Defining and Evaluating Network Communities based on Ground-truth

Jaewon Yang
Stanford University
crucis@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Nodes in real-world networks, such as social, information or technological networks, organize into communities where edges appear with high concentration among the members of the community. Identifying communities in networks has proven to be a challenging task mainly due to a plethora of definitions of a community, intractability of algorithms, issues with evaluation and the lack of a reliable gold-standard ground-truth.

We study a set of 230 large social, collaboration and information networks where nodes explicitly define group memberships. We use these groups to define the notion of ground-truth communities. We then propose a methodology which allows us to compare and quantitatively evaluate different definitions of network communities on a large scale. We choose 13 commonly used definitions of network communities and examine their quality, sensitivity and robustness. We show that the 13 definitions naturally group into four classes. We find that two of these definitions, Conductance and Triad-participation-ratio, consistently give the best performance in identifying ground-truth communities.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database Applications – *Data mining*

General Terms: Experimentation.

Keywords: Network communities, Community scores, Social and information networks.

1. INTRODUCTION

Networks are a natural way to represent social [27, 18], biological [21], technological [15], and information [6] systems. Nodes in such networks organize into densely linked groups that are commonly referred to as *network communities*, clusters or modules [9]. There are many reasons why networks organize into densely linked communities. For example, society is organized into groups, families, villages and associations [5, 10]. On the World Wide Web, topically related pages link more densely among themselves [6]. And, in metabolic networks, densely linked clusters of nodes are related to functional units, such as pathways and cycles [21].

Identifying community structure in networks [13, 4, 23, 7] has proven to be a challenging task [8, 11, 17, 16] due to three reasons: There exist multiple definitions of network communities [3, 22]; Even if we would agree on a single common definition, the formalizations of community detection lead to NP-hard problems [23]; And, the lack of reliable "ground-truth" makes evaluation extremely difficult.

Currently the performance of community detection methods is evaluated by manual inspection. For each detected community an effort is made to interpret it as a "real" community by identifying a common property or external attribute shared by all the members of the community. For example, when examining communities in a collaboration network of scientists, we might discover that such communities correspond to sets of scientists working in various areas of science [20]. Communities in social networks organize around social circles, work places, common hobbies, interests and affiliations [5, 10, 27]. Obviously, such anecdotal evaluation procedure requires extensive manual effort, is non-comprehensive and is limited to small networks as one cannot examine and attempt to interpret all the extracted communities of a large network.

Defining an appropriate notion of gold-standard ground-truth addresses the above issues. Using the ground-truth communities allows for *quantitative* and *large-scale* evaluation and comparison of different community detection methods. Such ability represents a significant step forward as the field can move beyond the current standard of anecdotal evaluation of communities to comprehensive evaluation of the performance of community detection methods.

The contributions of the present work are two fold. First, we describe a set of 230 large social, information and collaboration networks where we have a reliable notion of ground-truth communities. Second, based on the ground-truth we quantitatively evaluate 13 commonly used definitions of network communities by examining their robustness and sensitivity.

In [16], we evaluated definitions of network communities on a large scale real-world networks. However, there are two crucial differences in our work here. First, [16] used *detected* communities by the Local Spectral method [1] for the evaluation. Using communities detected by a specific community detection method would introduce a bias introduced by the detection method. In this paper, our evaluation is free from such bias as we adopt ground-truth communities which are explicitly declared by individual nodes. Second, [16] provides *qualitative* evaluation by showing the Network Characteristic Plot [17] for each definition of communities. Here we aim to quantify the robustness and sensitivity of the definitions of communities to compare which definitions are better than others.

Present work: Ground-truth communities. Next we describe our notion of ground-truth communities and argue why it corresponds to "real" communities.

Generally, after communities are identified in a given network, the essential next step is to interpret them by identifying a common

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
The 2nd MDS Workshop '12 (MDS'12), August 12, 2012, Beijing, China.
Copyright 2012 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

external property that all the members share and around which the community organizes [5]. Thus, the goal of network community detection is to identify sets of nodes with a common (often external/latent/unobserved) property based only the network connectivity structure. A "common property" can be common attribute, affiliation, role, property, or function [10, 27]. We use such common properties of nodes to define ground-truth communities. In our work we distinguish between network *communities* and *groups*. A community is defined structurally (*i.e.*, a set of nodes extracted by the community detection algorithm), while a group is defined based on nodes sharing a property around which the nodes organize in the network (*e.g.*, belonging to a common interest based group, sharing common affiliation) [5, 10].

Present work: Networks with ground-truth. We gathered 230 networks from a number of different domains and research areas where nodes explicitly state their ground-truth community memberships. The size of the networks ranges from hundreds of thousand to hundreds of millions of nodes and edges. The networks represent a wide range of edge densities, numbers of explicitly defined communities, as well as sizes and amounts of community overlap. Our collection consists of social, collaboration and information networks, where we consider various notions of ground-truth.

For example, in online social networks (like, Orkut, LiveJournal, Friendster and 225 different Ning networks) we consider explicitly defined interest based groups (*e.g.*, fans of Lady Gaga, members of a group on knitting, students of the same high school) as ground-truth communities. Such groups serve as organizing principles of nodes in social networks. They are created on specific topics, interests, hobbies, affiliations, and geographical regions. Each node can belong to zero, one or more groups. We also consider the Amazon product co-purchasing network where groups are defined by hierarchically nested product categories. Here all members of the same ground-truth community share a common function or purpose. Last, in the scientific collaboration network of DBLP we use publication venues as proxies for ground-truth research communities. Thus, for each of these datasets, we have both a network and a set of ground-truth communities. Note that we are careful to define ground-truth communities based on common affiliation, social circle, role, activity, interest, function, or some other property around which networks organize into communities [5, 10].

