Statistical Properties of Community Structure in Large Social and Information Networks

Jure Leskovec* Kevin J. Lang† Anirban Dasgupta† Michael W. Mahoney†
*Carnegie Mellon University †Yahoo! Research
jure@cs.cmu.edu {langk, anirban, mahoney}@yahoo-inc.com

ABSTRACT

A large body of work has been devoted to identifying community structure in networks. A community is often though of as a set of nodes that has more connections between its members than to the remainder of the network. In this paper, we characterize as a function of size the statistical and structural properties of such sets of nodes. 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, and we study over 70 large sparse real-world networks taken from a wide range of application domains. Our results suggest a significantly more refined picture of community structure in large real-world networks than has been appreciated previously.

Our most striking finding is that in nearly every network dataset we examined, we observe tight but almost trivial communities at very small scales, and at larger size scales, the best possible communities gradually "blend in" with the rest of the network and thus become less "community-like." This behavior is not explained, even at a qualitative level, by any of the commonly-used network generation models. Moreover, this behavior is exactly the opposite of what one would expect based on experience with and intuition from expander graphs, from graphs that are well-embeddable in a low-dimensional structure, and from small social networks that have served as testbeds of community detection algorithms. We have found, however, that a generative model, in which new edges are added via an iterative "forest fire" burning process, is able to produce graphs exhibiting a network community structure similar to our observations.

Categories and Subject Descriptors: H.2.8 Database Management: Database applications – Data mining

General Terms: Measurement; Experimentation.

Keywords: Social networks; Graph partitioning; Community structure; Conductance; Random walks.

1. INTRODUCTION

In this paper, we explore from a novel perspective several questions related to identifying meaningful communities in social and information networks, and we come to several surprising conclusions that have theoretical and practical implications for community detection.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

1.1 Overview of our approach

At the risk of oversimplifying the large 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 kind of interaction between pairs of those entities. For example, 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 amongst 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) A objective function or metric is chosen to formalize this idea of groups with more intra-group than inter-group 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 (communities) are then 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 social 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 will follows the usual path in this paper. For point (3), we choose a natural and widely-adopted notion of community goodness called *conductance*, also known as the normalized cut metric [6, 31, 16]. Since there exist a rich suite of both theoretical and practical algorithms to optimize this quantity [32, 20, 4, 17, 37, 10], we can for point (4) compare and contrast several methods to approximately optimize it.

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 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 communities 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*. Intuitively, the network community profile plot measures the quality of "best" community as a function of community size in a network. To measure the quality of a community we use *conductance* [6]. By this measure, the best communities are densely linked sets of nodes attached to the rest of the network via few edges. Fig. 1(a) gives a typical NCP plot.

We compare our results across over 70 large social and information networks, numerous commonly-studied small social networks, and also expanders and low-dimensional mesh-like objects. We also 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. 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.

1.2 Summary of our results

Main Empirical Findings: Our results suggest a rather detailed and somewhat counterintuitive picture of the community structure in large networks. Several qualitative properties of community structure are nearly universal:

- Up to a size scale, which empirically is roughly 100 nodes, there not only exist well-separated communities, but also the slope of the network community profile plot is generally sloping downward. (See Fig. 1(a).) This latter point suggests, and empirically we often observe, that smaller communities can be combined into meaningful larger communities.
- At size scale of 100 nodes, we often observe the global minimum of the network community profile plot. (Although these are the "best" communities in the entire graph, they are usually connected to the remainder of the network by just a single edge.)
- Above the size scale of roughly 100 nodes, the network community profile plot gradually increases, and thus there is a nearly *inverse* relationship between community size and community quality. (See Fig. 1(a).) This upward slope suggests, and empirically we often observe, that as a function of increasing size, the best possible communities as they grow become more and more "blended into" the remainder of the network.

This last point is particularly significant, and it is our main empirical finding: at larger and larger size scales 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 (less well-expressed/separated). 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. This is important if one asserts that cut and density based intuitions will find "true" communities.

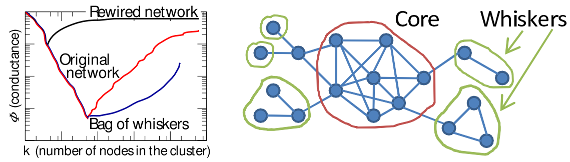


Figure 1: (a) Typical NCP plot. (b) Network structure as suggested by our experiments.

We have also examined in detail the structure of our social and information networks. We have observed that an "jellyfish" or "octopus" model [33, 7] provides a rough first approximation to 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 which contains a constant fraction of the nodes, and then there are a large number of relatively small "whiskers" that are only tenuously connected to the core. (See Fig. 1(b).)

