property that the number of children increases



(a) A toy network model ... (b) ... and its community profile plot.

Figure 19. Schematic picture of the properties of a network responsible for the upward-sloping community profile plot. (a) This toy model is designed so that the optimal conductance cuts are achieved by cutting nodes from the top of the tree. (b) The minimum of the NCP plot is achieved by cutting the single top node, and then larger and larger cuts have gradually worse and worse conductance values.

monotonically with the level of the node. This property is indicative of the fact that as the size of a cluster grows, the number of neighbors that it has also increases. The key insight in this construction is that the best conductance cuts first cut at the top of the growing tree and then gradually work their way "down" the tree, starting with the small subtrees and moving gradually down the levels, as depicted in Figure 19(a).

Thus, intuitively, one can think of small well-separated communities—those below the n_T size scale that consist of subsets of the small trees—starting to grow, and as they pass the n_T size scale and become bigger and bigger, they blend in more and more with the central part of the network, which (since it exhibits certain expander-like properties) does not have particularly well defined communities. Note (more generally) that if there are n_T nodes in the small tree at the top of the graph, then the dip in the NCP plot in Figure 19(b) is of depth $2/(n_T + 1)$. In particular, if $n_T = \Theta(\log n)$, then the depth of this cut is $\Theta(1/\log n)$.

Intuitively, the NCP plot increases, since the "cost" per edge for every additional edge inside a cluster increases with the size of the cluster. For example, in cut A in Figure 19(a), the "price" for having three internal edges is to cut six edges, i.e., two edges cut per edge inside. To expand the cluster by just a single edge, one has to move one level down in the tree (toward the cut B), where now the price for a single edge is four edges, and so on.

6.4. A More Realistic Model of Network Community Structure

The question arises now as to whether we can find a simple generative model that can explain both the existence of small well-separated whisker-like clusters

and also an expander-like core whose best clusters get gradually worse as the purported communities increase in size. Intuitively, a satisfactory network-generation model must successfully take into account the following two mechanisms:

(a) The model should produce a relatively large number of relatively small—but still large when compared to random graphs—well-connected and distinct whisker-like communities. (This should reproduce the downward part of the community profile plot and the minimum at small size scales.)

(b) The model should produce a large expander-like core, which may be considered as consisting of intermingled communities, perhaps growing out from the whisker-like communities, the boundaries of which get less and less well defined as the communities get larger and larger and as they gradually blend in with rest of the network. (This should reproduce the gradual upward-sloping part of the community profile plot.)

The so-called *forest fire model* [Leskovec et al. 05b, Leskovec et al. 07b] captures exactly these two competing phenomena. The forest fire model is a model of graph generation (which generates directed graphs—an effect we will ignore) in which new edges are added via a recursive "burning" mechanism in an epidemic-like fashion. Since the details of the recursive burning process are critical to the model's success, we explain it in some detail.

To describe the forest fire model of [Leskovec et al. 05b, Leskovec et al. 07b], let us fix two parameters, a *forward burning probability* p_{f} and a *backward burning probability* p_{r}. We start the entire process with a single node, and at each time step $t > 1$, we consider a new node v that joins the graph G_t constructed thus far. The node v forms out-links to nodes in G_t as follows:

(i) Node v first chooses a node w, which we will refer to as a "seed" node or an "ambassador" node, uniformly at random and forms a link to w.

(ii) Node v selects x out-links and y in-links of w that have not yet been visited. (x and y are two geometrically distributed random numbers with means $p_{\mathrm{f}}/(1-p_{\mathrm{f}})$ and $p_{\mathrm{r}}/(1-p_{\mathrm{r}})$, respectively. If not enough in-links or out-links are available, then v selects as many as possible.) Let $w_1, w_2, \ldots, w_{x+y}$ denote the nodes at the other ends of these selected links.

(iii) Node v forms out-links to $w_1, w_2, \ldots, w_{x+y}$, and then applies step (ii) recursively to each of $w_1, w_2, \ldots, w_{x+y}$, except that nodes cannot be visited a second time during the process.

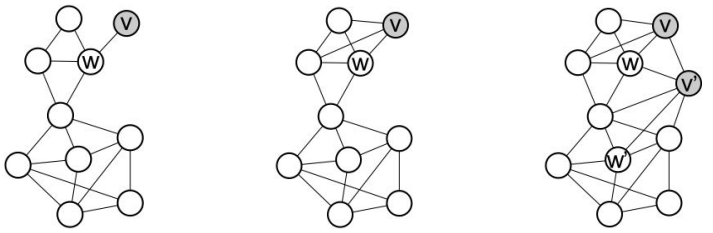


Figure 20. The forest fire burning process. Left: a new node v joins the network and selects a seed node w. Middle: v then attaches itself by recursively linking to w's neighbors, w's neighbor's neighbors, and so on, according to the "forest fire" burning mechanism described in the text. Right: a new node v' joins the network, selects seed w', and recursively adds links using the same "forest fire" burning mechanism. Notice that if v' causes a large "fire," it links to a large number of existing nodes. In this way, as potential communities around node w grow, the NCP plot is initially decreasing, but then larger communities around w gradually blend in with the rest of the network, increasing the NCP plot.

Thus, burning of links in the forest fire model begins at node w, spreads to $w_1, w_2, \ldots, w_{x+y}$, and proceeds recursively until the process dies out. One can view such a process intuitively as corresponding to a model in which a person arrives at a party and first meets an ambassador, who then introduces him or her around. If the person creates a small number of friendships, these will likely be from the ambassador's "community," but if the person happens to create many friendships, then these will likely go outside the ambassador's circle of friends. In this way, the ambassador's community might gradually become intermingled with the rest of the network.

Two properties of this model are particularly significant. First, although many nodes might form one or a small number of links, certain nodes can produce large conflagrations, burning many edges and thus forming a large number of out-links before the process ends. Such nodes will help generate a skewed out-degree distribution, and they will also serve as "bridges" that connect formerly disparate parts of the network. This will help make the NCP plot gradually increase. Second, there is a locality structure in that as each new node v arrives over time, it is assigned a "center of gravity" in some part of the network, i.e., at the ambassador node w, and the manner in which new links are added depends sensitively on the local graph structure around node w. Not only does the probability of linking to other nodes decrease rapidly with distance to the current ambassador, but because of the recursive process, regions with a higher density of links tend to attract new links.

Figure 20 illustrates this. Initially, there is a small community around node w. Then, node v joins and using the forest fire mechanism locally attaches to nodes

in the neighborhood of seed node w. The growth of the community around w corresponds to the downward part of the NCP plot. However, if a node v' then joins and causes a large fire, this has the effect of larger and larger communities around w blending into and merging with the rest of the network.

Not surprisingly, however, the forest fire model is sensitive to the choice of the burning probabilities p_f and p_b. We have experimented with a wide range of network sizes and values for these parameters, and in Figure 21, we show the community profile plots of several 10,000-node forest fire networks generated with $p_b = 0.3$ and several different values of p_f. The first thing to note is that since we are varying p_f over the six plots in Figure 21, we are viewing networks with very different densities. Next, notice that if, for example, $p_f = 0.33$ or $p_f = 0.35$, then we observe a very natural behavior: the conductance nicely decreases, reaches the minimum somewhere between 10 and 100 nodes, and then slowly but not too smoothly increases. Not surprisingly, it is in this parameter region where the forest fire model has been shown to exhibit realistic time-evolving graph properties such as densification and shrinking diameters [Leskovec et al. 05b, Leskovec et al. 07b].

Next, also notice that if p_f is too low or too high, then we obtain qualitatively different results. For example, if $p_f = 0.26$, then the community profile plot gradually decreases for nearly the entire plot. For this choice of parameters, the forest fire does not spread well, since the forward burning probability is too small and the network is extremely sparse and is treelike with just a few extra edges, and so we get large well-separated "communities" that get better as they get larger. On the other hand, when the burning probability is too high, e.g., $p_f = 0.40$, then the NCP plot has a minimum and then rises extremely rapidly. For this choice of parameters, if a node that initially attached to a whisker successfully burns into the core, then it quickly establishes many successful connections to other nodes in the core. Thus, the network has relatively large whiskers that failed to establish such a connection and a very expander-like core, with no intermediate region, and the increase in the community profile plot is quite abrupt.

We have examined numerous other properties of the graphs generated by the forest fire model and have found them to be broadly consistent with the social and information networks we have examined. One property, however, that is of particular interest is what the whiskers look like. Figure 22 shows an example of several whiskers generated by the forest fire model if we choose $p_b = 0.30$ and $p_f = 0.37$. They are larger and better structured than the treelike whiskers from the random graph model of Section 6.2. Also notice that they all look plausibly community-like with a core of the nodes densely linked among themselves, and the bridge edge then connects the whisker to the rest of the network.

