Vertex Neighborhoods, Low Conductance Cuts, and Good Seeds for Local Community Methods

David F. Gleich*
Purdue University
Computer Science Department
dgleich@purdue.edu

C. Seshadhri†
Sandia National Laboratories‡
Livermore, CA
scomand@sandia.gov

ABSTRACT

The communities of a social network are sets of vertices with more connections inside the set than outside. We theoretically demonstrate that two commonly observed properties of social networks, heavy-tailed degree distributions and large clustering coefficients, imply the existence of vertex neighborhoods (also known as egonets) that are themselves good communities. We evaluate these neighborhood communities on a range of graphs. What we find is that the neighborhood communities can exhibit conductance scores that are as good as the Fiedler cut. Also, the conductance of neighborhood communities shows similar behavior as the network community profile computed with a personalized PageRank community detection method. Neighborhood communities give us a simple and powerful heuristic for speeding up local partitioning methods. Since finding good seeds for the PageRank clustering method is difficult, most approaches involve an expensive sweep over a great many starting vertices. We show how to use neighborhood communities to quickly generate a small set of seeds.

Categories and Subject Descriptors

I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms

Algorithms, Theory

Keywords

clustering coefficients, triangles, egonets, conductance

*Supported by NSF CAREER award 1149756-CCF.
†Supported by the Sandia LDRD program (project 158477) and the applied mathematics program at the Dept. of Energy.
‡Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

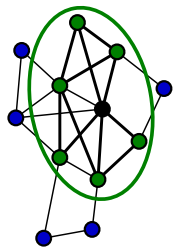
KDD'12, August 12–16, 2012, Beijing, China.

1. INTRODUCTION

Community detection, loosely speaking, is any process that takes a graph or network and picks out sets of related nodes. An incredibly variety of techniques exist for this single task, which has a variety of names as well: community detection, graph clustering, and graph partitioning. Thus, we use community and cluster interchangeably. For more information about approaches for this problem, see the recent survey by Schaeffer [33]. In many techniques, a community is a set with a good score under a quality measure that reflects the connectivity between the set and the rest of the network. Common measures are based on density of local edges, deviance from a random null model, the behavior of random walks, or graph cuts. Mostly, these measures are NP-hard to optimize.

To keep this manuscript simple, we shall evaluate communities using their *conductance score*. Schaeffer identified this measure as one of the most important cut-based measures and it has been studied extensively in a variety of disciplines [11, 17, 35]. Work by Leskovec et al. has recently demonstrated that, although different quality measures produce differences in terms of specific communities, strong communities persist under a variety of measures [26].

A vertex neighborhood of a vertex v is the set of vertices directly connected to v via an edge and v itself. For example, see the green and black vertices at right. *What we show here is that the presence of two commonly observed properties of modern information networks – a large global clustering coefficient [38] and a heavy tailed degree distribution [5] – implies the existence of vertex neighborhoods with good conductance scores.* We make this statement precise in Theorem 4.6, and the discussion surrounding it in Section 4. These results can be seen as an extension of the simple observation that, in the extreme case when the global clustering coefficient of a network is 1, the network must be a union of cliques. Neighborhoods define ideal communities in this case. We mathematically show that this argument can be extended to the case when the graph has a heavy tailed degree distribution and a large clustering coefficient. The significance of this finding is that robust community detection need not employ complicated algorithms. Instead, a straightforward approach that just involves counting triangles – a function that is easy to im-



plement in MapReduce [12] and easy to approximate [21], suffices to identify communities. It is intriguing that arguably the two most important measurable quantities of social networks imply that communities are very easy to find. This may lead to more mathematical work explaining the success of community detection algorithms, given that the problem is in general NP-hard. We note that unfortunately, our theoretical bounds reflect a worst case behavior and are weaker than required for practical use. Consequently, in the remainder of the paper we explore the utility of neighborhood communities empirically on 15 public real-world networks including collaboration networks, social networks, technological networks, and web networks – see Section 5 for a discussion.

Our empirical investigation of neighborhood clusters begins in Section 6. We first exhibit the conductance scores for the set of neighborhood communities for a few graphs (e.g. Figure 2). We find that neighborhood communities reflect the shape of the network community plot observed by Leskovec et al. [24, 25] at small size scales. We next compare the best neighborhood communities to those discovered by four other procedures: the Fiedler community, the best personalized PageRank community (§2.3), the best network whisker (§2.3), and the best clusters from METIS [18]. In one third of the cases, a neighborhood community is as good as the best of any of the other algorithms.

Motivated by the success of the neighborhood communities at small size scales, we explore using the best vertex neighborhoods as *seeds* for the Andersen-Chung-Lang algorithm in Section 7. Here, we find that these procedures, when seeded with an easy-to-identify set of neighborhood communities, produce larger clusters that decay as expected by the results in Leskovec et al. [24, 25]

The technical discussion of the manuscript begins next by introducing our notation and precisely defines the quantities we examine, such as clustering coefficients, due to variability in the definitions of these measures. We also discuss the Andersen-Chung-Lang personalized PageRank clustering scheme [2] and the network whiskers from Leskovec et al. [24, 25]. We utilize the latter two algorithms as reference points for the success of our community detection.

We make all of our algorithm and experimental code, the majority of the data for the experiments, and many extra figures that did not fit into the paper available:

`www.cs.purdue.edu/homes/dgleich/codes/neighborhoods`

These codes are easy to use. Given the adjacency matrix of a network `A`, the single Matlab command

```
>> ncpneighs(A)
```

will produce a figure like those in this paper.

Summary of contributions.

- We theoretically motivate the study of neighborhood communities by showing they often have a low conductance in graphs with a heavy tailed degree distribution and large clustering coefficients.
- We empirically evaluate these neighborhood communities and find them comparable to those communities found by other algorithms at small size scales.
- We find a small set of neighborhood communities that can be grown into larger communities using a PageRank based community detection algorithm. The results match those communities found with a more expensive sweep over all communities.

Table 1: A summary of the notation.