Main Modeling Results: The observed properties of the network community profile plot are 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, copying, and hierarchical network models. Moreover, this behavior is qualitatively different than what is observed in networks with an underlying mesh-like or manifold-like geometry (which is significant 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 often used as testbeds for community detection algorithms (which may have implications for the applicability of these methods to detect large community-like structures in 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.

We, however, make the following modeling 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 a consequence of the extreme sparsity of the data.
- A "forest fire" generative model [21], 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.

Intuitively, the structure of the *whiskers* (See Fig. 1(b).), which are not unlike small social networks that have been extensively studied, are 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 helps to determine the upward part of the network community profile plot.

• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [5]
Epinions	75,877	405,739	Trust network [28]
CA-DBLP	317,080	1,049,866	Co-authorship [5]
• Information (citation) networks			
Cit-hep-th	27,400	352,021	Arxiv `hep-th` [14]
AmazonProd	524,371	1,491,793	Amazon products [8]
• Web graphs			
Web-google	855,802	4,291,352	Google web graph
Web-wt10g	1,458,316	6,225,033	TREC WT10G
• Bipartite affiliation (authors-to-papers) networks			
Atp-DBLP	615,678	944,456	DBLP [21]
AtM-Imdb	2,076,978	5,847,693	Actors-to-movies
• Internet networks			
AsSkitter	1,719,037	12,814,089	Autonom. sys.
Gnutella	62,561	147,878	P2P network [29]

Table 1: Some of the network datasets we studied.

2. BACKGROUND AND OVERVIEW

In this section, we will provide background on our data and methods. There exist a large number of reviews on topics related to those discussed in this paper. For example, see the reviews on community identification [24, 9], graph and spectral clustering [13, 30], and the monographs on spectral graph theory and complex networks [6, 7].

2.1 Network datasets

We have examined a large number of real-world complex networks. Table 1 gives a subset of the networks that we use in this paper. (We refer to the extended version of the paper [23] for a complete list of networks.) In all cases, we consider networks as undirected, and we extract the largest connected component. We have grouped the networks into 5 categories: social networks which consist of on-line social networks and co-authorship networks of computer science (DBLP) and various areas of physics; information networks which contain citation networks of physics and blogosphere; web-graphs which contain networks with nodes representing web-pages and hyperlinks being the edges; bipartite social affiliation networks which contain mainly authors-to-papers networks of computer science and physics; and finally, internet networks which consist of autonomous systems network and Gnutella P2P file sharing network.

Table 1 also shows the number of nodes and edges in each network. The sizes of the networks we have studied 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 [23]. In addition, 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 a network of movie ratings from Netflix [23]—and most of the other networks, including the purely social networks, have average degree around 10 (median degree of 6). In total, we have examined over 100 different networks, including over 70 large real-world social and information networks, making this, to our knowledge, the largest and most comprehensive study of such networks. (We will make data available via a link from the first author's web page.)

2.2 Clusters and communities in networks

If $G = (V, E)$ denotes 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 degrees of 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, if A is the adjacency matrix of the graph G, then:

$$\phi(S) = \frac{\sum_{i\in S, j\notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}} \tag{6}$$

where $A(S) = \sum_{i\in S}\sum_{j\in V} A_{ij}$, in which case the conductance of the graph G is

$$\phi_G = \min_{S\subset V} \phi(S). \tag{7}$$

Thus, the conductance of a set provides a measure for the quality of the cut $(S, \overline{S})$, or relatedly the goodness of a community S. Indeed, it is often noted that communities should be thought of as sets of nodes with more and/or better intra-connections than inter-connections. When 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 how community-like is a set of nodes, it is commonly noted—*e.g.*, see [31] and [16]—that conductance captures the "gestalt" notion of clustering [36], and so it has been widely-used for graph clustering and community detection [13, 30].

3. 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.

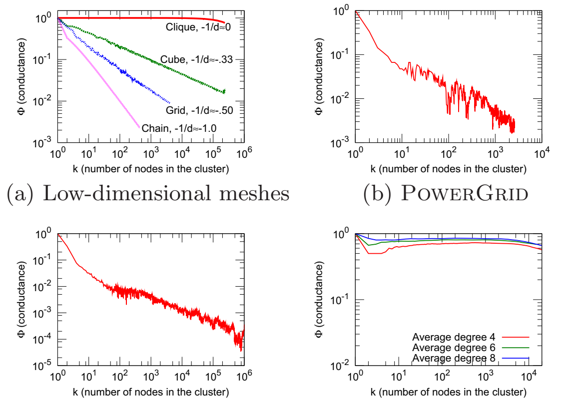
3.1 The network community profile plot

In order to resolve more finely community structure in large networks, 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 size of the purported community. 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. That is,

$$\Phi(k) = \min_{S\subset V, |S|=k} \phi(S). \tag{8}$$

where $|S|$ denotes the cardinality of the set S and where the conductance $\phi(S)$ of S is given by (6). Since this quantity is intractable to compute, we employ well-studied approximation algorithms for the Minimum Conductance Cut Problem to compute different approximations to the NCP plot. We employ two procedures: first, Metis+MQI, *i.e.*, the graph partitioning package Metis [17] followed by the flow-based MQI post-processing procedure MQI [19], which taken together returns sets that have very good conductance values; and second, the Local Spectral Algorithm [3], which returns sets that are somewhat "regularized" (more internally "coherent") but that often have worse conductance values.