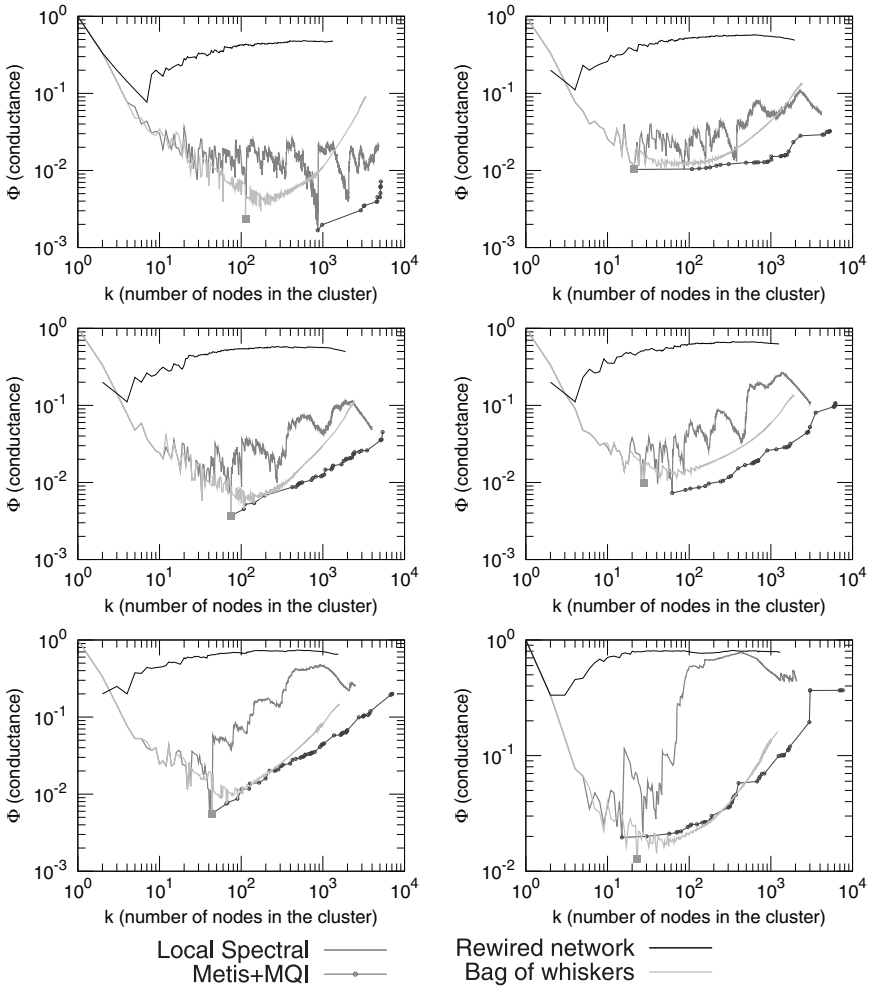


Figure 21. (Best viewed in color; see [Leskovec et al. 08a].) Community profile plots for the forest fire model at various parameter settings. The backward burning probability is $p_b = 0.3$, and we increase (left to right, top to bottom) the forward burning probability $p_f = \{0.26, 0.31, 0.33, 0.35, 0.37, 0.40\}$. Note that the largest and smallest values for p_f lead to less-realistic community profile plots, as discussed in the text.

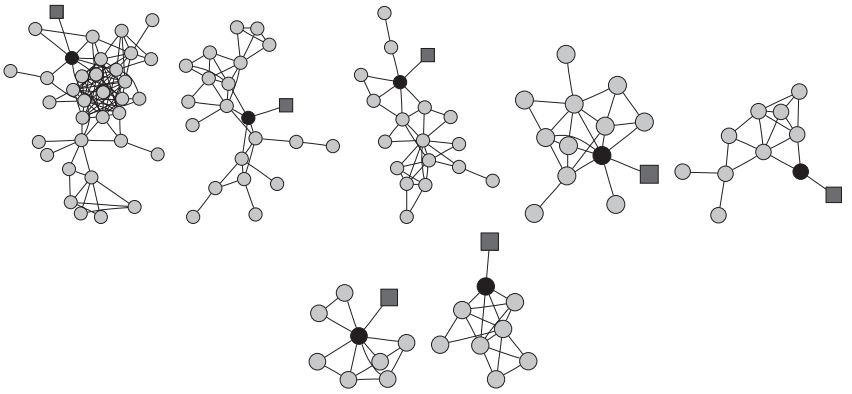


Figure 22. Examples of whiskers from a simulation of the forest fire model with parameter settings $p_f = 0.37$ and $p_b = 0.3$. The dark gray square node belongs to the network core, and by cutting the edge connecting it with the black circular node we separate the community of circles from the rest of the network (depicted as a dark gray square).

We conclude by noting that there has also been interest in developing hierarchical graph generation models, i.e., models in which a hierarchy is given and the linkage probability between pairs of nodes decreases as a function of their distance in the hierarchy [Ravasz et al. 02, Ravasz and Barabási 03, Chakrabarti et al. 04, Abello 04, Leskovec et al. 05a, Clauset et al. 06, Xuan et al. 06, Leskovec and Faloutsos 07]. The motivation for this comes largely from the intuition that nodes in social networks form small relatively tight groups that then further join into larger groups, and so on. As Figures 17(c) and 17(d) make clear, however, such models do not immediately lead to a community structure similar to what we have observed and which has been reproduced by the forest fire model. On the other hand, although there are significant differences between hierarchical models and the forest fire model, it is noted in [Leskovec et al. 05b, Leskovec et al. 07b] that there are similarities. In particular, in the forest fire model a new node v is assigned an ambassador w as an entry point to the network. This is analogous to a child having a parent in the hierarchy, which helps to determine how that node links to the remainder of the network. Similarly, many hierarchical models have a connection probability that decreases exponentially in the hierarchical tree distance. In the forest fire model, the probability that a node v will burn along a particular path to another node u' is exponentially small in the path length, although the analogy is not perfect, since there may exist many possible paths.

7. Discussion

In this section, we discuss several aspects of our main results in a broader context. In particular, in Section 7.1, we compare several data sets in which there is some notion of "ground truth" community, and we also describe several broader nontechnical implications of our results. Then, in Section 7.3, we describe recent work on community detection and identification. Finally, in Section 7.4, we discuss several technical and algorithmic issues and questions raised by our work.

7.1. Comparison with "Ground Truth" and Sociological Communities

In this subsection, we examine the relationship between network communities of the sort we have been discussing so far and some notion of "ground truth." When considering a real network, one hopes that the output of a community-finding algorithm will be "real" communities that exist in some meaningful sense in the real world. For example, in the karate club network in Figure 5(a), the cut found by the algorithm corresponds in some sense to a true community, in that it splits the nodes almost precisely as they split into two newly formed karate clubs.

In this section, we take a different approach: we take networks in which there are explicitly defined communities, and we examine how well these communities are separated from the rest of the network. In particular, we examine a minimum-conductance profile of several network data sets, where we can associate with each node one or more community labels that are exogenously specified. Note that we are overloading the term "community" here, since in this context the term might mean one of two things: first, it can refer to groups of nodes with good conductance properties; and second, it can refer to groups of nodes that belong to the same self-defined or exogenously specified group.

We consider the following five data sets:

- LiveJournal12 [Backstrom et al. 06]: LiveJournal is an online blogging site where users can create friendship links to other users. In addition, users can create groups that other users can then join. In LiveJournal, there are 385,959 such groups, and a node belongs to 3.5 groups on average. Thus, in addition to the information in the interaction graph, we have labels specifying those groups with which a user is associated, and thus we may view each such group as determining a "ground truth" community.

- CA-DBLP [Backstrom et al. 06]: We considered a coauthorship network in which nodes are authors and there is an edge if authors coauthored at least one paper. Here, publication venues (e.g., journals and conferences)

can play the role of "ground truth" communities. That is, an author is a member of a particular group or community if he or she published at a particular conference or in a particular journal. In our DBLP network, there are 2,547 such groups, with a node belonging to 2.6 on average.

- AMAZONALLPROD [Clauset et al. 04]: This is a network of products that are commonly purchased together at amazon.com. (Intuitively, one might expect that, for example, gardening books are frequently purchased together, so the network structure might reflect a well-connected cluster of gardening books.) Here, each item belongs to one or more hierarchically organized categories (book, movie genres, product types, etc.), and products from the same category define a group that we will view as a "ground truth" community. Items can belong to 49,732 different groups, and each item belongs to 14.3 groups on average.

- ATM-IMDB: This network is a bipartite actors-to-movies network composed of IMDB data, and an actor A is connected to movie B if A appeared in B. For each movie we also know the language and the country where it was produced. Countries and languages may be taken as "ground truth" communities or groups, where every movie belongs to exactly one group and actors belong to all groups to which movies that they appeared in belong. In our data set, we have 393 language groups and 181 country groups.

- EMAIL-INSIDE and EMAIL-INOUT [Leskovec et al. 07b]: This is an email communication network from a large European research organization conducting research in the natural sciences: physics, chemistry, biology, and computer science. Each of 986 members of the organization belongs to exactly one of 45 departments, and we use the department memberships to define "ground truth" communities.

Although none of these notions of "ground truth" is perfect, many community-finding algorithms use precisely this form of anecdotal evaluation: a network is taken, network communities are found, and then the correspondence of network communities to "ground truth" communities is evaluated. Note, in contrast, that we are evaluating how "ground truth" communities behave at different size scales with respect to our methodology, rather than examining how the groups we find relate to "ground truth" communities. Furthermore, note that the notions of "ground truth" are not all the same—we might expect that people publish papers across several different venues in a very different way from that in which actors appear in movies from different countries. More detailed statistics for each of these networks may be found in Tables 1, 2, and 3.