$n = \|V\|$	the number of vertices
$m = \|E\|$	the number of edges
d_v	the degree of vertex v
f_d	the number of vertices of degree d
W	the set of wedges in a graph
W_v	the set of wedges centered at vertex v
κ	the global clustering coefficient
$\bar{C}$	the mean local clustering coefficient
C_v	the local clustering coefficient for vertex v
$N_r(v)$	the set of vertices within distance r or v
$E(S,T)$	the set of edges between S and T
$\mathtt{cut}(S)$	the size of the cut around vertex set S
$\mathtt{vol}(S)$	the sum of degrees (volume) of vertices in S
$\mathtt{edges}(S)$	twice the number of edges among vertices in S
$\phi(S)$	the conductance of vertex set S

2. FORMAL SETTING AND NOTATION

All of the key notation is summarized in Table 1, and we briefly review it here. Let $G = (V, E)$ be a loop-less, undirected, unweighted graph. We denote the number of vertices by $n = |V|$ and the number of edges by $m = |E|$. In terms of the adjacency matrix, m is half the number of non-zeros entries. For a vertex v, let d_v be the degree of v. For any positive integer d, let f_d be the number of vertices of degree d, that is, the frequency of d in the degree distribution. The maximum degree is denoted by $d_{\max}$. Let $D_r(v)$ to be the distance r-neighborhood of v. This is the set of vertices whose shortest path distance from v is exactly r. Then, we define the ball of distance r around v, denoted by $N_r(v)$, as the set $\bigcup_{i \le r} D_r(v)$.

2.1 Clustering coefficients

A *wedge* is an unordered pair of edges that share an endpoint. The *center* of the wedge is the common vertex between the edges. A wedge $\{(s,t),(s,u)\}$ is *closed* if the edge (t,u) exists, and is open otherwise. We use W to denote the set of wedges in G, and W_v for the set of wedges centered at V. Note that $|W_v| = \binom{d_v}{2}$. We set $p_v = |W_v|/|W|$.

Social networks often have large *clustering coefficients* [38]. Because of the varying definitions of this term that are used, we will denote by κ the *global clustering coefficient*. This quantity is a normalized count of triangles that we interpret as the probability that a uniform random wedge w is closed:

$$\kappa = \Pr_{w \sim W}[w \text{ is closed}] = \frac{\text{number of closed wedges}}{|W|}$$

In terms of triangles, $\kappa = 3 \cdot \text{number of triangles}/|W|$. For any vertex v, C_v is the *local clustering coefficient of* v, or the probability that a uniform random wedge w from W_v is closed. Formally,

$$C_v = \Pr_{w \sim W_v}[w \text{ is closed}] = \frac{\text{number of closed wedges in } W_v}{|W_v|}$$

2.2 Cuts and Conductance

Given a set of vertices S, the set $\bar{S}$ is the complement set, $\bar{S} = V \backslash S$. For disjoint sets of vertices S, T, $E(S,T)$ denotes the edges between S and T. For convenience, we denote *the size of the cut induced by a set* $|E(S,\bar{S})|$ by $\mathtt{cut}(S)$.

The conductance of a cluster (a set of vertices) measures the probability that a one-step random walk starting in that

cluster leaves that cluster. Let $\texttt{vol}(S)$ denotes the sum of degrees of vertices in S and $\texttt{edges}(S)$ denotes twice the number of edges among vertices in S so that

$$\texttt{edges}(S) = \texttt{vol}(S) - \texttt{cut}(S).$$

Then the conductance of set S, denoted $\phi(S)$, is

$$\phi(S) = \frac{\texttt{cut}(S)}{\min\bigl(\texttt{vol}(S), \texttt{vol}(\bar{S})\bigr)}.$$

Conductance is measured with respect to the set S or $\bar{S}$ with smaller volume, and is the probability of picking an edge from the smaller set that crosses the cut. Because of this property, conductance is preserved on taking complements: $\phi(S) = \phi(\bar{S})$. For this reason, when we refer to the number of vertices in a set of conductance ϕ, we always use the smaller set $\min(|S|, |\bar{S}|)$. Figure 1 shows a few communities and their associated cuts and conductance scores from our methods and two points of comparison.

2.3 Finding good conductance communities

We briefly review three ways of identifying a community with a good conductance score.

Fiedler set. The well-known Cheeger inequality defines a bound between the second smallest eigenvalue of the normalized Laplacian matrix and the set of smallest conductance in a graph [11]. Formally,

$$(1/2)\lambda_2 \leq \min_{S \subset V} \phi(S) \leq \sqrt{2\lambda_2}$$

where λ_2 is the second smallest eigenvalue of the normalized Laplacian. The proof is constructive. It identifies a set of vertices that obeys the upper-bound using a *sweep* cut. This is the smallest conductance cut among all cuts induced by ordering vertices by increasing values of $\sqrt{d_v}x_v$, where x_v is the component of the eigenvector associated with λ_2. This is the same idea used in normalized cut procedures [35]. We refer to the set identified by this procedure as the *Cheeger* community or *Fiedler* community. The latter term is based on Fiedler's work in using the second smallest eigenvalue of the combinatorial Laplacian matrix [14]. Figure 1b shows the Fiedler community for the Les Misérables network.

Personalized PageRank communities. Another successful scheme for community detection based on conductance uses personalized PageRank vectors. A personalized PageRank vector is the stationary distribution of a random walk that follows an edge of the graph with probability α and "teleports" back to a fixed seed vertex with probability $1-\alpha$. We use $\alpha = 0.99$ in all experiments. The essence of the induced community is that an inexact personalized PageRank vector, computed via an algorithm that "pushes" rank round the graph, will identify good bottlenecks nearby a seed vertex. These bottlenecks can be formalized in a Cheeger-like bound [2]. The procedure to find a personalized PageRank community is: i) specify a value of α, a seed vertex v, and a desired cluster size σ; ii) solve the personalized PageRank problem using the algorithm from [2] until a degree-weighted tolerance of $\tau = 1/(10\sigma)$; and iii) sweep over all cuts induced by the ordering of the personalized PageRank vector (normalized by degrees) and choose the best. Personalized PageRank communities (PPR communities, for short) were used to identify an interesting empirical property of communities in large networks [24, 25]. To generate these plots, those authors examined a range of values of σ for a large number of vertices of the graph and summarized the best communities found at any size scale in a network community plot. Figure 1c shows the best personalized PageRank community for the network of character interactions in Les Misérables.