Present work: Methodology and findings. Communities are often studied structurally based on the connectivity patterns between members. Ground-truth communities allow us to examine how well various structural definitions of network communities correspond to real groups (*i.e.*, ground-truth communities). A good structural definition of a network community would be such that it would detect connectivity patterns that correspond to real groups. This means that we can evaluate different structural definitions based on their ability to identify connectivity structure of ground-truth communities.

We choose 13 commonly used definitions of network communities and examine their quality, sensitivity and robustness. Each such definition corresponds to a scoring function that scores a set of nodes based on the connectivity pattern between them. A higher score means that a set of nodes more closely resembles the connectivity pattern of a community. By comparing correlations of scores that different definitions assign to ground-truth communities, we find that the 13 definitions naturally group into four distinct clusters. We distinguish definitions that consider: (1) only internal community connectivity, (2) only external connectivity of the nodes to the rest of the network; (3) both internal and external community connectivity, and (4) network modularity.

Dataset	N	E	C	S	A
LiveJournal	4.0M	34.9M	311,782	40.06	3.09
Friendster	117.7M	2,586.1M	1,449,666	26.72	0.32
Orkut	3.0M	117.2M	8,455,253	34.86	95.9
Ning (225 nets)	7.0M	35.5M	137,177	46.89	0.92
Amazon	0.33M	0.92M	49,732	99.86	14.83
DBLP	0.42M	1.34M	2,547	429.79	2.56

Table 1: 230 social, collaboration and information networks with explicit ground-truth communities. N: number of nodes, E: number of edges, C: number of communities, S: average community size, A: community memberships per node. Ning statistics are aggregated over 225 different networks.

We then consider an axiomatic approach and define four intuitive properties of good communities. Intuitively, a "good" community is cohesive, compact, and internally well connected while being also well separated from the rest of the network. This allows us to characterize what connectivity patterns of ground-truth communities a given definition detects and which ones it misses. We also measure the robustness of community scoring functions based on four different types of randomized perturbations of ground-truth communities. Overall, our results show that notions of network community that are based on triadic closure [26] and the conductance [24] best capture the structure of ground-truth communities.

To the best of our knowledge our work is the first to use social and information networks with explicit community memberships to define evaluation methodology for comparing network community detection algorithms based on their accuracy on real data. We believe that the present work will bring more rigor and improve the standard for the evaluation of community detection methods.

2. COMMUNITY SCORING FUNCTIONS AND DATA SETS

We start by describing the network datasets with ground-truth communities. Then we continue with outlining 13 commonly used definitions of network communities.

Networks with ground-truth communities. Overall we consider 230 large social, collaboration and information networks, where for each network we have a graph and a set of ground-truth communities. We identify networks where nodes explicitly state their ground-truth community memberships. Members of these ground-truth communities share properties or attributes, common purpose or function. We did our best to identify networks in which such ground-truth communities can be reliably defined and identified. Networks that we study come from a variety of domains. Their size ranges from hundreds of thousand to hundreds of millions of nodes and billions of edges. Table 1 gives the dataset statistics.

First we consider online social networks (the LiveJournal blogging community [2], the Friendster online network [18], and the Orkut social network [18]) where users create explicit groups to which other users then join. Communities range from small to very large and are created based on specific topics, interests, hobbies and geographical regions. For example, LiveJournal categorizes communities into the following types: culture, entertainment, expression, fandom, gaming, life/style, life/support, sports, student life and technology. Similarly, in other social networks considered in this study users define topical communities that others then join. Each user can join to one or more communities. We define that each such explicit community is a ground-truth community. Similarly, we have a set of 225 different online social networks [12] that are all hosted on the Ning software platform. It is important to note that each Ning network is a separate social network — an independent website with a separate user community. For example, the NBA team Dallas Mavericks, rapper 50 Cent, and diabetes patients network TuDiabetes all use Ning to host their separate social networks. After joining a specific network, users then create and join topical communities. For example, communities in TuDiabetes network focus around specific types of diabetes, parenting children with diabetes,emotional and social support, different geographical regions, and similar. We focus our study on 225 different networks that each have at least 10,000 members.

The second type of network data we consider is the Amazon product co-purchasing network [15]. The nodes of the network represent products and edges link commonly co-purchased products. Each product (*i.e.*, node) belongs to one or more hierarchically organized product categories and products from the same category define a group which we view as a ground-truth community. This means members of the same Amazon community share a common function or role. Each level of the product hierarchy defines a set of hierarchically nested and overlapping communities.

Finally, we also consider the DBLP collaboration network [2] where nodes represent authors and edges connect authors that have co-authored a paper. Here we use publication venues (*i.e.*, conferences) as ground-truth communities which serve as proxies for highly overlapping scientific communities around which the network then organizes.

All our networks are complete and publicly available: LiveJournal [2], Friendster [1], Orkut [18], Ning [12], Amazon [15] and DBLP [2].[2] For each of these networks we identified a sensible way of defining ground-truth communities that serve as organizational units of these networks. The results we present here are consistent and robust across a wide range of networks and across an even wider range of groups. This gives further evidence that our approach is general and well-founded. Our work is consistent with the premise that is implicit in all community detection works: members of "real" communities share some (latent/unobserved) property or affiliation that serves as an organizing principle of the nodes and makes them well-connected in the network.

Data preprocessing. To represent all networks in a consistent way we drop edge directions and consider each network as an unweighted undirected static graph. Because members of the group may be disconnected in the network, we consider each connected component of the group as a separate ground-truth community. However, we allow ground-truth communities to be nested and to overlap (*i.e.*, a node can be a member of multiple groups at once).