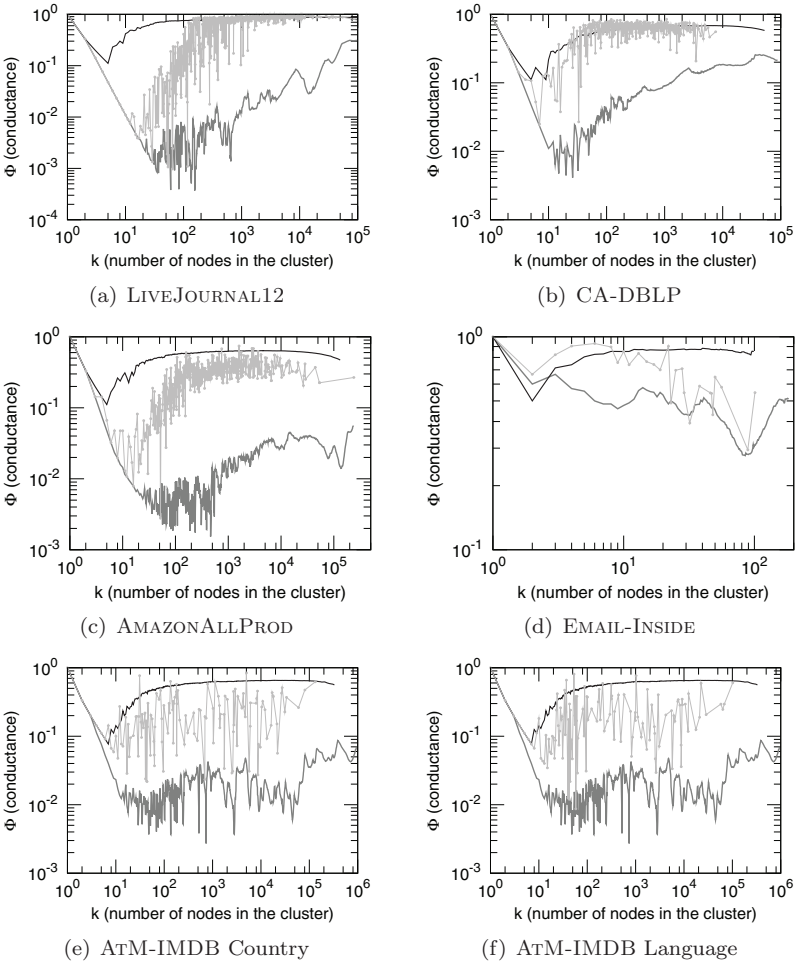
To examine the quality of "ground truth" communities in these network data sets, we take all groups and measure the conductance of the cut that separates that group from the rest of the network. Thus, we generated NCP plots in the following way. For every "ground truth" community, we measured the conductance of the cut separating it from the rest of the graph, from which we obtained a scatter plot of community size versus conductance. Then, we took the lower envelope of that plot, i.e., for every integer k we found the conductance value of the community of size k with the lowest conductance. Figure 23 shows the results for these network data sets; the figure also shows the NCP plot obtained using the local spectral algorithm on both the original network (plotted in dark gray) and the rewired network (plotted in black).

Several observations can be made:

- The conductance of "ground truth" communities follows that for the network communities up to nodes of size 10 to 100, i.e., larger communities get successively more community-like. As "ground truth" communities get larger, their conductance values tend to get worse and worse, in agreement with network communities discovered with graph-partitioning approximation algorithms. Thus, the qualitative trend we observed in nearly every large sparse real-world network (of the best communities blending in with the rest of the network as they grow in size) is seen to hold for small "ground truth" communities.

- One might expect that the NCP plot for the "ground truth" communities (the light gray curves) will be somewhere between the NCP plot of the original network (dark gray curves) and that for the rewired network (black curves), and this is seen to be the case in general. The NCP plot for network communities goes much deeper and rises more gradually than for "ground truth" communities. This is also consistent with our general observation that only small communities tend to be dense and well separated, and to separate large groups one has to cut disproportionately many edges.

- For the two social networks we studied (LiveJournal12 and CA-DBLP), larger "ground truth" communities have conductance scores that get quite "random," i.e., they are as well separated as they would be in a randomly rewired network (light gray and black curves overlap). This is likely associated with the relatively weak and overlapping notion of "ground truth" that we associated with those two network data sets. On the other hand, for the AmazonAllProd and AtM-IMDB networks, the general trend still remains, but large "ground truth" communities have conductance scores that lie well below those of the rewired network curve.



(a) LiveJournal12 (b) CA-DBLP

(c) AmazonAllProd (d) Email-Inside

(e) AtM-IMDB Country (f) AtM-IMDB Language

Figure 23. (Best viewed in color; see [Leskovec et al. 08a].) Network community profile plots for explicitly "ground truth" communities (light gray), compared with those for the original network (dark gray) and a rewired version of the network (black): (a) LiveJournal12; (b) CA-DBLP; (c) AmazonAllProd; (d) Email-Inside; and (e), (f) AtM-IMDB.

Our email network illustrates a somewhat different point. The NCP plot for Email-Inside should be compared with that for Email-InOut, which is displayed in Figure 7. The Email-Inside email network is rather small, and so it has a decreasing community profile plot, in agreement with the results for small social networks. Since communication is mainly focused between the members

of the same department, both network and "ground truth" communities are well expressed. Next, compare the NCP plot of EMAIL-INSIDE with that of EMAIL-INOUT (Figure 7). We see that the NCP plot of EMAIL-INSIDE slopes downward (since we consider communication only inside the organization), but as soon as we consider communication inside the organization and to the outside world (EMAIL-INOUT, or alternatively, see EMAIL-ENRON), then we see a completely different and more familiar picture—the NCP plot drops and then slowly increases.

This suggests that the organizational structure (e.g., departments) manifests itself in the internal communication network, but as soon as we put the organization into a broader context (i.e., how it communicates with the rest of the world), then the internal department structure seems to disappear.

7.2. Connections and Broader Implications

In contrast to numerous studies of community structure, we find that there is a natural size scale to communities. Communities are relatively small, with sizes only up to about one hundred nodes. We also find that above size of about one hundred, the "quality" of communities gets worse and worse, and communities more and more "blend into" the network. Eventually, even the existence of communities (at least when viewed as sets with stronger internal than external connectivity) is rather questionable. We show that large social and information networks can be decomposed into a large number of small communities and a large dense and intermingled network "core"—we empirically establish that the "core" contains on average 60% of the nodes and 80% of all edges. But as demonstrated by Figure 13, the "core" itself has a nontrivial structure—in particular, it has a core–whisker structure that is analogous to the original complete network.

The Dunbar Number. Our observation on the limit of community size agrees with [Dunbar 98], which predicted that roughly 150 is the upper limit on the size of a well-functioning human community. Moreover, [Allen 04] gives evidence that online communities have around 60 members, and on-line discussion forums start to break down at about 80 active contributors. Church congregations, military companies, divisions of corporations, all are close to the number 150 [Allen 04]. We are thus led to ask, why, above this size, is community quality inversely proportional to its size? And why are NCP plots of small and large networks so different?

Previous studies mainly focused on small networks (e.g., see [Danon et al. 05]), which are simply not large enough for the clusters to gradually blend into one another as one looks at larger size scales. Our results do not disagree with the

literature at small sizes. But it seems that in order to make our observations, one needs to look at large networks. It is only when Dunbar's limit is passed that we find large communities blurring and eventually vanishing. A second reason is that previous work did not measure and examine the *network community profile* of cluster size versus cluster quality.

Common Bond versus Common Identity Communities. Dunbar's explanation aligns well with the common bond versus common identity theory of group attachment [Ren et al. 07] from social psychology. Common identity theory makes predictions about people's attachment to the group as a whole, while common bond theory predicts people's attachment to individual group members. The distinction between the two refers to people's different reasons for being in a group. Because they like the group as a whole we get identity-based attachment, or because they like individuals in the group we get bond-based attachment. Anecdotally, bond-based groups are based on social interaction with others, personal knowledge of them, and interpersonal attraction to them. On the other hand, identity-based groups are based on common identity of its members, e.g., liking to play a particular online game or contributing to Wikipedia. It has been noted that bond communities tend to be smaller and more cohesive [Back 51], since they are based on interpersonal ties, while identity communities are focused around a common theme or interest. See [Ren et al. 07] for a very good review of the topic.

Translating this to our context, the bond versus identity communities mean that small, cohesive, and well-separated communities are probably based on common bonds, while bigger groups may be based on common identity, and it is hard to expect such big communities to be well separated or well expressed in a network sense. This further means that the transition between common bond (i.e., maintaining close personal ties) and common identity (i.e., sharing a common interest or theme) occurs at around one hundred nodes. It seems that at this size the cost of maintaining bond ties becomes too large, and the group either dies or transitions into a common identity community. It would be very interesting as a future research topic to explore differences in community network structure as the community grows and transitions from a common bond to a common identity community.

Edge Semantics. Another explanation could be that in small, carefully collected networks, the semantics of edges is very precise, while in large networks we know much less about each particular edge, especially, for example, when members of an online social network have very different criteria for calling someone a friend. Traditionally, social scientists used questionnaires to "normalize" the links by making sure that each link has the same semantics/strength.