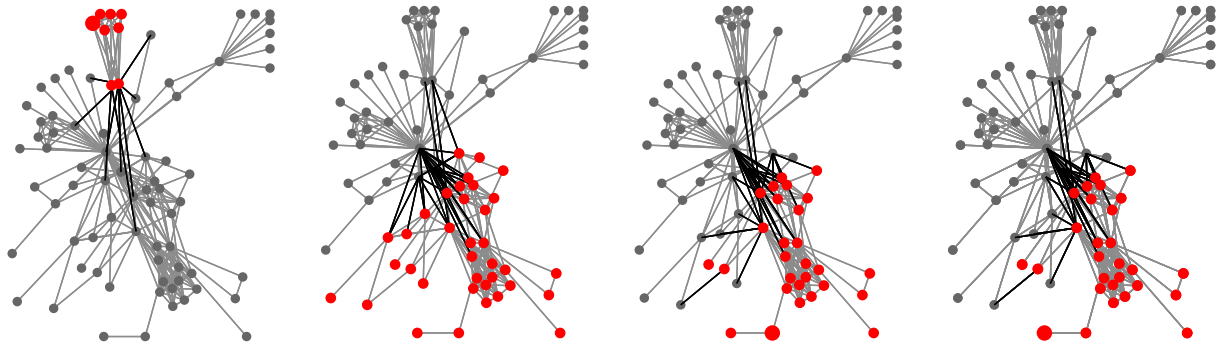
Whisker communities. Perhaps the best point of comparison with our approach are the *whisker communities* defined by Leskovec et al. [24,25]. These communities are small dense subgraphs connected by a single edge. They can be found by looking at any subgraph connected to the largest biconnected component by a single edge. Note that the largest biconnected component is not necessarily a 2-core of the graph. (Here, a graph k-core is a subset of vertices where all nodes have degree at least k [34].) Leskovec et al. observed that many of these subgraphs are rather dense. Each subgraph has a cut of exactly one, and consequently, a productive means of finding sets with low conductance is to sort these subgraphs by their volume. The best whisker cut is the single subgraph with largest volume.

3. RELATED WORK

Egonets, homophily, and structural holes. In the context of social networks, vertex neighborhoods are often called egonets because they reflect the the state of the network as perceived by a single vertex. Their analysis is a key component in the study of social networks [37], especially in terms of data collection. Studies of these networks often focus on the theory of structural holes, which is the notion that an individual can derive an advantage from serving as a bridge between disparate groups [9]. These bridge roles are interesting because they contradict homophily in social ties. Homophily, or the principle that similar individuals form ties, is the mechanism that is expected to produce networks with large local clustering coefficients [28]. These social theories have prompted the development of new methods to tease apart some of these effects in real-world networks [22], and to develop network models that capture structural holes [19].

Clustering and communities. Vertex neighborhoods often play a role in other techniques to find community or clustering structure in a network. Overlap in the neighborhood sets of vertices is a common vertex similarity metric used to guide graph clustering algorithms [33]. Other schemes utilize vertex neighborhoods as good *seed sets* for local techniques to grow communities [16,32]. We explore using a carefully chosen set of neighborhoods for this purpose in our final empirical discussion (§7). Perhaps the most closely related work is a recent idea to utilize the connected components of egonets, after their ego vertex is removed, to produce a good set of overlapping communities [31]. Our theoretical results establish that these ideas are highly likely to succeed in networks with local clustering and heavy tailed degree distributions.

Graph properties. Much of the modern work on networks rests on surprising empirical observations about the structure of real world connections. For instance, information networks were found to have a heavy tailed in the degree distribution [5,13]. These same networks were also found to have considerable local structure in the form of large clustering coefficients [38], but retained a small global diameter. Our theory shows that a third potential observation – the existence of vertex neighborhood with low conductance – is in fact implied by these other two properties. We formally show that heavy tailed degree distributions and high clustering coefficients imply the existence of large dense cores.



(a) Best neighborhood size=8, cut=10, ϕ=0.15
(b) Fiedler community size=36, cut=29, ϕ=0.13
(c) Best PPR community size=28, cut=31, ϕ=0.12
(d) Best seeded community size=28, cut=31, ϕ=0.12

Figure 1: A series of vertex sets and their associated sizes and conductance score on the graph of characters from *Les Misérables* [20]. The best neighborhood and best seeded community are discussed further in §6 and §7, respectively. See §2.3 for more on the Fiedler and PPR communities. Finding (d) is faster than (c).

Anomaly detection. Predictable behavior in the structure of egonets makes them a useful tool for detecting anomalous patterns in the structure of the network. For instance, Akoglu et al. [1] compute a small collection of measures on each egonet, such as the average degree and largest eigenvalues. Outliers in this space of vertices are often rather anomalous vertices. Our work is, in contrast, a precise statement about the regularity of the egonets, and says that we always expect a large egonet to be a good community.

Summary. Although we are not the first to study neighborhood based communities, the relationship between the local clustering, heavy tailed degree distributions, and large neighborhoods with small conductance does not appear to have been noticed before.

4. THEORETICAL JUSTIFICATION FOR NEIGHBORHOOD COMMUNITIES

The aim of this section is to provide some mathematical justification for the success of neighborhood cuts. Our aim is to show that heavy tailed degree distributions and large clustering coefficients imply the existence of neighborhood cuts with low conductance and large dense cores. *We stress that the exact bounds obtained are weak and only hold when the clustering coefficient is extremely large. Nonetheless, the proofs give significant intuition into* why *neighborhoods are good communities.* The theory is truly tested with the experimental work in the next sections.

We begin with the extreme case when the value of κ is 1, when *every* wedge is closed, for the following simple claim.