Community scoring functions. We now proceed to discuss various scoring functions that characterize how community-like is a given set of nodes. The idea is that given a community scoring function, one can then find sets of nodes with high score and consider these sets as communities. All scoring functions build on the intuition that communities are sets of nodes with many connections between the members and few connections to the rest of the network. There are many possible ways to mathematically formalize this intuition. We gather 13 commonly used and representative formalizations of scoring functions, or equivalently, 13 definitions of a network community. Some scoring functions are well known in the literature, while others are proposed for the first time.

We consider a function $f(S)$ that, given a set of nodes S, characterizes how community-like is the connectivity of these nodes. Let $G(V, E)$ be an undirected graph with $n = |V|$ nodes and $m = |E|$ edges. Let S be the set of nodes, where n_S is the number of nodes in S, $n_S = |S|$; m_S the number of edges in S, $m_S = |\{(u, v) \in E : u \in S, v \in S\}|$; and c_S, the number of edges on the boundary of S, $c_S = |\{(u, v) \in E : u \in S, v \notin S\}|$; and $d(u)$ is the degree of node u. We consider 13 scoring functions $f(S)$ that capture the notion of quality of a network community S. The experiments we present later in the section show that the scoring functions naturally group into four classes:

(A) Scoring functions based on the internal connectivity of S:

- **Internal density:** $f(S) = \frac{m_S}{n_S(n_S-1)/2}$ is the internal edge density of the node set S [22].
- **Edges inside:** $f(S) = m_S$ is the number of edges between the members of S [22].
- **Average degree:** $f(S) = \frac{2m_S}{n_S}$ is the average internal degree of the members of S [22].
- **Fraction over median degree (FOMD):** $f(S) = \frac{|\{u : u \in S, |\{(u,v) : v \in S\}| > d_m\}|}{n_S}$ is the fraction of nodes of S that have internal degree higher than d_m, where d_m is the median value of $d(u)$ of all the nodes in V.
- **Triangle participation ratio (TPR):** $f(S) = \frac{|\{u : u \in S, \{(v,w) : v, w \in S, (u,v) \in E, (u,w) \in E, (v,w) \in E\} \neq \varnothing\}|}{n_S}$ is the fraction of nodes in S that belong to a triad.

(B) Scoring functions based on the external connectivity of S:

- **Expansion** measures the number of edges per node that point outside the cluster: $f(S) = \frac{c_S}{n_S}$ [22].
- **Cut Ratio** is the fraction of existing edges (out of all possible edges) leaving the cluster: $f(S) = \frac{c_S}{n_S(n-n_S)}$ [7].

(C) Scoring functions that combine internal and external connectivity of S:

- **Conductance:** $f(S) = \frac{c_S}{2m_S+c_S}$ measures the fraction of total edge volume that points outside the cluster [24].
- **Normalized Cut:** $f(S) = \frac{c_S}{2m_S+c_S} + \frac{c_S}{2(m-m_S)+c_S}$ [24].
- **Maximum-ODF (Out Degree Fraction):** $f(S) = \max_{u \in S} \frac{|\{(u,v) \in E : v \notin S\}|}{d(u)}$ is the maximum fraction of edges of a node in S that point outside S [6].
- **Average-ODF:** $f(S) = \frac{1}{n_S} \sum_{u \in S} \frac{|\{(u,v) \in E : v \notin S\}|}{d(u)}$) is the average fraction of edges of nodes in S that point out of S [6].
- **Flake-ODF:** $f(S) = \frac{|\{u : u \in S, |\{(u,v) \in E : v \in S\}| < d(u)/2\}|}{n_S}$ is the fraction of nodes in S that have fewer edges pointing inside than to the outside of the cluster [6].

(D) Scoring function based on a network model:

- **Modularity:** $f(S) = \frac{1}{4}(m_S - E(m_S))$ is the difference between m_S, the number of edges between nodes in S and $E(m_S)$, the expected number of such edges in a random graph with identical degree sequence [19].

Experimental result: Four classes of scoring functions. Given 13 community scoring functions we investigate relationship between different scoring functions. We performed the following experiment: For each of the 10 million ground-truth communities in our networks, we compute the score using each of the 13 scoring functions. We then create a correlation matrix between scoring functions and threshold it. Fig. 1 shows connections between scoring functions with correlation ≥ 0.6 (on the LiveJournal network). We observe four connected components of scores which results in

[1] http://archive.org/details/friendster-dataset-201107

[2] All networks and the corresponding ground-truth communities are available at http://snap.stanford.edu/data

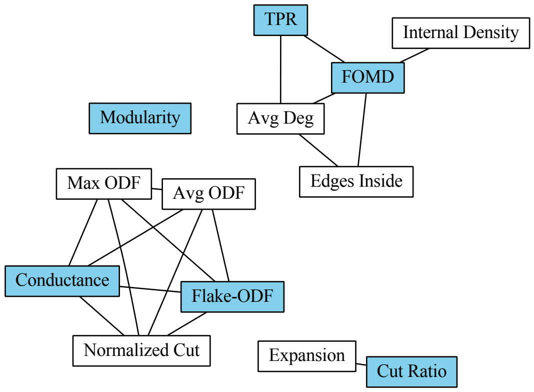


Figure 1: Correlations between community scoring functions.

the grouping above. Overall, none of the scoring functions are negatively correlated, which means that none of the scoring functions systematically disagree. Interestingly, Modularity is not correlated with any other scoring function (Average degree is the most correlated at 0.05 correlation). We observe similar correlation of scoring function in all other data sets as well.

The experiment demonstrates that even though many different notions of network community have been proposed, in the end these notions are heavily correlated. Essentially there are only 4 different notions of community scoring functions as revealed by Fig. 1. In the rest of the paper we only present results for the 6 representative scoring functions (blue nodes in Figure 1).

