An Empirical Analysis of Communities in Real-World Networks

Chuan Sheng Foo
Computer Science Department
Stanford University
csfoo@cs.stanford.edu

ABSTRACT

Little work has been done on the characterization of community structure as opposed to the design of algorithms for detecting communities. In this work, we perform an analysis of real-world communities by computing a variety of features and comparing their distributions to those of features computed on randomly generated communities. We find that real-world communities are more internally well-connected as compared to their random counterparts, and that conductance does not necessarily differentiate real from random communities.

1. INTRODUCTION

Much work in the field of networks has focused on the characterization and modelling of various properties of real-world networks. This includes properties such as power-law degree distributions, the small-world property of networks, densification of networks, and shrinking diameters [6]. The characterization of these global network properties has led to the development of network models that are able to generate new networks that accurately replicate the characteristics of their real-world counterparts [5]. In addition, recent work has also focused on discovering the local structure of networks, investigating the process of how new connections are formed in real-world networks [4].

In comparison, far less work has been done on the intermediate level of organization in networks, which involves the characterizion of groups or communities in networks. Instead, most work on the community level of network organization has been focused on algorithms for detecting communities in networks. Such approaches tend to be unsupervised, and focus on finding groups of nodes that maximize some pre-defined quality score that is thought to accurately describe the properties of a community (see [2] for a recent review and discussion).

Ideally, one would like to have a community detection algorithm that could automatically learn the salient characteristics of given sample communities in networks, and use this information to detect other novel communities. In other words, a *supervised* approach to community detection that uses empirically derived properties of communities to perform detection. The key advantage to such an approach lies precisely in this use of empirically derived properties of communities. This eliminates the need to incorporate predefined ideas of what communities look-like, that may not be an accurate representation of reality, into the design of the detection algorithm.

This work is a first step towards the goal of supervised community detection, by providing an empirical analysis of various communities in real-world networks. It may be seen as an extension of the work in [7], which analyzes community structure in real-world networks, where communities are defined by subsets of nodes with low conductance. In comparison, we analyze *ground-truth communities* through the computation of various features on communities and the comparison of their distributions with various randomized baselines. The reasoning behind this approach is that features which discriminate between real and randomized communities are those which define a community.

Our analysis yielded greater insight into the structure of a community, showing that a common intution that a community is a group of nodes in a network that are more internally well-connected than to the rest of the network is not entirely true. Specifically, we discovered that *communities are not necessarily easier to separate from the rest of the network as compared to their randomized counterparts* and that *true communities are indeed more internally well-connected than their randomized counterparts.* Besides having ramfications for community detection algorithms based on these assumptions of community structure, our discoveries also suggest new hypothesis about the structure of communities that we describe in the discussion section.

2. METHODOLOGY

In order to understand the structure of real-world communities, our approach was to attempt to find various features of communities that would allow us to discriminate them from random subsets of nodes. We therefore computed a variety of features on communities, and then compared them to the values of the same features but on “randomized” versions of the communities. We will first describe the features that we computed and then describe the process of how “randomized” communities were constructed for the purposes of comparison.

2.1 Features used to describe communities

Let $G(V, E)$ be an undirected graph with $n = |V|$ nodes and $m = |E|$ edges. A *community* is then defined to be a subset of nodes $S \subseteq V$ in G. Then, we may define the following quantities: the number of nodes in S, $n_S = |S|$; the number of edges in S, $m_S = |\{(u, v) : u \in S, v \in S\}|$; and c_S, the number of edges on the boundary of S, $c_S = |\{(u, v) : u \in S, v \notin S\}|$. Also, let $d(v)$ denote the degree of node v.

We computed the following 17 features for each community S:

- **Edges inside:** m_s, the number of edges inside the community
- **Edges cut:** c_S, number of edges needed to be removed to disconnect nodes in S from the rest of the network.
- **Expansion:** $\frac{c_S}{n_S}$, measures the number of edges per node that point outside the community
- **Internal density:** $1 - \frac{m_S}{n_S(n_S-1)/2}$ is the internal edge density of the community S
- **Conductance:** $\frac{c_S}{2m_S+c_S}$ measures the fraction of total edge volume that points outside the community [9, 3].
- **Normalized Cut:** $\frac{c_S}{2m_S+c_S} + \frac{c_S}{2(m-m_S)+c_S}$ [9].
- **Cut Ratio:** $\frac{c_S}{n_S(n-n_S)}$, the fraction of all possible edges leaving the community
- **Maximum-ODF (Out Degree Fraction):** $\max_{u \in S} \frac{|\{(u,v):v \notin S\}|}{d(u)}$, the maximum fraction of edges of a node pointing outside the community [1].
- **Average-ODF:** $\frac{1}{n_S} \sum_{u \in S} \frac{|\{(u,v):v \notin S\}|}{d(u)}$, the average fraction of nodes' edges pointing outside the community [1].
- **Flake-ODF:** $\frac{|\{u:u \in S, |\{(u,v):v \in S\}| < d(u)/2\}|}{n_S}$, the fraction of nodes in S that have less edges pointing inside than to the outside of the community [1].
- **Volume:** $\sum_{u \in S} d(u)$, sum of degrees of nodes in S.
- **Number of open triads**
- **Number of closed triads**
- **Ratio of number of open triads to closed triads**
- **Clustering coefficient**
- **Size of largest connected component in S**
- **Size of the largest connected component as a fraction of the size of S**