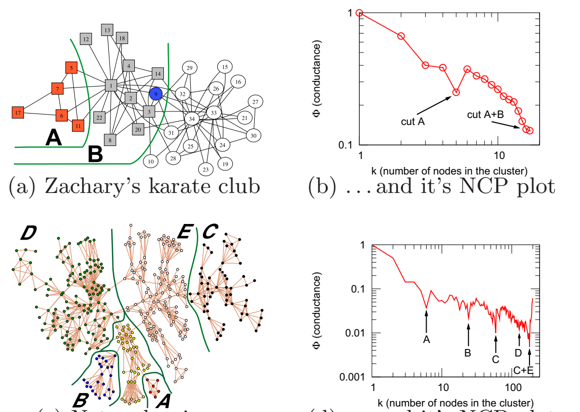
Just as the conductance of a set of nodes provides a quality measure of that set as a community, the shape of the NCP plot provides insight into the community structure of a graph. For example, the magnitude of the conductance tells us how well clusters of different sizes are separated from the rest of the network. One might hope to obtain some sort of



(a) Low-dimensional meshes (b) PowerGrid

(c) RoadNet-CA (d) Expander: dense G_{nm}

Figure 2: NCP plots for networks that "live" in low-dimensional spaces and for an expander-like graph.



(a) Zachary's karate club (b) ...and it's NCP plot

(c) Network science (d) ...and it's NCP plot

Figure 3: Depiction of several small social networks that are common test sets for community detection algorithms and their network NCP plots.

"smoothed" measure of the notion of the best community of size k, *e.g.*, by considering a 95-th percentile, rather than a minimum. We have not defined such a measure since there is no obvious way to average meaningfully over all subsets of size k. Although Metis+MQI finds sets of nodes with extremely good conductance value, 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 [32], the Local Spectral Algorithm finds sets that are "tighter" and more "well-rounded" and thus in many ways more community-like.

3.2 Community profile plots for expander, low-dimensional, and small social networks

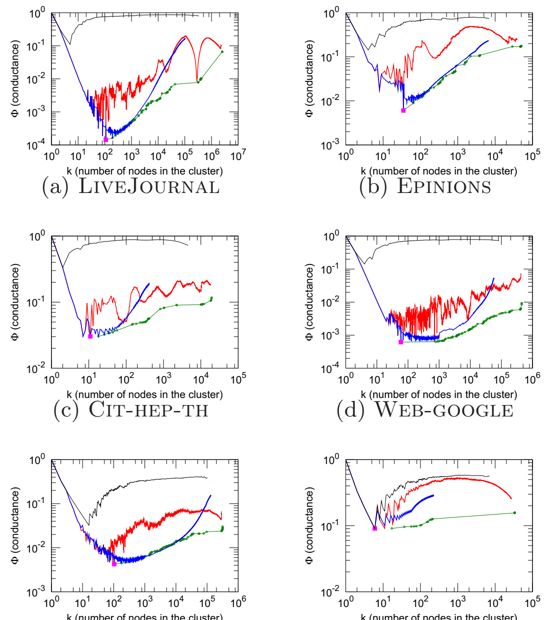
The NCP plot behaves in a characteristic manner for graphs that are "well-embeddable" into a low-dimensional geometric structure. To illustrate this, consider Figure 2. The NCP plot is steadily downward sloping as a function of the number of nodes in the smaller cluster. Moreover, the curves are straight lines with a slope equal to $-1/d$, where d is the dimensionality of the underlying grids. In particular, as the underlying dimension increases then the slope of the NCP plot gets less steep. Of course, this is a manifestation of the isoperimetric (*i.e.*, surface area to volume) phenomenon. 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 2(b) shows the NCP plot for a power grid network of Western States Power Grid [34], and Figure 2(c) shows the NCP plot for a road network of California. Finally, in contrast, Figures 2(d) shows NCP plots for a G_{nm} graph with $100,000$ nodes and average degrees of 4, 6, and 8, *i.e.*, graphs that are very good expanders. The NCP plot is roughly flat, which we also observed in Figure 2(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 [15].