3. EVALUATION OF COMMUNITY SCORING FUNCTIONS

The second main purpose of the paper is to develop evaluation methodology for network community detection. Based on ground-truth communities we now aim to compare and evaluate different community scoring functions.

Community goodness metrics. Our goal is to rank different definitions of a network community (*i.e.*, community scoring functions) by their ability to detect ground-truth communities. We adopt an axiomatic approach and proceed as follows. First, we define four community "goodness" metrics that formalize the intuition that "good" communities are both compact and well connected internally while being relatively well-separated from the rest of the network.

The difference between community scoring functions from Section 2 and the goodness metrics defined above is that a community scoring function quantifies how community-like a set is, while a goodness metric in an axiomatic way quantifies a desirable property of a community. A set with high goodness metrics does not necessarily correspond to a community, but a set with high community score should have a high value on one or more goodness metrics. In other words, the goodness metrics shed light on various (in many cases mutually exclusive) aspects of the network community structure.

Using the notation from Section 2, we define four goodness metrics $g(S)$ for a node set S:

- **Separability** captures the intuition that good communities are well-separated from the rest of the network [24, 7], meaning that they have relatively few edges pointing from set S to the rest of the network. Separability measures the ratio between the internal and the external number of edges of S: $g(S) = \frac{m_S}{c_S}$.
- **Density** builds on intuition that good communities are well connected [7]. One way to capture this is to characterize the fraction of the edges (out of all possible edges) that appear between the nodes in S, $g(S) = \frac{m_S}{n_S(n_S-1)/2}$.
- **Cohesiveness** characterizes the internal structure of the community. Intuitively, a good community should be internally well and evenly connected, *i.e.*, it should be relatively hard to split a community into two sub communities. We characterize this by the conductance of the internal cut. Formally, $g(S) = \max_{S' \subset S} \phi(S')$ where $\phi(S')$ is the conductance of S' measured in the induced subgraph by S. Intuitively, conductance measures the ratio of the edges in S' that point outside the set and the edges inside the set S'. A good community should have high cohesiveness (high internal conductance) as it should require deleting many edges before the community would be internally split into disconnected components [16].
- **Clustering coefficient** is based on the premise that network communities are manifestations of locally inhomogeneous distributions of edges, because pairs of nodes with common neighbors are more likely to be connected with each other [26].

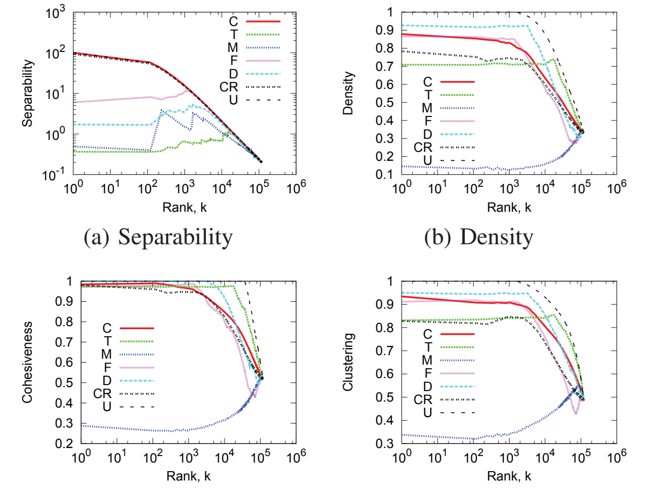
Experimental setup. We are interested in quantifying how "good" are the communities chosen by a particular scoring function $f(S)$ by evaluating their goodness metric. We formulate our experiments as follows: For each of 230 networks, we have a set of ground-truth communities S_i. For each community scoring function $f(S)$, we rank the ground-truth communities by the decreasing score $f(S_i)$. We measure the cumulative running average value of the goodness metric $g(S)$ of the top-k ground-truth communities (under the ordering induced by $f(S_i)$).

The intuition for the experiments is the following. A perfect community scoring function would rank the communities in the decreasing order of the goodness metric and thus the cumulative running average of the goodness metric would decrease monotonically with k. While if a hypothetical community scoring function would randomly rank the communities, then the cumulative running average would be a constant function of k.

Experimental results. Figure 2(a) shows the results by plotting the cumulative running average of separability for LiveJournal[3] ground-truth communities ranked by each of the six community scoring functions. Curve "U" presents the upper bound, *i.e.*, it plots the cumulative running average of separability when ground-truth communities are ordered by decreasing separability. We observe that Conductance (C) and Cut Ratio (CR) give near optimal performance, *i.e.*, they nearly perfectly order the ground-truth communities from the most separable to the least separable. On the other hand, we observe that Triad participation ratio (T) and Modularity (M) essentially score ground-truth communities in the inverse order of separability (especially for $k < 100$), which means that they both prefer densely linked sets of nodes.

Similarly, Figures 2(b), (c), and (d) show the cumulative running average of community density, cohesiveness and clustering coefficient. We observe that all scoring functions (except Modularity) rank denser, more cohesive and more clustered ground-truth communities higher. For the density metric, the Fraction over median degree (D) score performs best for high values of k followed by Conductance (C) and Flake-ODF (F). In terms of cohesiveness and clustering coefficient, the Triad participation ratio (T) score gives by far the best results. In all cases the only exception seems to be

[3]For brevity we show plots for the LiveJournal network. Qualitatively similar results are obtained for all other datasets (Appendix).



(a) Separability (b) Density

(c) Cohesiveness (d) Clustering coefficient

Figure 2: Cumulative average of goodness metrics for LiveJournal communities ranked by each of the six community scoring functions.