Evidence in Previous Work. There has also been some evidence that hints at the findings we make here. For example, [Clauset et al. 04] analyzed community structure of a graph related to AMAZONALLPROD, and they found that around 50% of the nodes belonged to the largest "miscellaneous" community. This agrees with the typical size of the network core, and one could conclude that the largest community they found likely corresponds to the intermingled core of the network, and most of the rest of the communities are whisker-like.

In addition, recently there have been several works hinting that the subject of network communities is more complex than might seem at first glance. For example, it has been found that even random graphs can have good modularity scores [Guimerà et al. 04]. Intuitively, random graphs have no community structure, but there can still exist sets of nodes with good community scores, at least as measured by modularity (due to random fluctuations about the mean). Moreover, very recently a study of robustness of community structure showed that the canonical example of presence of community structure in networks [Zachary 77] may have no significant community structure [Karrer et al. 07].

More General Thoughts. Our work also raises an important question of what is a natural community size and whether larger communities (in a network sense) even exist. It seems that when community size surpasses some threshold, the community becomes so diverse that it stops existing as a traditionally understood "network community." Instead, it blends in with the network, and intuitions based on connectivity and cuts seem to fail to identify it. Approaches that consider both the network structure and node attribute data might help to detect communities in these cases.

Also, conductance seems like a very reasonable measure that satisfies intuition about community quality, but we have seen that if one worries only about conductance, then bags of whiskers and other internally disconnected sets have the best scores. This raises interesting questions about cluster compactness, regularization, and smoothness: What is a good definition of compactness? What is the best way to regularize these noisy networks? And how should this be connected to the notion of community separability?

A common assumption is that each node belongs to exactly one community. Our approach does not make such an assumption. Instead, for each given size, we independently find the best set of nodes, and "communities" of different sizes often overlap. As long there is a boundary between communities (even if boundaries overlap), cut- and edge-density-based techniques (such as modularity and conductance) may have the opportunity to find those communities. However, it is the absence of clear community boundaries that makes the NCP plot go upward.

7.3. Relationship to Community Identification Methods

A great deal of work has been devoted to finding communities in large networks, and much of this has been devoted to formalizing the intuition that a community is a set of nodes that has more and/or better intralinkages between its members than interlinkages with the remainder of the network. Very relevant to our work is [Kannan et al. 04], which analyzes spectral algorithms and describes a community concept in terms of a bicriterion depending on the conductance of the communities and the relative weight of intercommunity edges. [Flake et al. 03] introduces a similar bicriterion that is based on network flow ideas, and [Flake et al. 00, Flake et al. 02] defined a community as a set of nodes that has more intra-edges than inter-edges. Similar edge-counting ideas were used in [Radicchi et al. 04] to define and apply the notions of a strong community and a weak community.

Within the "complex networks" community, [Girvan and Newman 02] proposed an algorithm that uses "centrality" indices to find community boundaries. Following this, [Newman and Girvan 04] introduced *modularity* as an a posteriori measure of the strength of community structure. Modularity measures inter- (and not intra-) connectivity, but it does so with reference to a randomized null model. Modularity has been very influential in the recent community-detection literature [Newman 04, Danon et al. 05], and one can use spectral techniques to approximate it [White and Smyth 05, Newman 06b]. On the other hand, [Guimerà et al. 04] and [Fortunato and Barthélemy 07] showed that random graphs have high-modularity subsets and that there exists a size scale below which communities cannot be identified. In part as a response to this, some recent work has had a more statistical flavor [Hastings 06, Reichardt and Bornholdt 07, Rosvall and Bergstrom 07, Karrer et al. 07, Newman and Leicht 07]. In light of our results, this work seems promising, both due to potential "overfitting" issues arising from the extreme sparsity of the networks, and also due to the empirically promising regularization properties exhibited by local spectral methods.

We have made extensive use of the local spectral algorithm of [Andersen et al. 06]. Similar results were originally proven in [Spielman and Teng 04a, Spielman and Teng 04b], which analyzed local random walks on a graph; see [Chung 07a, Chung 07c, Chung 07b] for an exposition of the relationships between these methods. [Andersen and Lang 06] showed that these techniques can find (in a scalable manner) medium-sized communities in very large social graphs in which there exist reasonably well defined communities. In light of our results, such methods seem promising more generally. Other recent work that has focused on developing local and/or near-linear-time heuristics for community

detection include [Clauset et al. 04, Wu and Huberman 04, Clauset 04, Bagrow and Bollt 05, Raghavan et al. 07].

In addition to this work we have cited, there exists work that views communities from a very different perspective. For example, [Kumar et al. 99] views communities as a dense bipartite subgraph of the web; [Gibson et al. 98] views them as consisting of a core of central authoritative pages linked together by hub pages; [Hopcroft et al. 03, Hopcroft et al. 04] are interested in the temporal evolution of communities that are robust when the input data to clustering algorithms that identify them are moderately perturbed; and [Palla et al. 05] views communities as a chain of adjacent cliques and focuses on the extent to which they are nested and overlap. The implications of our results for this body of work remain to be explored.

7.4. Relationship to Other Theoretical Work

In this subsection, we describe the relationship between our work and recent work with similar flavor in graph partitioning, algorithms, and graph theory.

Recent work has focused on the expansion properties of power-law graphs and the real-world networks they model. For example, [Mihail et al. 06], as well as [Gkantsidis et al. 03], studied Internet routing at the level of autonomous systems (AS), and showed that the preferential attachment model and a random graph model with power-law degree distributions each have good expansion properties if the minimum degree is greater than 2 or 3, respectively. This is consistent with empirical results, but as we have seen, the AS graphs are quite unusual when compared with nearly every other social and information network we have studied. On the other hand, Estrada has made the observation that although certain communication, information, and biological networks have good expansion properties, social networks do not [Estrada 06]. This is interpreted as evidence that such social networks have good small highly cohesive groups, a property that is not attributed to the biological networks that were considered. From the perspective of our analysis, these results are interesting, since it is likely that these small highly cohesive groups correspond to sets near the global minimum of the network community profile plot. Reproducing deep cuts was also a motivation for the development of the geometric preferential attachment models of [Flaxman et al. 04, Flaxman et al. 07]. Note, however, that the deep cuts they obtain arise from the underlying geometry of the model and thus are nearly bisections.

Consider also recent results on the structural and spectral properties of very sparse random graphs. Recall that the G_{np} random graph model [Bollobás 85]

consists of those graphs on n nodes in which there is an edge between every pair of vertices with probability p, independently. Recall also that if $p \in (1/n, \log n/n)$, then a typical graph in G_{np} has a giant component, i.e., a connected subgraph consisting of a constant fraction of the nodes, but the graph is not fully connected [Bollobás 85]. (If $p < 1/n$, then a typical graph is disconnected and there does not exist a giant component, while if $p > \log n/n$, then a typical graph is fully connected.) As noted, for example, in [Feige and Ofek 05], this latter regime is particularly difficult to analyze, since with fairly high probability there exist vertices with degrees that are much larger than their expected degree. As reviewed in Section 6.2, however, this regime is not unlike that in a power-law random graph in which the power-law exponent β is in $(2, 3)$ [Chung and Lu 01, Lu 01, Chung and Lu 06a].

[Chakrabarti et al. 07] defined the "min-cut" plot, which has similarities with our NCP plot. The authors used a different approach in which a network was recursively bisected and then the quality of the obtained clusters was plotted as a function of size; and the "min-cut" plots were used only as yet another statistic to assess how realistic synthetically generated graphs are. Note, however, that the "min-cut" plots have qualitatively similar behavior to our NCP plots, i.e., they initially decrease, reach a minimum, and then increase.

Of particular interest to us are recent results on the mixing time of random walks in this $p \in (1/n, \log n/n)$ regime of the G_{np} (and the related G_{nm}) random graph model. [Benjamini et al. 06] and [Fountoulakis and Reed 07b, Fountoulakis and Reed 07a] have established rapid mixing results by proving structural results about these very sparse graphs. In particular, they proved that these graphs may be viewed as a "core" expander subgraph whose deletion leaves a large number of "decorations," i.e., small components such that a bounded number are attached to any vertex in the core. The particular constructions in their proofs are complicated, but they have a similar flavor to the core-and-whiskers structure we have empirically observed. Similar results were observed in [Fernholz and Ramachandran 07], whose analysis separately considered the 2-core of these graphs and then the residual pieces. The authors show that a typical longest shortest path between two vertices u and v (by which we mean the maximum over all such pairs) consists of a path of length $O(\log n)$ from u to the 2-core, then a path of length $O(\log n)$ across the 2-core, and finally a path of length $O(\log n)$ from the 2-core to v. Again, this is reminiscent of the core-and-whiskers properties we have observed. In all these cases, the structure is very different from that of traditional expanders [Hoory et al. 06], which we also empirically observe. Eigenvalues of power-law graphs have also been studied in [Mihail and Papadimitriou 02, Chung et al. 03a, Chung et al. 03b, Chung et al. 04, Flaxman et al. 05].

8. Conclusion

We have investigated statistical properties of community-like sets of nodes in large real-world social and information networks. We discovered that community structure in these networks is very different from what we expected from experience with small networks and from what commonly used models would suggest.