CLAIM 4.1. *Suppose the global clustering coefficient of G is 1. Then G is the union of disjoint cliques.*

PROOF. Consider two vertices u and v that are connected. Suppose the shortest path distance between them is $\ell > 1$. Then the shortest path has at least 3 distinct vertices (including u and v). Take the last three vertices on this path, v_1, v_2, v. This forms a wedge at v_2, and must be closed (since the clustering coefficient is 1). Hence, the edge (v_1, v) exists and there exists a path between u and v of length less than ℓ, contradicting the assumption. Hence, any two connected vertices have a shortest path distance of 1, i.e., are connected by an edge and the graph is a disjoint union of cliques. □

Note that the neighborhood of any vertex in the above claim forms a clique disconnected from the rest of G. Therefore, all neighborhoods form perfect communities, in this extremely degenerate case. We prove this for more general settings.

Define $p_v = |W_v|/|W|$, and observe that it forms a distribution over the set of vertices. We show a simple claim about these values. We remind the reader that κ is the global clustering coefficient of G.

CLAIM 4.2. $\sum_v p_v C_v = \kappa$

PROOF.

$$\sum_v p_v C_v = \sum_v \frac{|W_v|}{|W|} \cdot \frac{\text{number of closed wedges in } W_v}{|W_v|}$$
$$= \frac{\sum_v (\#\text{ closed wedges in } W_v)}{|W|} = \kappa. \quad \square$$

We come to an important lemma. This argues that, on average, neighborhood cuts must have low conductance.

LEMMA 4.3.

$$\sum_v \left(p_v \frac{\mathtt{cut}(N_1(v))}{|W_v|} \right) = 2(1-\kappa)$$

PROOF. We express the sum of $\mathtt{cut}(N_1(v))$ as a double summation, and perform some algebraic manipulations.

$$\sum_v \mathtt{cut}(N_1(v)) = \sum_v \sum_{u \in N_1(v)} |N_1(u) \setminus (N_1(v) \cup \{v\})|$$
$$= \sum_u \sum_{v \in N_1(u)} |N_1(u) \setminus (N_1(v) \cup \{v\})|$$
$$= \sum_u \sum_{v \in N_1(u)} (\#\text{ open wedges centered at } u \text{ involving edge } (u,v))$$
$$= 2 \sum_u (\#\text{ open wedges centered at } u)$$
$$= 2(1-\kappa)|W|$$

We complete the proof with the following simple observation:

$$\sum_v \left(p_v \frac{\mathtt{cut}(N_1(v))}{|W_v|} \right) = \frac{\sum_v \mathtt{cut}(N_1(v))}{|W|}. \quad \square$$

So far, the discussion and proofs hold for any graph with clustering coefficient κ. We now bring in the heavy-tailed degree distribution of G. Since we are performing an asymptotic analysis, we will use $o(1)$ to denote any quantity that becomes negligible as the graph size increases. We will choose β to be a constant less than 1. The asymptotics hold for any β, but from a practical standpoint, we think of β as a constant such that most edges are incident to a vertex of degree at least $d_{\max}^{\beta}$, 2/3 is usually a reasonable value.

Let us formalize the notion that the degree distribution is heavy tailed. This is sometimes expressed as a power law, but that is a fairly contested assertion. Power laws often do not fit real-world degree distributions, but they nonetheless give a rough guide. So we will assume that there is a power law that *approximately bounds* the degree distribution. We formalize this as *Condition* (*).

CONDITION (*) (i) For all d, $f_d \in [\alpha_1 n/d^{\gamma}, \alpha_2 n/d^{\gamma}]$, for some α_1, α_2, and $\gamma < 3$. (ii) The global clustering coefficient of G is κ.

CLAIM 4.4. *Suppose G satisfies Condition (*). Let S be the set of vertices with degrees more than $d_{\max}^{\beta}$. Then, $\sum_{v \in S} p_v = 1 - o(1)$.*

PROOF. We can set $p_v = (2|W_v|)/(2|W|)$. We have $2|W| = \sum_{d>1}^{d_{\max}} d(d-1)f_d$ and $2|W_S| := \sum_{v \in S} 2|W_v| = \sum_{d=d_{\max}^{\beta}}^{d_{\max}} d^2 f_d$. Note that $\sum_{v \in S} p_v = 2|W_S|/2|W|$. It suffices to show that $2|W| - 2|W_S| = o(|W|)$. By the heavy tail condition,

$$2|W| = \sum_v 2|W_v| = \sum_{d>1}^{d_{\max}} d(d-1)f_d$$

$$\geq (\alpha_1 n/2) \sum_{d>1}^{d_{\max}} d^{2-\gamma} \geq (\alpha_1 n/2^{4-\gamma}) d_{\max}^{3-\gamma}$$

$$2(|W| - |W_S|) = \sum_{d>1}^{d_{\max}^{\beta}} d(d-1)f_d \leq \alpha_2 n \sum_{d>1}^{d_{\max}^{\beta}} d^{2-\gamma} \leq \alpha_2 n d_{\max}^{\beta(3-\gamma)}$$

Note that $3-\gamma > 0$ and $\beta < 1$. So the latter is asymptotically smaller than the former, i.e. $2|W| - 2|W_S| = o(|W|)$. □

We next state a cute result that can be derived from our calculations. It is not applied anywhere later, but shows that graphs with heavy tails and large clustering coefficients have large cores. See the online version for the proof.

THEOREM 4.5. *Consider a graph G satisfying Condition (*). There exists a k-core in G for $k \geq \kappa d_{\max}^{\beta}/2$.*

We come to our main theorem that proves the existence of a neighborhood cut with low conductance. When $\kappa = 1$, we get back the statement of Claim 4.1, since we have a set of conductance 0. But this theorem also gives non-trivial bounds for (very) large values of κ. As we mentioned earlier, when κ becomes small, this bound is not useful any longer. Again, it is the intuition of this theorem that is important, not the exact constant that it provides.

THEOREM 4.6. *Consider a graph G satisfying Condition (*). There exists a neighborhood cut with conductance at most $4(1-\kappa)/(3-2\kappa)$.*

PROOF. The proof uses the probabilistic method, given the bounds of Lemma 4.3 and Claim 4.2. Suppose we choose a vertex v according to the probability distribution given by p_v. Let X denote the random variable $\texttt{cut}(N_1(v))/|W_v|$, so $\mathbf{E}[X] = 2(1-\kappa)$ (Lemma 4.3). By Markov's inequality, $\Pr[X > 4(1-\kappa)] \leq 1/2$.