Interestingly, a steadily decreasing downward NCP plot is also seen for small social networks that have been extensively studied for validating community detection algorithms. Two examples are shown in Figures 3. For these networks, the interpretation is the hierarchical organization, where smaller communities are sparsely embedded in larger communities. Empirically we observe that local minima in the NCP plot correspond to sets of nodes that are plausible communities. Consider, *e.g.*, Zachary's karate club [35], an extensively-analyzed social network [24, 26]. Figure 3(a) depicts the karate club network, and Figure 3(b) shows its NCP plot. Note that Cut B, which separates the graph roughly in half, has better conductance value than Cut A (note also community A is included in B). This corresponds with the intuition about the NCP plot derived from studying low-dimensional graphs. The karate network corresponds well with 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. In a similar manner, Figure 3(c) depicts Newman's network of 379 scientists who conduct research on networks [25]. In this latter case, we see a hierarchical structure, in which the community defined by Cut C is included in a larger community that has better conductance value.

3.3 Community profile plots of large social and information networks

We have examined NCP plots for over 70 real-world social and information networks, and in Figure 4 we present NCP plots for six of these. The most striking feature is that the NCP plot is steadily increasing for nearly its entire range.

Consider, the NCP plot for the LiveJournal social network in Figure 4(a), and focus first on the red curve, which presents the results of Local Spectral Algorithm. Up to a size scale, which empirically is roughly 100 nodes, the slope of the NCP plot is generally sloping downward. At that size scale, we observe the global minimum of the NCP plot (denoted by a purple square). 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. Moreover, they are barely connected to the rest of the graph, *e.g.*, they are typically connected to the rest of the nodes by 1 (or 2, or perhaps 3—we will return to this issue in Section 4) edges. Above the size scale of roughly 100 nodes, the NCP plot gradually increases over several orders of magnitude.



(a) LiveJournal (b) Epinions (c) Cit-hep-th (d) Web-google (e) Atp-DBLP (f) Gnutella

Figure 4: [Best viewed in color.] NCP plots for a representative sample of large networks. Red curves plot the Local Spectral Algorithm; green curves plot Metis+MQI; blue curves plot the Bag of Whiskers Heuristic; and black curves plot the Local Spectral Algorithm applied to a randomly rewired network.

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. (The green curve plots the Metis+MQI, and the blue curve the results of Bag of Whiskers Heuristic, as described in Section 4.3.) Note that both axes in Figure 4 are logarithmic, and thus the upward trend of the NCP plot is over a wide range of size scales.

The black curve in Figure 4(a) plots the Local Spectral Algorithm applied to a *rewired version* of the LiveJournal network, *i.e.*, to a random graph conditioned on the same degree distribution as the original network. Interestingly, the rewired network also has an initially decreasing and then increasing/flattening NCP plot. Several things should be noted. (1) The original LiveJournal network has considerably more structure, *i.e.*, deeper/better cuts, than its rewired version, even up to the largest size scales. (2) Relative to the original network, the "best" community in the rewired graph, *i.e.*, the global minimum of the conductance curve, shifts upward and towards the left. This means that in rewired networks the best conductance clusters get smaller and have worse conductance scores. (3) The sets at and near the minimum are small trees that are connected to the core of the random graph by a single edge. (4) After the small dip at a very small size scale (≈ 10 nodes), the NCP plot increases to its high level rather quickly. This is due to the absence of structure in the (expander-like) core.

We have observed qualitatively similar results in other large social and information networks we have examined. Several additional examples are presented in Figure 4: another social network, (Epinions, in Fig. 4(b)); an information/citation network (Cit-hep-th, in Fig. 4(c)); a Web graph (Web-google, in Fig. 4(d)); a Bipartite affiliation network (Atp-DBLP, in Fig. 4(e)); and an Internet network (Gnutella, in Fig. 4(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 than each other—some of these differences are hidden due to the definition of the NCP plot, whereas others are evident. An example of the latter is that even the best cuts in Gnutella 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.

These findings mean that best-expressed network communities are rather small, their size being practically independent of network size (ca. 100 nodes). Moreover, as the community size grows the community blends into the rest of the network, which makes them very difficult to detect using cut-based ideas. (We come back to this in Section 7.)

4. MORE STRUCTURAL OBSERVATIONS

Next we describe the results of examining the networks in greater detail to understand which structural properties are responsible for the observed properties of the NCP plot.

4.1 General statistics on our network datasets

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 around 100 nodes that are attached to the remainder of the network via a small number of edges. In particular, a large fraction of the network is made out of nodes that are not in the (2-edge-connected) core, *i.e.*, they are in components attached to the core of the network via a *single* edge. For example, the core of Epinions network contains only 47% of the nodes and 80% of the edges. Averaging over all our networks, we see that the network core contains around only 60% of the nodes and 80% of the edges of the original network. This is somewhat akin to the so-called "Jellyfish" model [33] and "Octopus" models [7], which we describe 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 to the rest of the network 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 datasets, we examine them in greater detail.