In particular, we defined a *network community profile plot (NCP plot)*, and we observed that good network communities exist only up to a size scale of ≈ 100 nodes. This agrees well with the observations of Dunbar. For size scales above ≈ 100 nodes, the NCP plot slopes upward as the conductance score of the best possible set of nodes gets gradually worse and worse as those sets increase in size. Thus, if the world is modeled by a sparse "interaction graph" and if a density-based notion such as conductance is an appropriate measure of community quality, then the "best" possible "communities" in nearly every real-world network we examined gradually become less and less community-like and instead gradually "blend in" with the rest of the network, as the purported communities steadily grow in size. Although this suggests that large networks have a *core–periphery* or *jellyfish* type of structure, where small "whiskers" connect themselves into a large dense intermingled network "core," we also observed that the "core" itself has an analogous core–periphery structure.

None of the commonly used network-generation models, including preferential attachment, copying, and hierarchical models, generates networks that even qualitatively reproduce this community structure property. We found, however, that a model in which edges are added recursively, via an iterative "forest fire" burning mechanism, produces remarkably good results. Our work opens several new questions about the structure of large social and information networks in general, and it has implications for the use of graph-partitioning algorithms on real-world networks and for detecting communities in them.

Acknowledgments. We thank Reid Andersen, Christos Faloutsos, and Jon Kleinberg for discussions, Lars Backstrom for data, and Arpita Ghosh for assistance with the proof of Theorem 6.1.

A conference proceedings version of this paper appeared in WWW 2008 as [Leskovec et al. 08b], and a technical report version appeared on the arXiv as [Leskovec et al. 08a].

References

[Abello 04] J. Abello. "Hierarchical Graph Maps." *Computers and Graphics* 28:3 (2004), 345–359.

[Ahn et al. 09] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. "Link Communities Reveal Multi-scale Complexity in Networks." Preprint, arXiv:0903.3178, March 2009.

[Aiello et al. 00] W. Aiello, F. R. K. Chung, and L. Lu. "A Random Graph Model for Massive Graphs." In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pp. 171–180. New York: ACM Press, 2000.

[Aiello et al. 01] W. Aiello, F. R. K. Chung, and L. Lu. "A Random Graph Model for Power Law Graphs." *Experimental Mathematics* 10 (2001), 53–66.

[Albert and Barabási 99] R. Z. Albert and A.-L. Barabási. "Emergence of Scaling in Random Networks." *Science* 286:5439 (1999), 509–512.

[Albert and Barabási 02] R. Z. Albert and A.-L. Barabási. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74 (2002), 47–97.

[Albert et al. 99] R. Z. Albert, H. Jeong, and A.-L. Barabási. "The Diameter of the World Wide Web." *Nature* 401 (1999), 130–131.

[Allen 04] Christopher Allen. "The Dunbar Number as a Limit to Group Sizes." *Life with Alacrity.* Availalbe at http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html, 2004.

[Andersen and Lang 06] R. Andersen and K. Lang. "Communities from Seed Sets." In *Proceedings of the 15th International Conference on World Wide Web*, pp. 223–232. New York: ACM Press, 2006.

[Andersen et al. 06] R. Andersen, F. R. K. Chung, and K. Lang. "Local Graph Partitioning Using PageRank Vectors." In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 475–486. Los Alamitos, CA: IEEE Press, 2006.

[Arora and Kale 07] S. Arora and S. Kale. "A Combinatorial, Primal–Dual Approach to Semidefinite Programs." In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pp. 227–236. New York: ACM Press, 2007.

[Arora et al. 04a] S. Arora, E. Hazan, and S. Kale. "$O(\sqrt{\log n})$ Approximation to Sparsest Cut in $\tilde{O}(n^2)$ Time." In *Proceedings of the 45th Annual Symposium on Foundations of Computer Science*, pp. 238–247. Los Alamitos, CA: IEEE Press, 2004.

[Arora et al. 04b] S. Arora, S. Rao, and U. Vazirani. "Expander Flows, Geometric Embeddings and Graph Partitioning." In *roceedings of the 36th Annual ACM Symposium on Theory of Computing*, pp. 222–231. New York: ACM Press, 2004.

[Babenko et al. 07] M. Babenko, J. Derryberry, A. Goldberg, R. Tarjan, and Y. Zhou. "Experimental Evaluation of Parametric Max-Flow Algorithms." In *Experimental Algorithms: 6th International Workshop, WEA 2007, Rome, Italy, June 6–8, 2007, Proceedings*, Lecture Notes in Computer Science 4525, pp. 256–269. Berlin: Springer, 2007.

[Back 51] K. W. Back. "Influence through Social Communication." *Journal of Abnormal and Social Psychology* 46 (1951), 9–23.

[Backstrom et al. 06] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. "Group Formation in Large Social Networks: Membership, Growth, and Evolution." In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: ACM Press, pp. 44–54, 2006.

[Bagrow and Bollt 05] J. P. Bagrow and E. M. Bollt. "Local Method for Detecting Communities." *Physical Review E* 72 (2005), 046108.

[Benjamini et al. 06] I. Benjamini, G. Kozma, and N. Wormald. "The Mixing Time of the Giant Component of a Random Graph." Preprint, arXiv:math/0610459, October 2006.

[Boccaletti et al. 06] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. "Complex Networks: Structure and Dynamics." *Physics Reports* 424:4–5 (2006), 175–308.

[Boguñá et al. 09] M. Boguñá, D. Krioukov, and K. C. Claffy. "Navigability of Complex Networks." *Nature Physics* 5 (2009), 74–80.

[Bollobás 85] B. Bollobás. *Random Graphs.* London: Academic Press, 1985.

[Bollobás and Riordan 04] B. Bollobás and O. M. Riordan. "Mathematical Results on Scale-Free Random Graphs." In *Handbook of Graphs and Networks*, edited by S. Bornholdt and H. G. Schuster, pp. 1–34. New York: Wiley, 2004.

[Borgattia and Everett 00] S. P. Borgattia and M. G. Everett. "Models of Core/Periphery Structures." *Social Networks* 21:4 (2000), 375–395.

[Brandes et al. 07] U. Brandes, M. Gaertler, and D. Wagner. "Engineering Graph Clustering: Models and Experimental Evaluation." *Journal of Experimental Algorithmics* 12:1 (2007), Article no. 1.1.

[Broder et al. 00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. "Graph Structure in the Web." In *9th International World Wide Web Conference*, pp. 309–320. Amsterdam: Elsevier, 2000.

[Burer and Monteiro 03] S. Burer and R. D. C. Monteiro. "A Nonlinear Programming Algorithm for Solving Semidefinite Programs via Low-Rank Factorization." *Mathematical Programming (Series B)* 95:2 (2003), 329–357.

[Chakrabarti and Faloutsos 06] D. Chakrabarti and C. Faloutsos. "Graph Mining: Laws, Generators, and Algorithms." *ACM Computing Surveys* 38:1 (2006), Article no. 2.

[Chakrabarti et al. 04] D. Chakrabarti, Y. Zhan, and C. Faloutsos. "R-MAT: A Recursive Model for Graph Mining." In *SDM '04: Proceedings of the 4th SIAM International Conference on Data Mining*, pp. 442–446. Philadelphia: SIAM, 2004.

[Chakrabarti et al. 07] D. Chakrabarti, C. Faloutsos, and Y. Zhan. "Visualization of Large Networks with Min-Cut Plots, A-Plots and R-MAT." *International Journal of Human-Computer Studies* 65:5 (2007), 434–445 .

[Cheeger 69] J. Cheeger. "A Lower Bound for the Smallest Eigenvalue of the Laplacian." In *Problems in Analysis: A Symposium in Honor of Salomon Bochner*, edited by Robert C. Gunning, pp. 195–199. Princeton: Princeton University Press, 1969.

[Cherkassky and Goldberg 95] B. V. Cherkassky and A. V. Goldberg. "On Implementing Push–Relabel Method for the Maximum Flow Problem." In *Integer Programming and Combinatorial Optimization: 4th International IPCO Conference Copenhagen, Denmark, May 29-31, 1995, Proceedings*, Lecture Notes in Computer Science 920, pp. 157–171. Berlin: Springer, 1995.

[Chung 97] F. R. K. Chung. *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics 92. Providence: American Mathematical Society, 1997.

[Chung 07a] F. R. K Chung. "Four Proofs of Cheeger Inequality and Graph Partition Algorithms." Paper presented at the Fourth International Congress of Chinese Mathematicians (ICCM), Hangzhou, China, December 17–22, 2007.

[Chung 07b] F. R. K. Chung. "The Heat Kernel as the Pagerank of a Graph." *Proceedings of the National Academy of Sciences* 104:50 (2007), 19735–19740.

[Chung 07c] F. R. K. Chung. "Random Walks and Local Cuts in Graphs." *Linear Algebra and Its Applications* 423 (2007), 22–32.

[Chung and Lu 01] F. R. K. Chung and L. Lu. "The Diameter of Sparse Random Graphs." *Advances in Applied Mathematics* 26:4 (2002), 257–279.

[Chung and Lu 02a] F. R. K. Chung and L. Lu. "The Average Distances in Random Graphs with Given Expected Degrees." *Proceedings of the National Academy of Sciences* 99:25 (2002), 15879–15882.