Set $\xi = 2\kappa - 1$, and set $\Pr[C_v < \xi] = p$. Let S_ξ denote the set $\{v | C_v < \xi\}$. (So $\sum_{v \in S_\xi} p_v = p$.) Now, starting with Claim 4.2,

$$\kappa = \sum_v p_v C_v = \sum_{v \in S_\xi} p_v C_v + \sum_{v \notin S_\xi} p_v C_v$$

$$< \xi \sum_{v \in S_\xi} p_v + \sum_{v \notin S_\xi} p_v = p\xi + (1-p)$$

$$\Longrightarrow p < (1-\kappa)/(1-\xi) = 1/2$$

By the union bound, the probability that $\texttt{cut}(N_1(v))/|W_v| > 4(1-\kappa)$ *or* $C_v < \xi$ is less than 1. Hence, there exists some vertex v such that $\texttt{cut}(N_1(v)) \leq 4(1-\kappa)|W_v|$ and $C_v \geq \xi$ (we can also show that $d_v \geq n^{\beta}$). Let E be the set of edges in the subgraph induced on $N_1(v)$. Since $C_v \geq \xi$, $|E| \geq \xi |W_v|$. We can bound the conductance of $N_1(v)$,

$$\frac{\texttt{cut}(N_1(v))}{|E| + \texttt{cut}(N_1(v))} \leq \frac{4(1-\kappa)|W_v|}{\xi|W_v| + 4(1-\kappa)|W_v|} = \frac{4-4\kappa}{3-2\kappa}.$$ □

Table 2: Datasets for our experiments. The five types are: collaboration networks, social networks, technological networks, and web graphs.

Graph	Verts	Edges	Avg. Deg.	Max Deg.	κ	$\bar{C}$
ca-AstroPh	17903	196972	22.0	504	0.318	0.633
email-Enron	33696	180811	10.7	1383	0.085	0.509
cond-mat-2005	36458	171735	9.4	278	0.243	0.657
arxiv	86376	517563	12.0	1253	0.560	0.678
dblp	226413	716460	6.3	238	0.383	0.635
hollywood-2009	1069126	56306653	105.3	11467	0.310	0.766
fb-Penn94	41536	1362220	65.6	4410	0.098	0.212
fb-A-oneyear	1138557	4404989	7.7	695	0.038	0.060
fb-A	3097165	23667394	15.3	4915	0.048	0.097
soc-LiveJournal1	4843953	42845684	17.7	20333	0.118	0.274
oregon2-010526	11461	32730	5.7	2432	0.037	0.352
p2p-Gnutella25	22663	54693	4.8	66	0.005	0.005
as-22july06	22963	48436	4.2	2390	0.011	0.230
itdk0304	190914	607610	6.4	1071	0.061	0.158
web-Google	855802	4291352	10.0	6332	0.055	0.519

5. DATA

Data for our empirical investigation comes from a variety of sources. See Table 2 for a summary of the networks and their basic statistics. All networks are undirected and were symmetrized if the original data were directed. Also, any self-loops in the networks were discarded. We only look at the largest connected component of the network. There are four types of networks here. We investigate forest-fire models [23] in the online figures.

Collaboration networks. In these networks, the nodes represent people. The edges represent collaborations, either via a scientific publication (ca-AstroPh [23], cond-mat-2005 [30], arxiv [8], dblp [6, 7]), an email (email-Enron [25]), or a movie (hollywood-2009 [6, 7]). These networks have

large global clustering coefficients and large mean clustering coefficients.

Social networks. The nodes are people again, and the edges are either explicit "friend" relationships (fb-Penn94 [36], fb-A [39], soc-LiveJournal [4]) or observed network activity over edges in a one-year span (fb-A-oneyear [39]).

Technological networks. The nodes act in a communication network either as agents (p2p-Gnutella25 [27]) or as routers (oregon2 [23], as-22july06 [29], itdk0304 [10]). The edges are observed communications between the nodes.

Web graphs. The nodes are web-pages, and the edges are symmetrized links between the pages [25].

6. EMPIRICAL NEIGHBORHOOD COMMUNITIES

6.1 Computation

We first show that we can adapt any procedure to compute all local clustering coefficients to compute the conductance scores for each neighborhood in the graph. Most of the work to compute a local clustering coefficient is performed when finding the number of triangles at the vertex. We can express the number of triangles with v as:

$$\texttt{edges}(N_1(v) \setminus \{v\})/2$$

because each edge among v's neighbors produces a triangle (recall that the edges function double-counts). Note also that $\texttt{edges}(N_1(v) \setminus \{v\})/2 = \texttt{edges}(N_1(v))/2 - d_v$. Then $\texttt{cut}(N_1(v)) = \texttt{vol}(N_1(v)) - \texttt{edges}(N_1(v))$. And so, given the number of triangles, we can compute the cut given the volume of the neighborhood as well. This is easy to do with any graph structure that explicitly stores the degrees.

6.2 Quality of neighborhood communities

We use Leskovec et al.'s [24] network community plot to show the information on all neighborhood communities simultaneously. These plots will help us understand if the neighborhood communities are high quality (low conductance), and how they compare to other community detection methods. Given the conductance scores from all the neighborhood communities and their size in terms of number of vertices, we first identify the best community at each size. The network community plot shows the relationship between best community conductance and community size on a log-log scale. In Leskovec et al., they found that these plots had a characteristic shape for modern information networks: an initial sharp decrease until the community size is between 100 and 1000, then a considerable rise in the conductance scores for larger communities. In our case, neighborhood communities cannot be any larger than the maximum degree plus one, and so we mark this point on the figures. We always look at the smaller side of the cut, so no community can be larger than half the vertices of the graph. We also mark this location on the plots. Each subsequent figure in this paper utilizes this size-vs-conductance plot, and we will continually layer information from new methods above results from old methods. The result are information-dense plots that need slightly more study than would be ideal, however, we point out the salient features in each plot in the text. Note also that we deliberately attempt to preserve the axes limits across figures to promote comparisons. However, some figures *have different axis limits* to exhibit the range of data.