4.2 "Whiskers" and the "core" in our networks

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*. To find 1-whiskers, we employ the following algorithm. Using a depth-first search algorithm, we find the largest 2-edge-connected component B of the graph G. (A graph is 2-edge-connected if the removal of any single edge does not disconnect the graph.) We then delete all the edges that have one of the end points in B. We call the connected components of this new graph G' 1-whiskers, since they correspond to largest subgraphs that can be disconnected from G by removing just a single edge.

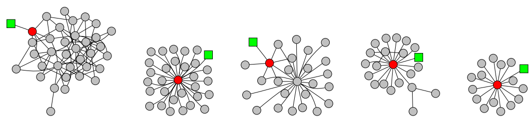


Figure 5: Five largest whiskers of Epinions network.

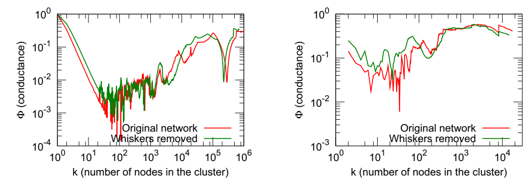
Not surprisingly, there is a wide range of whisker sizes and shapes. Empirically, 1-whisker distribution is heavy-tailed, with the largest whisker size ranging from around less than 10 to well above 100. (See extended version [23] for plots.) The largest whiskers in co-authorship and citation networks have around 10 nodes, whiskers in bipartite graphs also tend to be small, and very large whiskers are found in a web graph. In rewired networks the whiskers tend to be much smaller than in the original network. A particularly noteworthy exception is found in the Autonomous systems networks and the GNUTELLA network. Here, 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: many large whiskers would have bad effects on the Internet connectivity in case of link failures.

Figure 5 shows the five largest whiskers of the EPINIONS social network. The whiskers have on the order of 50 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 datasets 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.

4.3 Bags and communities of whiskers

Empirically, if one looks at the sets of nodes achieving the minimum in the NCP plot (usually the green Metis+MQI curve), then before 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 them are responsible for the "best" conductance sets of different sizes, we have developed the *Bag-of-Whiskers Heuristic*. Suppose we have a set $W = \{w_1, w_2, \ldots\}$ 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 a dynamic programming heuristic to get an approximate solution to this problem. This way, we find a cluster of a particular size that is composed solely from whiskers. Figure 4 (blue curve) shows the results of Bag-of-Whiskers.

First, notice that the largest whisker (denoted with purple square) is the lowest point in all plots. This means that the best conductance community is in a sense trivial as it is connected via just a single edge, and in addition a very simple heuristic can find it. Second, note that above that size scale the Bag-of-Whiskers finds sets of extremely good conductance. Third, this heuristic often agrees with the results from Metis+MQI. This means that the best communities are indeed *disconnected*. Thus, if one only cares about finding good cuts then best cuts in these large sparse graphs are obtained by composing unrelated disconnected pieces, which suggests that community goodness scores need to be reevaluated by also considering the community "coherence".



(a) LIVEJOURNAL (b) EPINIONS

Figure 6: [Best viewed in color.] NCP plots with (in red) and without (in green) 1-whiskers, for two of the six networks shown Figure 4.

4.4 Networks with no whiskers

One might wonder whether we see something different if we consider a network in which these barely-connected pieces have been removed. Thus, we found all whiskers and removed them from the network, using procedure described in Sec. 4.2, *i.e.*, we kept the largest 2-edge-connected component. Again, we computed the NCP plots in Figure 6.

Notice that whisker removal does not change the NCP plot much: the plot shifts slightly upward, but the general trends remain the same. Upon examination, the global minimum occurs with a "whisker" that is connected by *two* edges to the rest of the network. Intuitively, the network core has a large number of barely connected pieces—connected now by two edges rather than by a single edge. Since the "volume" for these pieces is similar to that for 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. RESULTS FROM OTHER ALGORITHMS

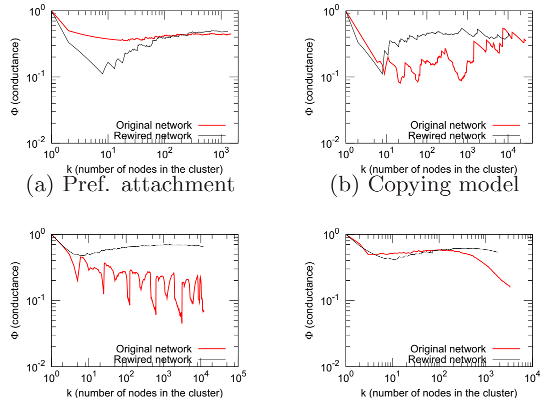
We we have employed a range of other algorithmic techniques to be confident that we are computing quantities fundamental to the networks we are considering, rather than artifacts of the heuristics and approximation algorithms we employ. Due to space limitations, much of this technical material and its associated discussion is omitted from this conference paper, but full details may be found in the journal version of this paper [23].