[Chung and Lu 02b] F. R. K. Chung and L. Lu. "Connected Components in Random Graphs with Given Expected Degree Sequences." *Annals of Combinatorics* 6:2 (2002), 125–145.

[Chung and Lu 03] F. R. K. Chung and L. Lu. "The Average Distances in a Random Graph with Given Expected Degrees." *Internet Mathematics* 1 (2003), 91–113.

[Chung and Lu 06a] F. R. K. Chung and L. Lu. *Complex Graphs and Networks*, CBMS Regional Conference Series in Mathematics 107. Providence: American Mathematical Society, 2006.

[Chung and Lu 06b] F. R. K. Chung and L. Lu. "The Volume of the Giant Component of a Random Graph with Given Expected Degrees." *SIAM Journal on Discrete Mathematics* 20 (2006), 395–411.

[Chung et al. 03a] F. R. K. Chung, L. Lu, and V. Vu. "Eigenvalues of Random Power Law Graphs." *Annals of Combinatorics* 7 (2003), 21–33.

[Chung et al. 03b] F. R. K. Chung, L. Lu, and V. Vu. "The Spectra of Random Graphs with Given Expected Degrees." *Proceedings of the National Academy of Sciences* 100:11 (2003), 6313–6318.

[Chung et al. 04] F. R. K. Chung, L. Lu, and V. Vu. "The Spectra of Random Graphs with Given Expected Degrees." *Internet Mathematics* 1 (2004), 257–275.

[Clauset 04] A. Clauset. "Finding Local Community Structure in Networks." *Physical Review E* 72 (2005), 026132.

[Clauset et al. 04] A. Clauset, M. E. J. Newman, and C. Moore. "Finding Community Structure in Very Large Networks." Preprint, arXiv:cond-mat/0408187, August 2004.

[Clauset et al. 06] A. Clauset, C. Moore, and M. E. J. Newman. "Structural Inference of Hierarchies in Networks." Preprint, arXiv:physics/0610051, October 2006.

[Clauset et al. 07] A. Clauset, C. R. Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data." Preprint, arXiv:0706.1062, June 2007.

[Colizza et al. 05] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. "Characterization and Modeling of Protein–Protein Interaction Networks." *Physica A Statistical Mechanics and Its Applications* 352 (2005), 1–27.

[Costa et al. 07] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. "Characterization of Complex Networks: A Survey of Measurements." *Advances in Physics* 56:1 (2007), 167–242.

[da Silva et al. 08] M. R. da Silva, H. Ma, and A.-P. Zeng. "Centrality, Network Capacity, and Modularity as Parameters to Analyze the Core-Periphery Structure in Metabolic Networks." *Proceedings of the IEEE* 96:8 (2008), 1411–1420.

[Danon et al. 05] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. "Comparing Community Structure Identification." *Journal of Statistical Mechanics: Theory and Experiment* 29:09 (2005), P09008.

[Dhillon et al. 07] I. S. Dhillon, Y. Guan, and B. Kulis. "Weighted Graph Cuts without Eigenvectors: A Multilevel Approach." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29:11 (2007), 1944–1957.

[Donath and Hoffman 72] W. E. Donath and A. J. Hoffman. "Algorithms for Partitioning Graphs and Computer Logic Based on Eigenvectors of Connection Matrices." *IBM Technical Disclosure Bulletin* 15:3 (1972), 938–944.

[Dorogovtsev and Mendes 02] S. N. Dorogovtsev and J. F. F. Mendes. "Evolution of Networks." *Advances in Physics* 51 (2002), 1079–1187.

[Dorogovtsev et al. 06] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. "k-Core Organization of Complex Networks." *Physical Review Letters* 96:4 (2006), 040601.

[Doyle and Carlson 00] J. Doyle and J. M. Carlson. "Power Laws, Highly Optimized Tolerance, and Generalized Source Coding." *Physical Review Letters* 84:24 (2000), 5656–5659.

[Doyle and Carlson 02] J. Doyle and J. M. Carlson. "Complexity and Robustness." *Proceedings of the National Academy of Sciences* 99 (2002), 2538–2545.

[Dunbar 98] R. Dunbar. *Grooming, Gossip, and the Evolution of Language*. Cambridge, MA: Harvard University Press, 1998.

[Estrada 06] E. Estrada. "Spectral Scaling and Good Expansion Properties in Complex Networks." *Europhysics Letters* 73 (2006), 649–655.

[Estrada 07] E. Estrada. "Topological Structural Classes of Complex Networks." *Physical Review E* 75 (2007), 016103.

[Estrada and Hatano 09] E. Estrada and N. Hatano. "Communicability Graph and Community Structures in Complex Networks." Preprint, arXiv:0905.4103, May 2009.

[Evans and Lambiotte 09] T. S. Evans and R. Lambiotte. "Line Graphs, Link Partitions, and Overlapping Communities." *Physical Review E* 80 (2009), 016105.

[Fabrikant et al. 02] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. "Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet. In *Automata, Languages and Programming: 29th International Colloquium, ICALP 2002 Málaga, Spain, July 8-13, 2002, Proceedings*, Lecture Notes In Computer Science 2380, pp. 110–122. Berlin: Springer, 2002.

[Faloutsos et al. 99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. "On Power-Law Relationships of the Internet Topology." In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pp. 251–262. New York: ACM Press, 1999.

[Feige and Ofek 05] U. Feige and E. Ofek. "Spectral Techniques Applied to Sparse Random Graphs." *Random Structures and Algorithms* 27 (2005), 251–275.

[Fernholz and Ramachandran 07] D. Fernholz and V. Ramachandran. "The Diameter of Sparse Random Graphs." *Random Structures and Algorithms* 31 (2007), 482–516.

[Fiduccia and Mattheyses 82] C. M. Fiduccia and R. M. Mattheyses. "A Linear-Time Heuristic for Improving Network Partitions." In *Proceedings of the 19th Design Automation Conference*, pp. 175–181. Piscataway, NJ: IEEE Press, 1982.

[Fiedler 73] M. Fiedler. "Algebraic Connectivity of Graphs." *Czechoslovak Mathematical Journal* 23:98 (1973), 298–305.

[Flake et al. 00] G. W. Flake, S. Lawrence, and C. L. Giles. "Efficient Identification of Web Communities." In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–160. New York: ACM Press, 2000.

[Flake et al. 02] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. "Self-Organization and Identification of Web Communities." *Computer* 35:3 (2002), 66–71.

[Flake et al. 03] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis. "Graph Clustering and Minimum Cut Trees." *Internet Mathematics* 1:4 (2003), 385–408.

[Flaxman et al. 04] A. D. Flaxman, A. M. Frieze, and J. Vera. "A Geometric Preferential Attachment Model of Networks." In *Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, Rome, Italy, October 16, 2004, Proceeedings*, Lecture Notes in Computer Science 3243, pp. 44–55. Berlin: Springer, 2004.

[Flaxman et al. 05] A. Flaxman, A. Frieze, and T. Fenner. "High Degree Vertices and Eigenvalues in the Preferential Attachment Graph." *Internet Mathematics* 2:1 (20050, 1–19.

[Flaxman et al. 07] A. D. Flaxman, A. M. Frieze, and J. Vera. "A Geometric Preferential Attachment Model of Networks II." In *Algorithms and Models for the Web-Graph: 5th International Workshop, WAW 2007, San Diego, CA, USA, December 11–12, 2007, Proceeedings*, Lecture Notes in Computer Science 4863, pp. 41–55. Berlin: Springer, 2007.

[Fortunato 10] S. Fortunato. "Community Detection in Graphs." *Physics Reports* 486 (2010), 75–174.

[Fortunato and Barthélemy 07] S. Fortunato and M. Barthélemy. "Resolution Limit in Community Detection." *Proceedings of the National Academy of Sciences* 104:1 (2007), 36–41.

[Fountoulakis and Reed 07a] N. Fountoulakis and B. Reed. "The Evolution of the Mixing Rate." Preprint, arXiv:math/0701474, January 2007.

[Fountoulakis and Reed 07b] N. Fountoulakis and B.A. Reed. "Faster Mixing and Small Bottlenecks." *Probability Theory and Related Fields* 137:3–4 (2007), 475–486.

[Gaertler 05] M. Gaertler. "Clustering." In *Network Analysis: Methodological Foundations*, edited by U. Brandes and T. Erlebach, pp. 178–215. New York: Springer, 2005.

[Gallo et al. 89] G. Gallo, M. D. Grigoriadis, and R. E. Tarjan. "A Fast Parametric Maximum Flow Algorithm and Applications." *SIAM Journal on Computing* 18:1 (1989), 30–55.

[Gehrke et al. 03] J. Gehrke, P. Ginsparg, and J. Kleinberg. "Overview of the 2003 KDD Cup." *SIGKDD Explorations* 5:2 (2003), 149–151.

[Gibson et al. 98] D. Gibson, J. Kleinberg, and P. Raghavan. "Inferring Web Communities from Link Topology." In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pp. 225–234. New York: ACM Press, 1998.

[Girvan and Newman 02] M. Girvan and M. E. J. Newman. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99:12 (2002), 7821–7826.