Scoring function	Separability	Density	Cohesiveness	Clustering
Conductance (C)	**1.0**	3.5	3.4	3.1
Flake-ODF (F)	3.9	3.6	3.5	4.3
FOMD (D)	4.9	3.0	2.9	2.9
TPR (T)	4.5	**2.3**	**2.1**	**1.2**
Modularity(M)	4.0	5.5	5.7	3.9
CutRatio (CR)	2.6	3.1	3.2	5.5

Table 2: Average scoring function rank for each goodness metric across all the 230 networks. Conductance gives the highest Separability while Triad participation ratio (TPR) scores best on the remaining metrics.

that the Modularity score actually ranks the communities in nearly reverse order of the goodness metric. Since the cumulative running average increases as a function of k, this means that communities with low density, cohesiveness or clustering coefficient are scored higher by the Modularity score. We note that these are all well-known issues of the Modularity score [8, 11] but they get further attenuated when tested on large networks.

The curves in Figure 2 also illustrate the ability of the community scoring functions to rank communities. To quantify this we perform the following experiment. For a given goodness metric g and for each community scoring function f, we measure the rank of each scoring function in comparison to other scoring functions at every value of k. For example, the score that consistently ranks communities with larger separability higher than any other score would have an average rank of 1. In Figure 2(a), the rank of Conductance at $k = 100$ is 1, Cut ratio 2, Flake-ODF 3, FOMD 4, Modularity 5, and TPR 6. Now for every k, we rank the scores and compute the average rank over all values up to k, which quantifies the ability of the scoring function to identify communities with high value of the goodness metric.

Our results show that Conductance and Triad Participation Ratio consistently give the best performance in ranking ground-truth communities. In particular, Table 2 shows the average rank for each score and each goodness metric. An average rank of 1 means that a particular score always outperforms other scores, while rank of 6 means that the score gives worst ranking out of all 6 scores. We observe that Conductance (C) performs best in terms of Separability but relatively bad in the other three metrics. For Density, Cohesiveness and Clustering coefficient, Triad participation ratio (T) gives by far best results and outperforms all other scoring functions. Perhaps not surprisingly, Triad participation ratio scores badly on Separability of ground-truth communities. Thus, Conductance is able to identify well-separated communities, but performs poorly in identifying dense and cohesive sets of nodes with high clustering coefficient. On the other hand, Triad participation ratio gives the worst performance in terms of Separability but scores the best for the other three metrics.

We conclude that depending on the network different definitions of network communities might be appropriate. When the network contains well-separated non-overlapping communities, Conductance is the best scoring function. When the network contains dense heavily overlapping communities, then the Triad participation ratio defines the most appropriate notion of a community.

Lastly, in Figure 2 we observe that the average goodness metric of the top k communities remains flat but then quickly degrades. We observe the same pattern in all our data sets. Thus, for the remainder of the paper we focus our attention to a set of the top 5,000 communities of each network based on the average rank over the 6 scoring functions.

4. ROBUSTNESS OF COMMUNITY SCORING FUNCTIONS

So far we have examined the ability of different scoring functions to rank ground-truth communities according to their goodness. In the following section, we evaluate community scoring functions using a set of perturbation strategies for communities. We develop a set of strategies to generate randomized perturbations of ground-truth communities, which allows us to investigate robustness and sensitivity of community scoring functions. Intuitively, a good community scoring function should be such that it is stable under small perturbations but degrades quickly as the perturbation get larger.

Our reasoning is as follows. We desire a community scoring function that scores well when evaluated on a ground-truth community but scores low when evaluated on a perturbed community. In other words, an ideal community scoring function should give a maximal value when evaluated on the ground-truth community. If we consider a slightly perturbed ground-truth community (*i.e.*, a node set that differs very slightly from the ground-truth community), we would want the score to be nearly as good as the score of the original ground-truth community. This would mean that the scoring function is robust to noise and small perturbations of the original community. However, if the ground-truth community is perturbed a lot and starts to resemble a random set of nodes, then a good scoring function should give it a low score.

Community perturbation strategies. We proceed by defining a set of community perturbation strategies. To vary the amount of perturbation, each perturbation strategy has a single parameter p that controls the intensity of the perturbation and thus the resemblance of the perturbed community to the ground-truth community. Given p and a ground-truth community defined by its members S, the community perturbation starts with S and then modifies it (*i.e.*, removes some nodes from S and adds others to it) by executing the perturbation strategy $p|S|$ times. We define the following perturbation strategies:

- NodeSwap perturbation is based on the mechanism where nodes at the boundary of the community swap memberships and so the community memberships diffuse from the original community through the network. We achieve this by picking a random edge (u, v) where $u \in S$ and $v \notin S$ and then swap the memberships (*i.e.*, remove u from S and add v). Note that NODESWAP preserves the size of S but if v is not connected to the nodes in S, then NODESWAP makes S disconnected.

- RANDOM takes community members and replaces them with random non-members. We pick a random node $u \in S$ and a random $v \notin S$ and then swap the memberships. Like NODESWAP, RANDOM maintains the size of S but may disconnect S. Generally, RANDOM will degrade the quality of S faster than NODESWAP, since NODESWAP only affects the "fringe" of the community.
- EXPAND perturbation grows the membership set S by expanding it at the boundary. We pick a random edge (u, v) where $u \in S$ and $v \notin S$ and add v to S. Adding v to S will generally decrease the quality of the community. EXPAND preserves the connectedness of S but increases the size of S.
- SHRINK is the opposite of EXPAND as it removes members from the community boundary. We pick a random edge (u, v) where $u \in S, v \notin S$ and remove u from S. Removing u will decrease the quality of S and result in a smaller community while preserving connectedness.