6. MODELS FOR NETWORK COMMUNITY STRUCTURE

In this section, we address modeling issues in order to understand the properties of generative models sufficient to reproduce the phenomena we have observed.

6.1 Commonly-used network models

We have studied a wide range of commonly-used network generative 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



(a) Pref. attachment (b) Copying model

(c) Barabasi Hierarchical (d) Geometric PA

Figure 7: [Best viewed in color.] NCP for networks from commonly network generation models. Red curves are Local Spectral Algorithm on the original network, and black curves are Local Spectral Algorithm applied to a randomly rewired network.

range of parameters, and in no case have we been able to reproduce our empirical observations, at even a qualitative level. In Figure 7, we summarize these results.

Figure 7(a) shows the NCP plot for a $10,000$ node network generated according to the original preferential attachment model [1], 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 (even more than the corresponding rewired graph), and thus the network that is generated is very expander-like at all size scales. In a different type of generative model edges are added via a copying mechanism [18]. Figure 7(b) shows the results for a network with $50,000$ nodes, generated with $m = 2$ and $\beta = 0.05$. Although the copying model aims to produce communities by linking a new node to neighbors of a existing node, this does not seem to be the right mechanism to reproduce the NCP plot since potential attachment nodes are all treated equally and since new nodes always create same number of edges.

Next, in Figure 7(c), we consider a network that was designed to have a recursively hierarchical community structure [27]. In this case, however, the NCP plot is sloping downwards, and the local dips in the plot correspond to multiples of the size of the basic module of the graph. Finally, Figure 7(d) shows the NCP plot for a geometric preferential attachment model [12]. This model aims to achieve a heavy-tailed degree distribution as well as deep cuts, 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 bi-partitions are fairly well-balanced [12].

6.2 A very sparse random graph model

We have studied a random graph model with given expected degrees, as described by Chung and Lu [7]. Let 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 $\max_i w_i^2 < \sum_k w_k$), be given. Then, in this random graph model, an edge between nodes i and j is added, independently, with probability $p_{ij} = w_i w_j / \sum_k w_k$. We use $G(\mathbf{w})$ to denote a random graph generated in this manner.

The special case of the $G(\mathbf{w})$ model in which $\mathbf{w}$ has a power law distribution is of interest to us here. Given the number of nodes n, the power-law exponent β, and the parameters w and $w_{\max}$, Chung and Lu [7] give the degree sequence for a power-law graph:

$$w_i = ci^{-1/(\beta-1)} \text{ for } i \text{ s.t. } i_0 \le i < n + i_0, \tag{9}$$

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:

$$c = \alpha w n^{1/(\beta-1)} \text{ and } i_0 = n\left(\alpha \frac{w}{w_{\max}}\right)^{\beta-1}, \tag{10}$$

where we have defined $\alpha = \frac{\beta-2}{\beta-1}$. In this case, one can verify that the number of vertices that have expected degree in the range $(k-1, k]$ is proportional to $k^{-\beta}$.

The following theorem will characterize the shape of the NCP plot for this $G(\mathbf{w})$ model when the degree distribution follows Equation (9), with $\beta \in (2, 3)$. The theorem makes two complementary claims: (1) the model has clusters of log size with logarithmically deep cuts; (2) once we get beyond this size scale there do not exist any such deep cuts.

THEOREM 1. *Consider the random power-law graph model $G(\mathbf{w})$, where $\mathbf{w}$ is given by Equation (9), where $w > 5.88$, and the power-law exponent β satisfies $2 < \beta < 3$. Then, 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 exists $c', \epsilon > 0$ such that there are no sets of size larger than $c' \log n$ having conductance smaller than ϵ.*

PROOF. See the journal version of this paper [23]. □

Recall that when $w \ge \frac{4}{e}$ and $\beta \in (2, 3)$ then a typical graph in this model is not fully connected but does have a giant component [7]. (The well-studied $G_{n,p}$ random graph model also has a similar regime when $p \in (1/n, \log n/n)$.) In addition, under certain conditions, the average distance between nodes is in $O(\log \log n)$ 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 [7], and numerous "whiskers" attached.

6.3 A more realistic model of network community structure

We have seen that commonly-studied models, including preferential attachment models, copying models, simple hierarchical models, and models in which there is an underlying mesh-like or manifold-like geometry are not the right way to think about the network community structure. We have also seen that the extreme sparsity of the networks might be responsible for the deep cuts at small sizes.