[Gkantsidis et al. 03] C. Gkantsidis, M. Mihail, and A. Saberi. "Conductance and Congestion in Power Law Graphs." In *Proceedings of the 2003 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 148–159. New York: ACM Press, 2003.

[Glasgow 00] University of Glasgow. "TREC Web Corpus: WT10g." Available at http://ir.dcs.gla.ac.uk/test_collections/wt10g.html, 2000.

[Goldberg and Rao 98] A. V. Goldberg and S. Rao. "Beyond the Flow Decomposition Barrier." *Journal of the ACM* 45 (1998), 783–797.

[Goldberg and Tarjan 88] A. V. Goldberg and R. E. Tarjan. "A New Approach to the Maximum-Flow Problem." *Journal of the ACM* 35 (1988), 921–940.

[Google 02] Google, Inc. "Google Programming Contest." Available at http://www.google.com/programming-contest/, 2002.

[Guattery and Miller 98] S. Guattery and G. L. Miller. "On the Quality of Spectral Separators." *SIAM Journal on Matrix Analysis and Applications* 19 (1998), 701–719.

[Guimerà et al. 04] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. "Modularity from Fluctuations in Random Graphs and Complex Networks." *Physical Review E* 70 (2004), 025101.

[Gulbahce and Lehmann 08] N. Gulbahce and S. Lehmann. "The Art of Community Detection." *BioEssays* 30:10 (2008), 934–938.

[Hastings 06] M. B. Hastings. "Community Detection as an Inference Problem." *Physical Review E* 74 (2006), 035102.

[Hendrickson and Leland 95] B. Hendrickson and R. Leland. "A Multilevel Algorithm for Partitioning Graphs." In *Proceedings of the 1995 ACM/IEEE Conference on Supercomputing (CDROM)*, Article no. 28. New York: ACM Press, 1995.

[Holme 05] P. Holme. "Core–Periphery Organization of Complex Networks." *Physical Review E* 72 (2005), 046111.

[Hoory et al. 06] S. Hoory, N. Linial, and A. Wigderson. "Expander Graphs and Their Applications." *Bulletin of the American Mathematical Society* 43 (2006), 439–561.

[Hopcroft et al. 03] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. "Natural Communities in Large Linked Networks." In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 541–546. New York: ACM Press, 2003.

[Hopcroft et al. 04] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. "Tracking Evolving Communities in Large Linked Networks." *Proceedings of the National Academy of Sciences* 101 (2004), 5249–5253.

[Jain et al. 99] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data Clustering: A Review." *ACM Computing Surveys* 31 (1999), 264–323.

[Jeong et al. 01] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. "Lethality and Centrality in Protein Networks." *Nature* 411 (2001), 41–42.

[Kannan et al. 04] R. Kannan, S. Vempala, and A. Vetta. "On Clusterings: Good, Bad and Spectral." *Journal of the ACM* 51:3 (2004), 497–515.

[Karrer et al. 07] B. Karrer, E. Levina, and M. E. J. Newman. "Robustness of Community Structure in Networks." Preprint, arXiv:0709.2108, September 2007.

[Karypis and Kumar 98a] G. Karypis and V. Kumar. "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs." *SIAM Journal on Scientific Computing* 20 (1998), 359–392.

[Karypis and Kumar 98b] G. Karypis and V. Kumar. "Multilevel k-Way Partitioning Scheme for Irregular Graphs." *Journal of Parallel and Distributed Computing* 48 (1998), 96–129.

[Kernighan and Lin 70] B. Kernighan and S. Lin. "An Effective Heuristic Procedure for Partitioning Graphs." *The Bell System Technical Journal* 49 (1970), 291–308.

[Khalil and Liu 04] A. Khalil and Y. Liu. "Experiments with PageRank Computation." Technical Report 603, Indiana University Department of Computer Science, December 2004.

[Khandekar et al. 06] R. Khandekar, S. Rao, and U. Vazirani. "Graph Partitioning Using Single Commodity Flows." In *Proceedings of the 38th annual ACM Symposium on Theory of Computing*, pp. 385–390. New York: ACM Press, 2006.

[Klimt and Yang 04] B. Klimt and Y. Yang. "Introducing the Enron Corpus," *First Conference on Email and Anti-Spam (CEAS) 2004 Proceedings*. Available at http://ceas.cc/2004/168.pdf, 2004.

[Kumar et al. 99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "Trawling the Web for Emerging Cyber-communities." *Computer Networks* 31:11 (1999), 1481–1493.

[Kumar et al. 00] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. "Stochastic Models for the Web Graph." In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pp. 57–65. Washington, DC: IEEE Computer Society, 2000.

[Kumar et al. 06] R. Kumar, J. Novak, and A. Tomkins. "Structure and Evolution of Online Social Networks." In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 611–617. New York: ACM Press, 2006.

[Lancichinetti and Fortunato 09] A. Lancichinetti and S. Fortunato. "Community Detection Algorithms: A Comparative Analysis." *Physical Review E* 80 (2009), 056117.

[Lang 04] K. Lang. "Finding Good Nearly Balanced Cuts in Power Law Graphs." Technical Report YRL-2004-036, Yahoo! Research Labs, Pasadena, CA, November 2004.

[Lang and Rao 93] K. Lang and S. Rao. "Finding Near-Optimal Cuts: An Empirical Evaluation." In *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 212–221. Philadelphia: SIAM, 1993.

[Lang and Rao 04] K. Lang and S. Rao. "A Flow-Based Method for Improving the Expansion or Conductance of Graph Cuts." In *Integer Programming and Combinatorial Optimization: 10th International IPCO Conference, New York, NY, USA, June 7–11, 2004. Proceedings*, Lecture Notes in Computer Science 3064, pp. 325–337. Berlin: Springer, 2004.

[Leighton and Rao 88] T. Leighton and S. Rao. "An Approximate Max-Flow Min-Cut Theorem for Uniform Multicommodity Flow Problems with Applications to Approximation Algorithms." In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pp. 422–431. Washington, DC: IEEE Computer Society, 1988.

[Leighton and Rao 99] T. Leighton and S. Rao. "Multicommodity Max-Flow Min-Cut Theorems and Their Use in Designing Approximation Algorithms." *Journal of the ACM* 46:6 (1999), 787–832.

[Leskovec and Faloutsos 07] J. Leskovec and C. Faloutsos. "Scalable Modeling of Real Graphs Using Kronecker Multiplication." In *Proceedings of the 24th International Conference on Machine Learning*, pp. 497–504. New York: ACM Press, 2007.

[Leskovec et al. 05a] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. "Realistic, Mathematically Tractable Graph Generation and Evolution, Using Kronecker Multiplication." In *Knowledge Discovery in Databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3–7, 2005, Proceedings*, Lecture Notes in Computer Science 3721, pp. 133–145. Berlin: Springer, 2005.

[Leskovec et al. 05b] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." In *Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 177–187. New York: ACM Press, 2005.

[Leskovec et al. 07a] J. Leskovec, L. A. Adamic, and B. A. Huberman. "The Dynamics of Viral Marketing." *ACM Transactions on the Web* 1:1 (2007), Article no. 5.

[Leskovec et al. 07b] J. Leskovec, J. Kleinberg, and C. Faloutsos. "Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data* 1:1 (2007), Article no. 2.

[Leskovec et al. 07c] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. "Patterns of Cascading Behavior in Large Blog Graphs." In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 551-556. Philadelphia: SIAM, 2007.

[Leskovec et al. 08a] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters." Preprint, arXiv:0810.1355, October 2008.

[Leskovec et al. 08b] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. "Statistical Properties of Community Structure in Large Social and Information Networks." In *Proceedings of the 17th International Conference on World Wide Web*, pp. 695–704. New York: ACM Press, 2008.

[Leskovec et al. 10a] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. "Kronecker Graphs: An Approach to Modeling Networks." *Journal of Machine Learning Research* 11 (2010), 985–1042.

[Leskovec et al. 10b] J. Leskovec, K. J. Lang, and M. W. Mahoney. "Empirical Comparison of Algorithms for Network Community Detection." In *Proceedings of the 19th International Conference on World Wide Web*, pp. 631–640. New York: ACM Press, 2010.

[Li et al. 06] L. Li, J. C. Doyle, and W. Willinger. "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications." *Internet Mathematics* 2:4 (2006), 431–523.

[Lu 01] L. Lu. "The Diameter of Random Massive Graphs." In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 912–921. Philadelphia: SIAM, 2001.

[Lusseau et al. 03] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. "The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations." *Behavioral Ecology and Sociobiology* 54 (2003), 396–405.

[Mihail and Papadimitriou 02] M. Mihail and C. H. Papadimitriou. "On the Eigenvalue Power Law." In *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, pp. 254–262. Berlin: Springer, 2002.

[Mihail et al. 06] M. Mihail, C. H. Papadimitriou, and A. Saberi. "On Certain Connectivity Properties of the Internet Topology." *Journal of Computer and System Sciences* 72:2 (2006), 239–251.

[Milo et al. 04] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. "On the Uniform Generation of Random Graphs with Prescribed Degree Sequences." Preprint, arXiv:cond-mat/0312028v2, May 2004.

[Mohar 91] B. Mohar. "The Laplacian Spectrum of Graphs." In *Graph Theory, Combinatorics, and Applications, Vol. 2*, edited by Y. Alavi, G. Chartrand, O. R. Oellermann, and A .J. Schwenk, pp. 871–898. New York: Wiley, 1991.