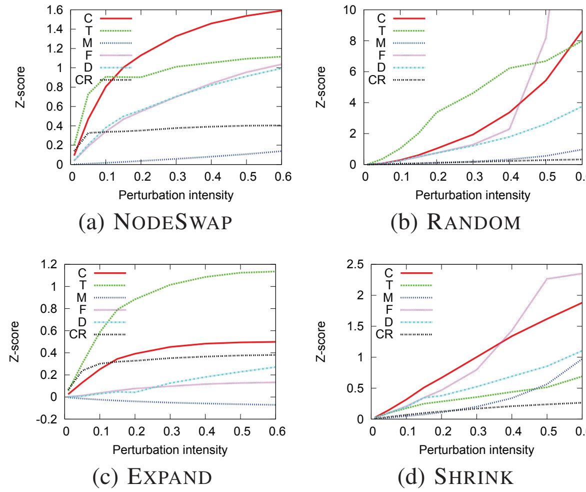
For a given community S, we let $h(S, p)$ denote a perturbed version of the community generated by the perturbation h with intensity p.

We now quantify the difference of the score between the unperturbed ground-truth community and the perturbed community. We use the Z-score, which measures in the units of standard deviation how much the scoring function changes as a function of perturbation intensity. Given ground-truth communities S_i, the Z-score $Z(f, h, p)$ of community scoring function f under perturbation strategy h with intensity p is:

$$Z(f, h, p) = \frac{E_i[f(S_i) - f(h(S_i, p))]}{\sqrt{Var_i[f(h(S_i, p))]}},$$

where $E_i[\cdot], Var_i[\cdot]$ are the mean and the variance over communities S_i, and $f(h(S_i, p))$ is the community score of perturbed S_i under perturbation h with intensity p. To measure $f(h(S_i, p))$, we run 20 trials of $h(S_i, p)$ and compute the average value of f. Z-score is the difference between the average community score of true communities $f(S_i)$ and the average community scores of perturbed communities $f(h(S_i, p))$ normalized by the standard deviation of community scores of perturbed communities. Since $f(h(S_i, p)$ are independent for each i, $E_i[f(h(S_i, p))]$ follows a Normal distribution by the Central Limit Theorem. Thus, $P(z < Z(f, h, p))$ gives the probability that $E_i[f(h(S_i, p))] > E_i[f(S_i)]$ where z is a standard normal random variable. We measure f so that lower values mean better communities, *i.e.*, we add a negative sign to TPR, Modularity and FOMD. High Z-scores mean that $E_i[f(S_i)]$ is likely to be smaller than $E_i[f(h(S_i, p))]$ and that S_i is better than $h(S_i, p)$ in terms of f.

Experimental results. For each of the 6 community scoring functions, we measure Z-score for perturbation intensity p ranging between 0.01 and 0.6. This means that we randomly swap between 1% and 60% of the community members and measure the Z-score for each of the 6 community scoring functions. For small p, small Z-scores are desirable since they would indicate that the scoring function is robust and does not change much under small perturbations of communities. For high perturbation intensities p, high Z-scores are preferred because this would suggest that the community scoring function is sensitive, *i.e.*, as the community becomes more "random" we want the community scoring function to significantly increase its value.



(a) NODESWAP (b) RANDOM (c) EXPAND (d) SHRINK

Figure 3: Z-scores as a function of the perturbation intensity of each of the 6 scoring functions for the LiveJournal communities. Conductance (C) and Triad participation ratio (T) best detect the perturbations of a ground-truth community.

Scoring function	NodeSwap	Random	Expand	Shrink
Conductance (C)	**1.06**	1.59	0.50	**0.45**
Flake-ODF (F)	0.51	1.15	0.11	0.41
FOMD (D)	0.18	0.57	0.19	0.12
TPR (T)	0.37	**1.85**	**0.74**	0.21
Modularity(M)	0.23	0.14	0.03	0.15
CutRatio (CR)	0.53	0.83	0.13	0.43

Table 3: Average absolute increment of the Z-score between small and large community perturbations over all the 230 networks. We bold the best performing scoring functions.

Figure 3 shows the Z-scores of LiveJournal communities as a function of perturbation intensity p. We plot the Z-score for each of the 6 community scoring functions and the four perturbation strategies. As expected, the Z-scores increase with p, which means that as the community gets more perturbed, and thus more "random", the value of the score tends to decrease. However, the faster the increase the more sensitive and thus the better the score is. For example, under the NODESWAP perturbation Conductance (C) exhibits the highest Z-score after $p > 0.2$, and it has the steepest curve. Triad participation ratio (T) also exhibits desirable behavior. On the other hand, Modularity (M) score does not change much as we perturb the ground-truth community. This means that Modularity has a difficult time distinguishing true communities from randomized sets of nodes. For the RANDOM perturbation, the Z-score of the Triad participation ratio (T) is the highest when $p > 0.4$. Conductance (C) and Flake-ODF (F) also exhibit relatively good sensitivity towards the RANDOM perturbations. For the EXPAND perturbation, Triad participation ratio (T) clearly performs best followed by Conductance (C) and Cut ratio (CR). Interestingly, Modularity (M) shows decreasing Z-score. This means that if we randomly expand the community (and thus increase its size), then Modularity keeps increasing. This phenomenon is known as the resolution limit of modularity [8] as it is unable to detect small communities. Last, for the SHRINK perturbation we observe a monotonic increase of Z-scores for all scoring functions. We note very similar results on all of the remaining datasets considered in this study (Appendix).