The question arises 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:

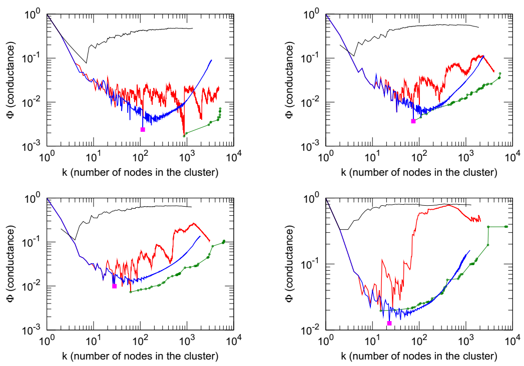


Figure 8: [Best viewed in color.] NCP plots for the Forest Fire Model at various parameter settings. The backward burning probability is $p_b = 0.3$, and we increase (left to right, top to bottom) the forward burning probability $p_f = \{0.26, 0.33, 0.35, 0.40\}$. Note that the largest and smallest values for p_f lead to less realistic community profile plots.

(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 thought of 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* [21, 22] captures exactly these two competing phenomena. The Forest Fire Model is a model of graph generation (that generates directed graphs—an effect we will ignore) in which new edges are added via a recursive "burning" mechanism in an epidemic-like fashion.

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. 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. See [21, 22] for details.

The Forest Fire Model is parameterized by a *forward burning probability* p_f and a *backward burning probability* p_b, and, not surprisingly, the behavior of the model is sensitive to the choice of p_f and p_b. We have experimented with a wide range of network sizes and values for these parameters, and in Figure 8, 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 the four plots in Figure 8, we are viewing networks with very different densities. Next, notice that if, *e.g.*, $p_f = 0.33$ 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 [21, 22]. Next, 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, the network is extremely sparse and is tree-like 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 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 which 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.

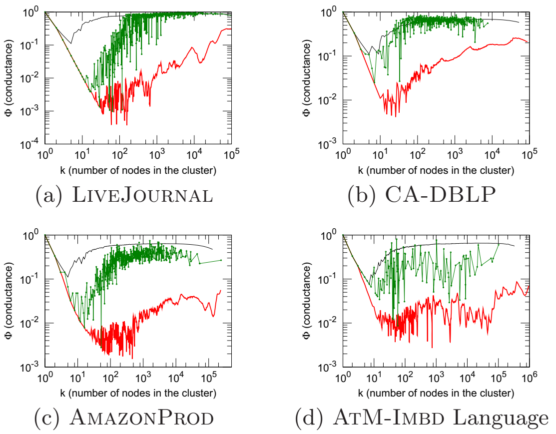
7. DISCUSSION

7.1 Comparison to ground truth communities

A common practice when evaluating community detection algorithms is to compare extracted communities with some notion of "ground truth" (in a hope that extracted and true communities correspond). We have examined four networks in which we have access to some notion of "ground truth".

- LiveJournal [5] is an online blogging community where users create and then join groups. We view each such group as defining a "ground truth" community.
- CA-DBLP [5] is a network in which nodes are authors and edges connect authors co-authoring at least one paper. Here, publication venues (*e.g.*, journals, conferences) play the role of "ground truth" communities.
- AmazonProd [8] is a network linking products often purchased together at `amazon.com`. Each item belongs to one or more hierarchically organized categories, and products from the same category define a group which is a "ground truth" community.
- AtM-IMDB is a bipartite actors-to-movies network. 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.

To examine the quality of "ground truth" communities in the these network datasets, one can take all groups and measure the conductance of the cut that separates the group from the rest of the network. Thus, we generated NCP plots in the following way. For every "ground truth" community, we measure its conductance, from which we obtain a scatter plot of community size versus conductance. Then, we take the lower-envelope of this plot, *i.e.*, for every k we find the conductance value of the community of size k that has



(a) LiveJournal (b) CA-DBLP

(c) AmazonProd (d) AtM-Imbd Language

Figure 9: [Best viewed in color.] NCP plots for explicitly "ground truth" communities (green), compared with that for the original network (red) and a rewired version of the network (black).

the lowest conductance. Figure 9 shows the results for these network datasets; the figure also shows the NCP plot obtained from using the Local Spectral Algorithm on both the original network (red) and on the rewired network (black).

Several things should be noted. First, the conductance of "ground truth" communities follows that for the network communities up to until size 10-100 nodes, *i.e.*, 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 "ground truth" communities. Second, one might expect that the NCP plot for the "ground truth" communities (the green curves) will be somewhere between the NCP plot of the original network (red curve) and that for the rewired network (black), 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 very 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. Third, for the two social networks we studied (LiveJournal 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 (green and black curves overlap). This is likely associated with the relatively weak and overlapping notion of "ground truth" we associated with those two network datasets. On the other hand, for AmazonProd and AtM-IMDB networks, the general trend still remains but large "ground truth" communities have conductance scores that lie well below the rewired network curve.