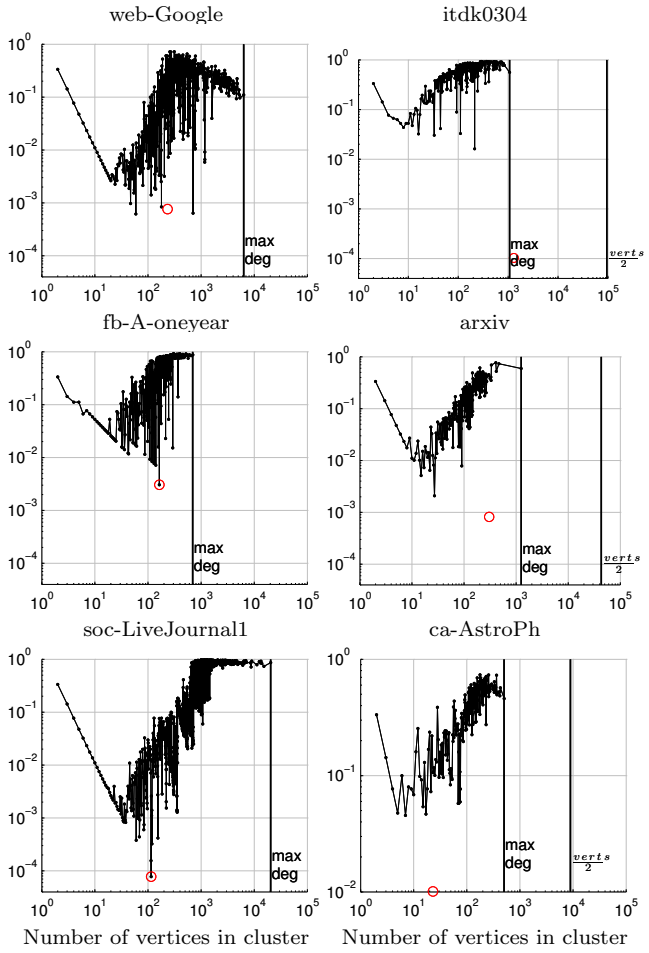


Figure 2: The best neighborhood community conductance at each size (black) and the Fiedler community (red). (Note the axis limits on ca-AstroPh).

First, we show these network community plots for six of the networks in Figure 2. These figures are representative of the best and worst of our results. Plots for other graphs are available on the website given in the introduction.

The three graphs on the left show cases where a neighborhood community *is or is nearby* the best Fiedler community (the red circle). The three graphs on the right highlight instances where the Fiedler community is much better than any neighborhood community. We find it mildly surprising that these neighborhood communities *can be* as good as the Fiedler community. The structure of the plot for both fb-A-oneyear and soc-LiveJournal1 is instructive. Neighborhoods of the *highest degree* vertices are not community-like – suggesting that these nodes are somehow exceptional. In fact, by inspection of these communities, many of them are nearly a star graph. However, a few of the large degree nodes define strikingly good communities (these are sets with a few hundred vertices with conductance scores of around 10^{-2}). This evidence concurs with the intuition from Theorem 4.6.

6.3 Comparison to PPR communities

Note that these plots show the same shape as observed by Leskovec et al. [24]. Consequently, in the next set of figures, and in the remainder of the empirical investigation,

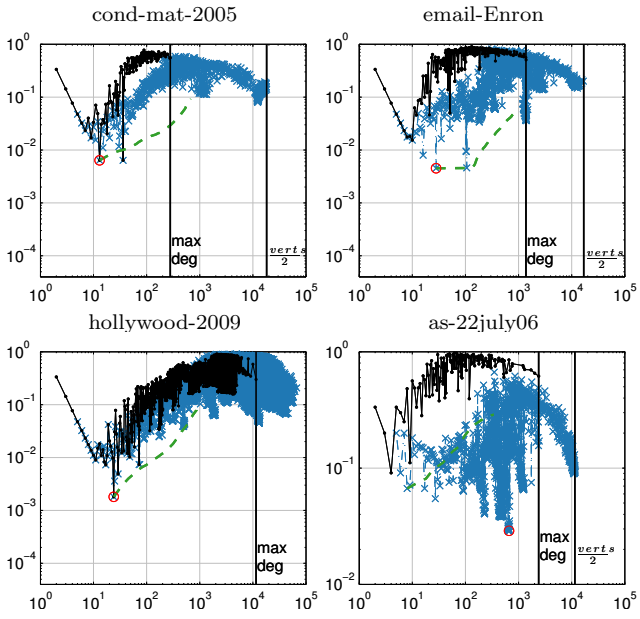


Figure 3: A comparison of neighborhood communities (black) personalized PageRank communities (blue), and whiskers (green).

we compare our neighborhood communities against those computed via the personalized PageRank community scheme employed in that work and described in Section 2.3. Figure 3 compares the neighborhood communities to those computed by running the local personalized PageRank algorithm for all vertices as described by Leskovec et al. [24]. We also show the behavior of the whisker communities in this plot as well. The plot adopts the same style of figure. The PageRank communities are in a deep blue color, and the whisker communities are show in a shade of green. Here, we see that the neighborhood communities show similar behavior at small size scales (less than 20 vertices), but the personalized PageRank algorithm is able to find larger communities of smaller, or similar conductance. In these four cases (which are representative of the others), a personalized PageRank community was also the Fiedler community.

6.4 The best community found

Based on this observation, we wanted to understand how the best community identified by a range of algorithms compares to the neighborhood communities. The results are shown in Table 3. We computed a set of communities with METIS by repeatedly calling the algorithm, asking it to use *more* partitions each time, and saved the best. See our online codes for the precise details of which partitions were used. The seeded community results are described in the next section.