Sensitivity of community scoring functions. We also quantify the sensitivity of community scoring functions by computing the increase of the Z-score between small ($p = 0.05$) and large perturbations ($p = 0.2$). As noted above, we prefer a community scoring function with fast increase of the Z-score as the community perturbation intensity increases. Table 3 displays the difference of the Z-score between a large and a small perturbation: $Z(f, h, 0.2) - Z(f, h, 0.05)$. We compute the average increment across all the 230 networks. A high value of increment means that the score is both robust and sensitive. The score is robust because even at small perturbation ($p = 0.05$) it maintains low Z-value, while at large perturbation ($p = 0.2$) it has high Z-value and thus the overall Z-score increment is high. We also observe that the Conductance is the most robust score under the NODESWAP and SHRINK. The Triad participation ratio (T) is the most robust under RANDOM and EXPAND. In both cases it is relatively closely followed by the Conductance.

5. CONCLUSION

The lack of reliable ground-truth gold-standard communities has made network community detection a very challenging task. In this paper, we studied a set of 230 different large social, collaboration and information networks in which we defined the notion of ground-truth communities by nodes *explicitly* stating their group memberships. The size of the networks ranges from hundreds of thousand to hundreds of millions of nodes and billions of edges. The networks represent a wide range of edge densities, numbers of explicitly defined communities, as well as sizes and amounts of community overlap.

We then developed an evaluation methodology for comparing network community detection algorithms based on their accuracy on real data and compared different definitions of network communities and examined their robustness. Our results demonstrate large differences in behavior of community scoring functions. We find that Conductance and Triad participation ratio perform best, while the Modularity score gives relatively poor results.

Our work here and the availability of ground-truth communities allows for a range of interesting future directions. For example, it would be interesting to investigate network community detection as a supervised (rather than an unsupervised) problem. The challenge here is that a network is a single object, which we cannot simply split into a training and a test dataset. Similarly, further examining the connectivity structure of ground-truth communities could lead to novel insights, models and algorithms for network community detection. Overall, we believe that the present work will bring more rigor and increase the standards for the evaluation of network community detection methods, and the datasets publicly released as a part of this work will benefit the research community.

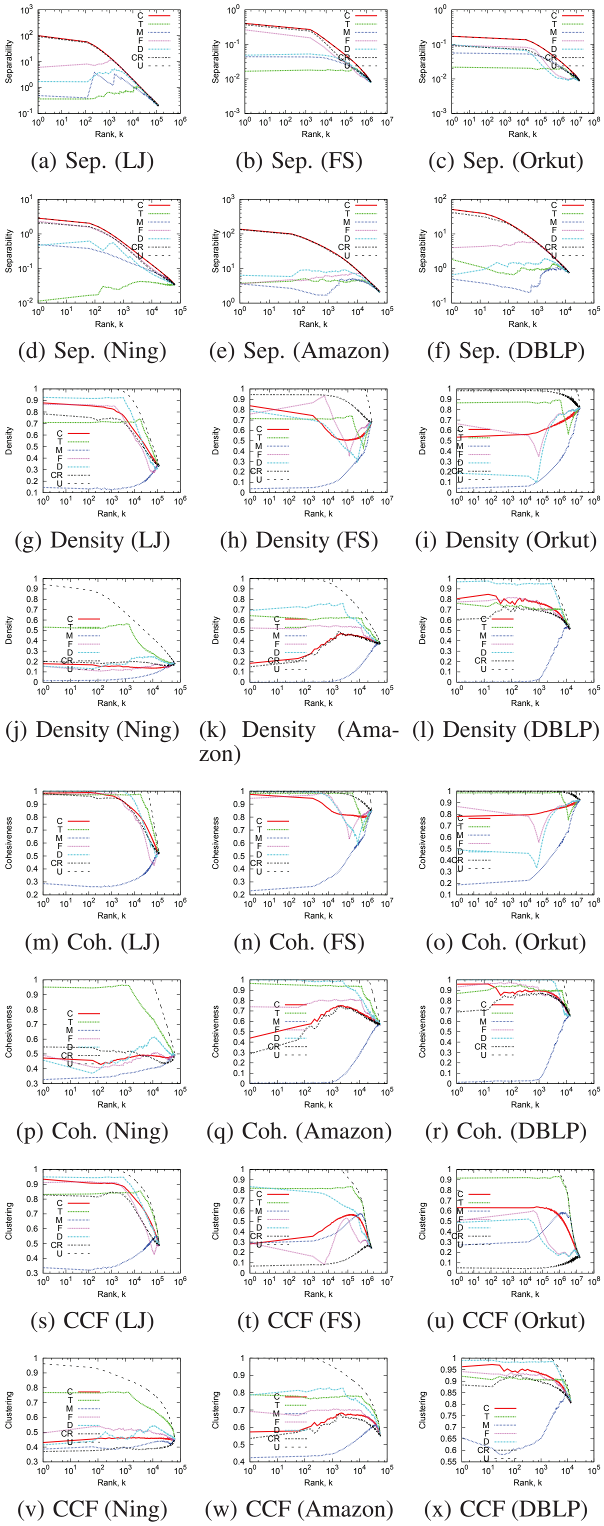
Acknowledgements. This research has been supported in part by NSF CNS-1010921, IIS-1016909, IIS-1149837, IIS-1159679, Albert Yu & Mary Bechmann Foundation, Boeing, Allyes, Samsung, Alfred P. Sloan and the Microsoft Faculty Fellowship.