7.2 Broader implications

In contrast to numerous studies of community structure, we find that the best communities are relatively small with sizes only up to about 100 nodes. We also find that above size of about 100, the "quality" of communities get worse and worse and communities more and more "blend into" the the graph. Eventually, even the existence of communities (at least when viewed as sets with stronger internal than external connectivity) is rather questionable. This seems to agree with Dunbar [11] who predicted that 150 is the upper limit on the size of a human community. Moreover, Allen [2] gives evidence that on-line 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 magic sum of 150 [2]. We are thus led to ask: Why 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 [9]), 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 literature at small sizes. But it seems that in order to make our observations one needs to look at large networks. Probably 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 vs. cluster quality.

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, *e.g.*, especially in when online people have very different criteria for calling someone a friend. Traditionally social scientists through questionnaires "normalized" the links by making sure each link has the same semantics/strength.

There has also been some evidence that hints towards the findings we make here. For example, Clauset et al. [8] analyzed community structure of the AmazonProd, and found that 50% of the nodes belonged to the largest "miscellaneous" community. This agrees with the typical size of the network core (as defined in Section 4.1), and one could conclude that the largest community they found corresponds to the intermingled core of the network, and the rest of the communities are whisker-like.

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 it becomes so diverse, that it stops existing as a traditionally understood "network community". It blends 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 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 only worries about conductance, then bags of whiskers and other internally disconnected sets have the best scores. This raises interesting questions about cluster coherence, regularization and smoothness: what is a good definition of coherence, and how should this be connected to the notion of community separability.

8. CONCLUSION

We investigated statistical properties of sets of nodes in large real-world social and information networks that could plausibly be interpreted as good communities, and we discovered that community structure in these networks is very different than what we expected from the literature and from what commonly-used models would suggest. The most striking example of this is that, in nearly every network dataset we examined, the conductance score of the best possible set of nodes gets gradually worse and worse as those sets increase in size. This suggests that that larger and larger clusters are "blended in" more and more with the rest of the network. Our interpretation is that if a concept like 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. 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.

Acknowledgement

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 1.

9. REFERENCES

[1] R. Z. Albert and A-L. Barabási. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[2] Christopher Allen. Life with alacrity: The Dunbar number as a limit to group sizes, `http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html`, 2004.

[3] 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*, pages 475–486, 2006.

[4] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In *STOC '04: Proceedings of the 36th annual ACM Symposium on Theory of Computing*, pages 222–231, 2004.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54, 2006.

[6] F.R.K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. AMS, 1997.

[7] F.R.K. Chung and L. Lu. *Complex Graphs and Networks*, volume 107 of *CBMS Regional Conference Series in Mathematics*. AMS, 2006.

[8] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. arXiv:cond-mat/0408187, August 2004.

[9] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 29(09):P09008, 2005.

[10] 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):1944–1957, 2007.

[11] Robin Dunbar. *Grooming, Gossip, and the Evolution of Language*. Harvard Univ Press, October 1998.

[12] A.D. Flaxman, A.M. Frieze, and J. Vera. A geometric preferential attachment model of networks. In *WAW '04: Proceedings of the 3rd Workshop On Algorithms And Models For The Web-Graph*, pages 44–55, 2004.

[13] M. Gaertler. Clustering. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations*, pages 178–215. Springer, 2005.

[14] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 KDD Cup. *SIGKDD Explorations*, 5(2), 2003.

[15] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43:439–561, 2006.

[16] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Jour. of the ACM*, 51(3), 2004.

[17] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20:359–392, 1998.

[18] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 2000.

[19] K. Lang and S. Rao. A flow-based method for improving the expansion or conductance of graph cuts. In *IPCO '04: Proceedings of the 10th International Conf. on Integer Programming and Combinatorial Optimization*, 2004.

[20] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787–832, 1999.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187, 2005.

[22] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transact. on Knowledge Discovery from Data*, 1(1), 2007.

[23] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. *Manuscript*.

[24] M.E.J. Newman. Detecting community structure in networks. *The European Physical J. B*, 38:321–330, 2004.

[25] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74, 2006.

[26] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, 2006.

[27] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67:026112, 2003.

[28] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *ISWC '03: Proceedings of the 2nd International Semantic Web Conference*, pages 351–368, 2003.

[29] 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):50–57, 2002.

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

[31] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transcations of Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[32] D.A. Spielman and S.-H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *FOCS '96: Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, pages 96–107, 1996.

[33] S.L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the internet topology. In *GLOBECOM '01: Global Telecommunications Conference*, pages 1667–1671, 2001.

[34] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[35] W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[36] C.T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20(1):68–86, 1971.

[37] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:311–331, 2004.