By-and-large, all methods, except neighborhoods, tend to identify similar communities as the best. The Penn94 graph shows a large difference where the Fiedler community is much larger than the best PageRank community *and* it has better conductance. When a neighborhood has conductance that's as good as the rest, then it is always a whisker as well. In the following full section, we explore using these neighborhood communities as *seeds* for the PageRank algorithms. These seeded communities are just as good as the other methods.

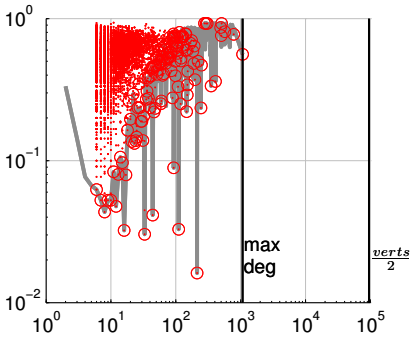


Figure 4: The conductance of locally minimal communities in the itdk0304 graph (red). Note that these capture most of the local minima (downward spikes) in the profile.

7. SEEDED COMMUNITIES

Many of the theorems about extracting local communities from seed sets [2, 3] require that the seed set itself be a good community. This is precisely what our theoretical results justify for neighborhood communities. Consequently, in this section, we look at *growing* the neighborhood communities using the local personalized PageRank community algorithm from a set of carefully chosen seeds.

One of the key problems with using the personalized PageRank community algorithms is that finding a good set of seeds is not easy. Leskovec et al. [25] used exhaustive enumeration over all seeds, and [15] describes a way to do this using the most popular videos on YouTube. Such a meaningful heuristic is not always available. We begin this section by empirically showing that there is an easy-to-identify set of neighborhood communities that are local extrema in the network community plot of the neighborhood communities.

First, some quick terminology: we say a neighborhood community is a local minima, or locally minimal, if the conductance of the neighborhood of a vertex is smaller than the conductance of any of the adjacency neighborhood communities. Formally,

$$\phi(N_1(v)) \leq \phi(N_1(w)) \quad \text{for all } w \text{ adjacent to } v$$

is true for any locally minimal communities. These can be found by looking at each edge in the graph once. We find there are only a small set of locally minimal communities with more than 6 vertices. Shown in Figure 4 are the conductance and sizes of the roughly 7000 communities identified by this measure for the itdk0304 graph. Indeed, among all of the graphs with at least 85,000 vertices, this heuristic picks out about 3% of the vertices as local minima. In the worst case, it picked out 100,000 seeds for soc-LiveJournal1. Increasing the minimum size to 10 vertices reduces this down to 50,000 seeds. We then use these locally minimal neighborhoods as seed sets for the personalized PageRank community detection procedure. Each locally minimal neighborhoods is grown by up to 50-times its volume by solving for communities using values of σ up to 50. In some experiments motivated by Theorem 4.5, we also found interesting results by looking at graph k-cores, and so we also explore growing the k-cores of a graph by up to 5 times their volume.

Figure 5 shows the results. In these figures, we leave the baseline neighborhood communities in for comparison.

Table 3: The best community detected by the six methods explored. The first (§6) and last (§7) are ours.

Graph	Neighborhood		Fiedler		PageRank		Whisker		Metis		Seeded	
	Cond.	Size	Cond.	Size	Cond.	Size	Cond.	Size	Cond.	Size	Cond.	Size
ca-AstroPh	0.0455	7	0.0101	23	0.0101	23	0.0101	23	0.0101	23	0.0101	23
email-Enron	0.0154	10	0.0045	28	0.0045	28	0.0045	28	0.0080	16	0.0045	28
cond-mat-2005	0.0064	13	0.0064	13	0.0064	13	0.0064	13	0.0154	11	0.0064	13
arxiv	0.0021	27	0.0008	303	0.0014	304	0.0021	27	0.0021	27	0.0019	306
dblp	0.0038	24	0.0038	25	0.0034	83	0.0038	25	0.0041	17	0.0034	83
hollywood-2009	0.0018	24	0.0018	24	0.0018	24	0.0018	24	0.0018	24	0.0018	24
Penn94	0.3333	2	0.1898	7191	0.1966	41	0.3333	2	0.1986	6923	0.2009	39
fb-A-oneyear	0.0031	164	0.0031	164	0.0031	164	0.0031	164	0.0090	56	0.0031	164
fb-A	0.0345	8	0.0084	647	0.0084	647	0.0133	38	0.0130	77	0.0084	647
soc-LiveJournal1	0.0001	115	0.0001	115	0.0001	115	0.0001	115	0.0001	115	0.0001	115
oregon2-010526	0.1368	12	0.0467	316	0.0438	318	0.1429	4	0.0553	3820	0.0443	319
p2p-Gnutella25	0.1429	10	0.0417	24	0.0417	24	0.0588	9	0.0417	24	0.0417	24
as-22july06	0.0909	4	0.0289	661	0.0286	654	0.0667	8	0.0296	657	0.0285	656
itdk0304	0.0162	213	0.0001	1306	0.0002	1188	0.0001	1306	0.0046	152	0.0002	1188
web-Google	0.0006	59	0.0008	234	0.0006	59	0.0006	59	0.0006	59	0.0006	59

This leaves the figures rather noisy, but the key insight is that the dark black line (neighborhood seeded communities) closely tracks the the outline of the pure-PageRank based community profile (light blue). That profile was computed by using every vertex in the graph as a seed (although, some vertices were skipped after 10 other clusters had already visited that vertex). This effect is most clearly illustrated by the email-Enron dataset. The dark black line identifies almost all of the local minima from the full PageRank sweep (there are a few it misses). A weakness of these minimal seeds for PageRank is that they may not capture the *largest* communities. However, we found that growing communities from a k-core instead (red line with circles) of a vertex neighborhood do seem to capture this region of the profile (e.g. arxiv), although ca-AstroPh is an exception.