6. REFERENCES

[1] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *FOCS '06*, pages 475–486, 2006.
[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06*, pages 44–54, 2006.
[3] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *Journal of Stat. Mech.: Theory and Experiment*, 29(09):P09008, 2005.
[4] I. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on PAMI*, 29(11):1944–1957, 2007.
[5] S. L. Feld. The focused organization of social ties. *American Journal of Sociology*, 86(5):1015–1035, 1981.
[6] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In *KDD '00*, pages 150–160, 2000.
[7] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010.
[8] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *PNAS*, 104(1):36–41, 2007.
[9] M. Girvan and M. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.
[10] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.
[11] R. Guimerà, M. Sales-Pardo, and L. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70:025101, 2004.
[12] S. Kairam, D. Wang, and J. Leskovec. The life and death of online groups: Predicting group growth and longevity. In *WSDM '12*, 2012.
[13] 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.
[14] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.
[15] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM TWEB*, 1(1), 2007.
[16] J. Leskovec, K. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW '10*, 2010.
[17] 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, 2009.
[18] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07*, pages 29–42, 2007.
[19] M. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.
[20] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 2004.
[21] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
[22] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *PNAS*, 101(9):2658–2663, 2004.
[23] S. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.
[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transcations on PAMI*, 2000.
[25] D. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *STOC '04*, pages 81–90, 2004.
[26] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.
[27] W. Zachary. An information flow model for conflict and fission in small groups. *J. of Anthropological Research*, 1977.

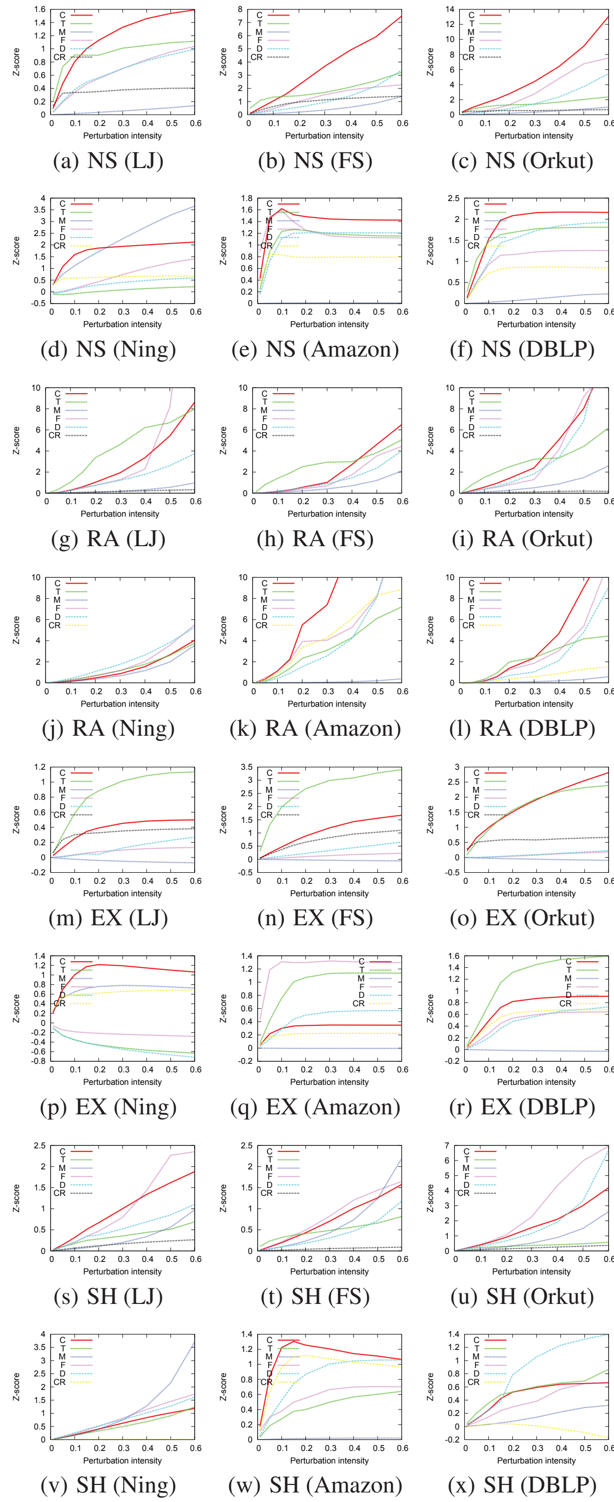
APPENDIX

A. APPENDIX



(a) Sep. (LJ) (b) Sep. (FS) (c) Sep. (Orkut)
(d) Sep. (Ning) (e) Sep. (Amazon) (f) Sep. (DBLP)
(g) Density (LJ) (h) Density (FS) (i) Density (Orkut)
(j) Density (Ning) (k) Density (Amazon) (l) Density (DBLP)
(m) Coh. (LJ) (n) Coh. (FS) (o) Coh. (Orkut)
(p) Coh. (Ning) (q) Coh. (Amazon) (r) Coh. (DBLP)
(s) CCF (LJ) (t) CCF (FS) (u) CCF (Orkut)
(v) CCF (Ning) (w) CCF (Amazon) (x) CCF (DBLP)

Figure 4: Average metrics of top k communities by the scores. Sep.: Separability, Coh.: Cohesiveness.



(a) NS (LJ) (b) NS (FS) (c) NS (Orkut)
(d) NS (Ning) (e) NS (Amazon) (f) NS (DBLP)
(g) RA (LJ) (h) RA (FS) (i) RA (Orkut)
(j) RA (Ning) (k) RA (Amazon) (l) RA (DBLP)
(m) EX (LJ) (n) EX (FS) (o) EX (Orkut)
(p) EX (Ning) (q) EX (Amazon) (r) EX (DBLP)
(s) SH (LJ) (t) SH (FS) (u) SH (Orkut)
(v) SH (Ning) (w) SH (Amazon) (x) SH (DBLP)

Figure 5: Z-score of 6 scores versus the perturbation intensity for each null model. NS: NodeSwap, RA: Random, EX: Expand, SH: Shrink.