[Molloy and Reed 95] M. Molloy and B. Reed. "A Critical Point for Random Graphs with a Given Degree Sequence." *Random Structures and Algorithms* 6 (1995), 161–180.

[Molloy and Reed 98] M. Molloy and B. Reed. "The Size of the Giant Component of a Random Graph with a Given Degree Sequence." *Combinatorics, Probability and Computing* 7 (1998), 295–305.

[Montgomery and Faloutsos 01] A.L. Montgomery and C. Faloutsos. "Identifying Web Browsing Trends and Patterns." *Computer* 34:7 (2001), 94–95.

[Netflix 09] Netflix, Inc. "Netflix Prize." Available at http://www.netflixprize.com/, 2009.

[Newman 03] M. E. J. Newman. "The Structure and Function of Complex Networks." *SIAM Review* 45 (2003), 167–256.

[Newman 04] M. E. J. Newman. "Detecting Community Structure in Networks." *The European Physical Journal B* 38 (2004), 321–330.

[Newman 05] M. E. J. Newman. "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physics* 46 (2005), 323–351.

[Newman 06a] M. E. J. Newman. "Finding Community Structure in Networks Uusing the Eigenvectors of Matrices." *Physical Review E* 74 (2006), 036104.

[Newman 06b] M. E. J. Newman. "Modularity and Community Structure in Networks." *Proceedings of the National Academy of Sciences* 103:23 (2006), 8577–8582.

[Newman 09] Mark Newman. "Network Data." Available at http://www-personal.umich.edu/~mejn/netdata/, 2009.

[Newman and Girvan 04] M. E. J. Newman and M. Girvan. "Finding and Evaluating Community Structure in Networks." *Physical Review E* 69 (2004), 026113.

[Newman and Leicht 07] M. E. J. Newman and E. A. Leicht. "Mixture Models and Exploratory Analysis in Networks." *Proceedings of the National Academy of Sciences* 104:23 (2007), 9564–9569.

[Oregon 97] University of Oregon Advanced Network Technology Center. "University of Oregon Route Views Project: Online Data and Reports." Available at http://www.routeviews.org, 1997.

[Palla et al. 05] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society." *Nature* 435:7043 (2005), 814–818.

[Qi et al. 06] Y. Qi, J. K. Seetharaman, and Z. B. Joseph. "Random Forest Similarity for Protein–Protein Interaction Prediction from Multiple Sources." In *Pacific Symposium on Biocomputing*, pp. 531–542. Ridge Edge, NJ: World Scientific, 2005.

[Radicchi et al. 04] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. "Defining and Identifying Communities in Networks." *Proceedings of the National Academy of Sciences* 101:9 (2004), 2658–2663.

[Raghavan et al. 07] U. Nandini Raghavan, R. Z. Albert, and S. Kumara. "Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks." *Physical Review E* 76 (2007), 036106.

[Ravasz and Barabási 03] E. Ravasz and A.-L. Barabási. "Hierarchical Organization in Complex Networks." *Physical Review E* 67 (2003), 026112.

[Ravasz et al. 02] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. "Hierarchical Organization of Modularity in Metabolic Networks." *Science* 297:5586 (2002), 1551–1555.

[Reichardt and Bornholdt 07] J. Reichardt and S. Bornholdt. "Partitioning and Modularity of Graphs with Arbitrary Degree Distribution." *Physical Review E* 76 (2007), 015102.

[Ren et al. 07] Y. Ren, R. Kraut, and S. Kiesler. "Applying Common Identity and Bond Theory to Design of Online Communities." *Organization Studies* 28:3 (2007), 377–408.

[Richardson 03] M. Richardson, R. Agrawal, and P. Domingos. "Trust Management for the Semantic Web." In *The SemanticWeb—ISWC 2003*, Lecture Notes in Computer Science 2870, pp. 351–368. Berlin: Springer, 2003.

[Ripeanu et al. 02] M. Ripeanu, I. Foster, and A. Iamnitchi. "Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design." *IEEE Internet Computing* 6:1 (2002), 50–57.

[Rosvall and Bergstrom 07] M. Rosvall and C. T. Bergstrom. "An Information-Theoretic Framework for Resolving Community Structure in Complex Networks." *Proceedings of the National Academy of Sciences* 104:18 (2007), 7327–7331.

[Rosvall and Bergstrom 08] M. Rosvall and C. T. Bergstrom. "Maps of Random Walks on Complex Networks Reveal Community Structure." *Proceedings of the National Academy of Sciences* 105:4 (2008), 1118–1123.

[Sampson 68] S. F. Sampson. "A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships." PhD thesis, Cornell University Deptartment of Sociology, 1968.

[Schaeffer 07] S. E. Schaeffer. "Graph Clustering." *Computer Science Review* 1:1 (2007), 27–64.

[Shi and Malik 00] J. Shi and J. Malik. "Normalized Cuts and Image Segmentation." *IEEE Transcations of Pattern Analysis and Machine Intelligence* 22:8 (2000), 888–905.

[Siganos et al. 06] G. Siganos, S. L. Tauro, and M. Faloutsos. "Jellyfish: A Conceptual Model for the Internet Topology." *Journal of Communications and Networks* 8 (2006), 339–350.

[Spielman and Teng 96] D. A. Spielman and S.-H. Teng. "Spectral Partitioning Works: Planar Graphs and Finite Element Meshes." In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, pp. 96–107. Los Alamitos, CA: IEEE Press, 1996.

[Spielman and Teng 04a] D. A. Spielman and S.-H. Teng. "Nearly-Linear Time Algorithms for Graph Partitioning, Graph Sparsification, and Solving Linear Systems." In *Proceedings of the 36th annual ACM Symposium on Theory of Computing* pp. 81–90. New York: ACM Press, 2004.

[Spielman and Teng 04b] D. A. Spielman and S.-H. Teng. "Nearly-Linear Time Algorithms for Graph Partitioning, Graph Sparsification, and Solving Linear Systems." Preprint, arXiv:cs/0310051v9, March 2004.

[Tauro et al. 01] S. L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. "A Simple Conceptual Model for the Internet Topology." In *GLOBECOM '01: IEEE Global Telecommunications Conference*, pp. 1667–1671. Los Alamitos, Ca: IEEE Press, 2001.

[Tenenbaum et al. 00] J. B. Tenenbaum, V. de Silva, and J. C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction." *Science* 290:5500 (2000), 2319–2323.

[von Luxburg 06] U. von Luxburg. "A Tutorial on Spectral Clustering." Technical Report 149, Max Plank Institute for Biological Cybernetics, August 2006.

[Wang et al. 09] P. Wang, M. C. González, C. A. Hidalgo, and A.-L. Barabási. "Understanding the Spreading Patterns of Mobile Phone Viruses." *Science* 324:5930 (2009), 1071–1076.

[Wasserman and Faust 94] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge, UK: Cambridge University Press, 1994.

[Watts and Strogatz 98] D. J. Watts and S. H. Strogatz. "Collective Dynamics of Small-World Networks." *Nature* 393 (1998), 440–442.

[White and Smyth 05] S. White and P. Smyth. "A Spectral Clustering Approach to Finding Communities in Graphs." In *SDM '05: Proceedings of the 5th SIAM International Conference on Data Mining*, pp. 76–84. Philadelphia: SIAM, 2005.

[Wu and Huberman 04] F. Wu and B. A. Huberman. "Finding Communities in Linear Time: A Physics Approach." *The European Physical Journal B* 38:2 (2004), 331–338.

[Xuan et al. 06] Q. Xuan, Y. Li, and T.-J. Wu. "Growth Model for Complex Networks with Hierarchical and Modular Structures." *Physical Review E* 73 (2006), 036105.

[Yang et al. 09] T. Yang, R. Jin, Y. Chi, and S. Zhu. "Combining Link and Content for Community Detection: A Discriminative Approach." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 927–936. New York: ACM Press, 2009.

[Zachary 77] W. W. Zachary. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 33 (1977), 452–473.

[Zahn 71] C. T. Zahn. "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters." *IEEE Transactions on Computers* C-20:1 (1971), 68–86.

[Zhao and Karypis 04] Y. Zhao and G. Karypis. "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering." *Machine Learning* 55 (2004), 311–331.

[Zinoviev 08] D. Zinoviev. "Topology and Geometry of Online Social Networks." Preprint, arXiv:0807.3996, July 2008.

[Zinoviev and Duong 09] D. Zinoviev and V. Duong. "Toward Understanding Friendship in Online Social Networks." Preprint, arXiv:0902.4658, February 2009.

Jure Leskovec, Computer Science Department, Stanford University, Stanford, CA 94305 (jure@cs.stanford.edu)

Kevin J. Lang, Yahoo! Research, 701 1st Ave., Sunnyvale, CA 94089 (langk@yahoo-inc.com)

Anirban Dasgupta, Yahoo! Research, 701 1st Ave., Sunnyvale, CA 94089 (anirban@yahoo-inc.com)

Michael W. Mahoney, Computer Science Department, Stanford University, Stanford, CA 94305 (mmahoney@cs.stanford.edu)

Received May 13, 2008; accepted February 24, 2010.