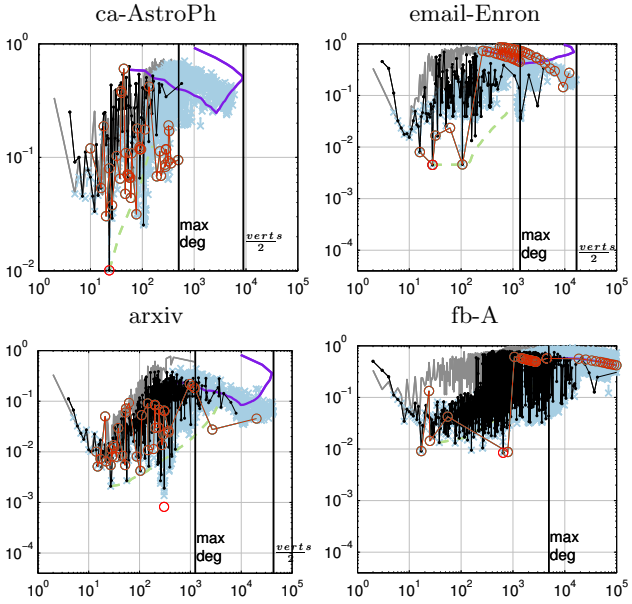


Figure 5: Network community plots with neighborhood communities (gray), PageRank communities (light blue), whiskers (green), k-cores (purple), locally minimal seed PageRank communities (black), and k-core seeded PageRank communities (red).

8. CONCLUDING DISCUSSIONS

We recap. Community detection is the problem of finding cohesive collections of nodes in a network. We formalize this as finding vertex sets with small conductance. Modern information networks have many distinctive properties, including a large clustering coefficient and a heavy-tailed degree distribution. We derive a set of theoretical results that show these properties imply that such networks will have vertex neighborhoods that are *themselves* sets of small conductance. Although our theoretical bounds are weak, they suggest measuring the conductance of vertex neighborhoods.

Algorithms to compute *all such conductance scores* are easy to implement by modifying a routine for computing local clustering coefficients. We evaluate these communities on a set of real-world networks. In summary, our results support the idea that there are many neighborhood communities which are *good communities* in a conductance sense. They may be smaller than desired, however.

We next investigate finding a set of locally minimal communities. These communities represent the *best of the neighborhood*. We find that these locally minimal communities, of which there are many fewer than vertices in the graph (usually around 3%), capture the local minimal in the network community profile plot. More importantly, they can be enlarged using a local personalized PageRank community detection procedure. Afterwards, the profile of these "grown" neighborhoods is strikingly close to the profile of the PageRank communities when seeded with all vertices individually. While we do not discuss timing due to the variability in the quality of implementations, this later procedure is much faster in our experiments.

These findings have implications for future studies in community detection. One explanation for the results with the PageRank seeds is that vertex neighborhoods form the base of *any* good community in the network. This idea may guide future research into social network communities.

9. REFERENCES

[1] L. Akoglu, M. McGlohon, and C. Faloutsos. oddball: Spotting anomalies in weighted graphs. In M. Zaki, J. Yu, B. Ravindran, and V. Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6119 of *Lecture Notes in Computer Science*, pages 410–421. Springer, 2010.

[2] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

[3] R. Andersen and K. J. Lang. Communities from seed sets. In *Proceedings of the 15th international conference on the World Wide Web*, pages 223–232. ACM Press, 2006.

[4] 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*, KDD '06, pages 44–54. ACM, 2006.

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

[6] P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th WWW2011*, pages 587–596, March 2011.

[7] P. Boldi and S. Vigna. The Webgraph Framework I: Compression techniques. In *Proceedings of the 13th international conference on the World Wide Web*, pages 595–602. ACM Press, 2004.

[8] F. Bonchi, P. Esfandiar, D. F. Gleich, C. Greif, and L. V. S. Lakshmanan. Fast matrix computations for pair-wise and column-wise commute times and katz scores. *Internet Mathematics*, To appear., 2011.

[9] R. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1995.

[10] T. CAIDA. http://www.caida.org/tools/measurement/skitter/router_topology/. Accessed in 2005.

[11] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1992.

[12] J. Cohen. Graph twiddling in a MapReduce world. *Computing in Science and Engineering*, 11(4):29–41, 2009.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29:251–262, August 1999.

[14] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.

[15] U. Gargi, W. Lu, V. Mirrokni, and S. Yoon. Large-scale community detection on YouTube for topic discovery and exploration. In *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[16] J. Huang, H. Sun, Y. Liu, Q. Song, and T. Weninger. Towards online multiresolution community detection in large-scale networks. *PLoS ONE*, 6(8):e23829, August 2011.

[17] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004.

[18] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, 1998.

[19] J. Kleinberg, S. Suri, E. Tardos, and T. Wexler. Strategic network formation with structural holes. In *Proceedings of the 9th ACM conference on Electronic commerce*, EC '08, pages 284–293. ACM, 2008.

[20] D. E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, 1993.

[21] M. Kolountzakis, G. Miller, R. Peng, and C. Tsourakakis. Efficient triangle counting in large graphs via degree-based vertex partitioning. In *Algorithms and Models for the Web-Graph*, volume 6516 of *Lecture Notes in Computer Science*, pages 15–24. Springer, 2010.

[22] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 601–610. ACM, 2010.

[23] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1:1–41, March 2007.

[24] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

[25] 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. *Internet Mathematics*, 6(1):29–123, September 2009.

[26] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 631–640. ACM, 2010.

[27] R. Matei, A. Iamnitchi, and P. Foster. Mapping the Gnutella network. *Internet Computing, IEEE*, 6(1):50–57, January 2002.

[28] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[29] M. Newman. http://www-personal.umich.edu/~mejn/netdata/, 2006.

[30] M. E. J. Newman. Scientific collaboration networks. II: Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001.

[31] B. S. Rees and K. B. Gallagher. Overlapping community detection by collective friendship group inference. In *International Conference on Advances in Social Network Analysis andMining*, pages 375–379, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[32] S. E. Schaeffer. *Algorithms for Nonuniform Networks*. PhD thesis, Helsinki University of Technology, 2006.

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

[34] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

[35] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, August 2000.

[36] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *arXiv*, cs.SI:1102.2166, 2011.

[37] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1994.

[38] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998.

[39] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, EuroSys '09, pages 205–218. ACM, 2009.