Table 1: Statistics for networks used in this work

Dataset	# Nodes	# Edges	# Communities
DBLP	851523	1135266	3050
Flickr	584207	2257488	14051
LinkedIn	7550955	29161896	147
LiveJournal	4847571	42851237	385959

2.2 Random baselines used for comparison

To determine which of the abovementioned features truly characterizes real communities, we sought to see if the distributions of these features would remain the same on communities that were randomly generated. To do so, we adopted the following 3 procedures to generate random communities:

- **Network rewiring:** We simply keep the node sets of the communities the same, but we rewire the underlying network using an edge switching procedure as described in [8]. This results in communities having different internal edge structure due to the rewired network.
- **Randomized memberships:** We randomly assign nodes in the graph to communities while keeping the number of communities, the number of members in each community and the number of communities each node is a member of the same as in the original data. This is achieved using a membership swapping technique, where we pick two random nodes and swap their community memberships for a randomly chosen community from each node.
- **Randomized memberships (neighbor constrained):** This is a similar procedure as the randomized memberships case, but where nodes are required to be neighbors in the graph. To do this, we pick a random node from the graph, and then pick one of its neighbors to execute the membership swap as before.

3. EXPERIMENTAL SETUP

We computed the 17 described statistics on communities and their randomized counterparts using the 3 methods previously described, on the following 4 datasets. Basic statistics of these datasets may be found in Table 1.

- **DBLP co-authorship network:** The DBLP network is constructed from publications at various Computer Science conferences and journals. Authors are nodes in the network, while edges are defined by co-authorship on a publication; each publication gives rise to a clique on the authors of the publication. Communities in this network are defined by the publication venue, *e.g.,* KDD, ICML, FOCS.
- **Flickr photo-sharing network:** Members of the Flickr network can post and tag photos. There is also an underlying social network of friendships, which we used as the network in this work. Communities are then defined by co-tagging – two members are in the same community if they have photos tagged by other members with the same tag; we did not consider tags provided by members on their own photos.

- **LinkedIn social network:** Here we have data on the underlying social network, as well as community data defined by the various industries that members work in. Due to the coarse definition of an industry, the communities in this dataset tend to be rather large (at least a few thousand members).
- **LiveJournal blogging network:** This is the network of bloggers and their friends. LiveJournal has a concept of communities which bloggers join based on similar interests. We used these communities as the ground-truth in this work. This dataset is probably the most useful of the 4 since the communities are indeed user-defined and not defined by us in this work.

We then plotted the average and median values of each feature versus the community size, as well as the distribution for each feature over various community size ranges. For each feature, we manually inspected the plots to determine if the feature was significant, by checking for a separation between the points for the original communities and their randomized counterparts.

4. RESULTS

4.1 Communities may not have low conductance

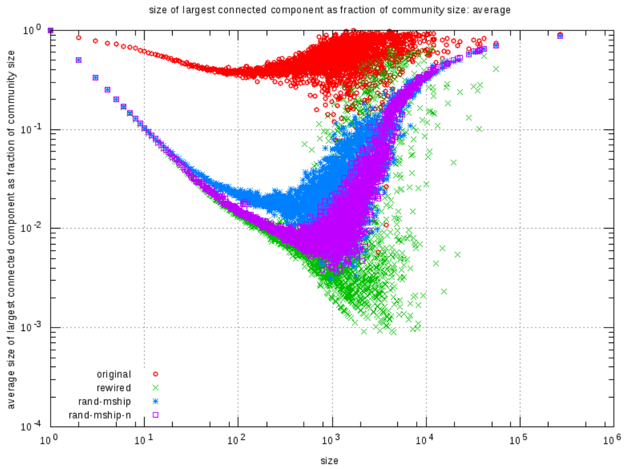
As there are many algorithms based on the assumption that communities are subgraphs that are "easily separated" from the network (for example, those based on graph-cuts), we decided to investigate if real communities have this property. One popular measure used to quantify this notion that a community is "easily separated" is *conductance* (see section 2.1 for a more precise definition) – if a community has low conductance, it is more easily separated from the rest of the network.

We found that the data does not entirely support this assumption. From the plots of the average conductance versus community size shown in Figure 1, we observe that the conductance of ground-truth communities in all the networks is not appreciably different from those of randomized communities in the Flickr and LiveJournal networks, across all community sizes.

We observe some separation in the DBLP plot, but this is probably an artefact of the way the network was constructed. In the construction of the network, all authors in a single publication are presumed to be connected, resulting in the network being composed of a network of overlapping cliques. This could result in a lower conductance score since there are likely to be more edges inside the communities due to the cliques. The separation of the LinkedIn plot is also possibly due to the coarse definitions of the industries which were used to define communities, and the properties of the network. As it is a social network used for professional networking, people are more likely to have friends within their own industries, and if the industries are coarsely defined, then there are few cross-community edges, resulting in a lower conductance.

4.2 Communities have greater internal connectivity

Figure 3: Plot of the average of the size of largest connected component in the community as a fraction of the community size, versus community size for the LiveJournal dataset. Each point in the plot gives the average fraction over all communities of the particular size given on the x-axis.



While real communities may not necessarily be more easily separated from the rest of the network than random communities, they have greater internal connectivity than their randomized counterparts. This is shown in the plots of Figure 2, which show the average number of edges inside the community versus community size. As opposed to the plots for conductance, here there is a clear separation between the true communities and their randomized counterparts. In all 4 networks, we see that across all community sizes, the true communities have a greater number of edges inside. We also see that communities on the rewired network tend to have the least number of edges inside, reflecting the fact that rewiring the network may not be a good random baseline as it destroys much of the network structure.

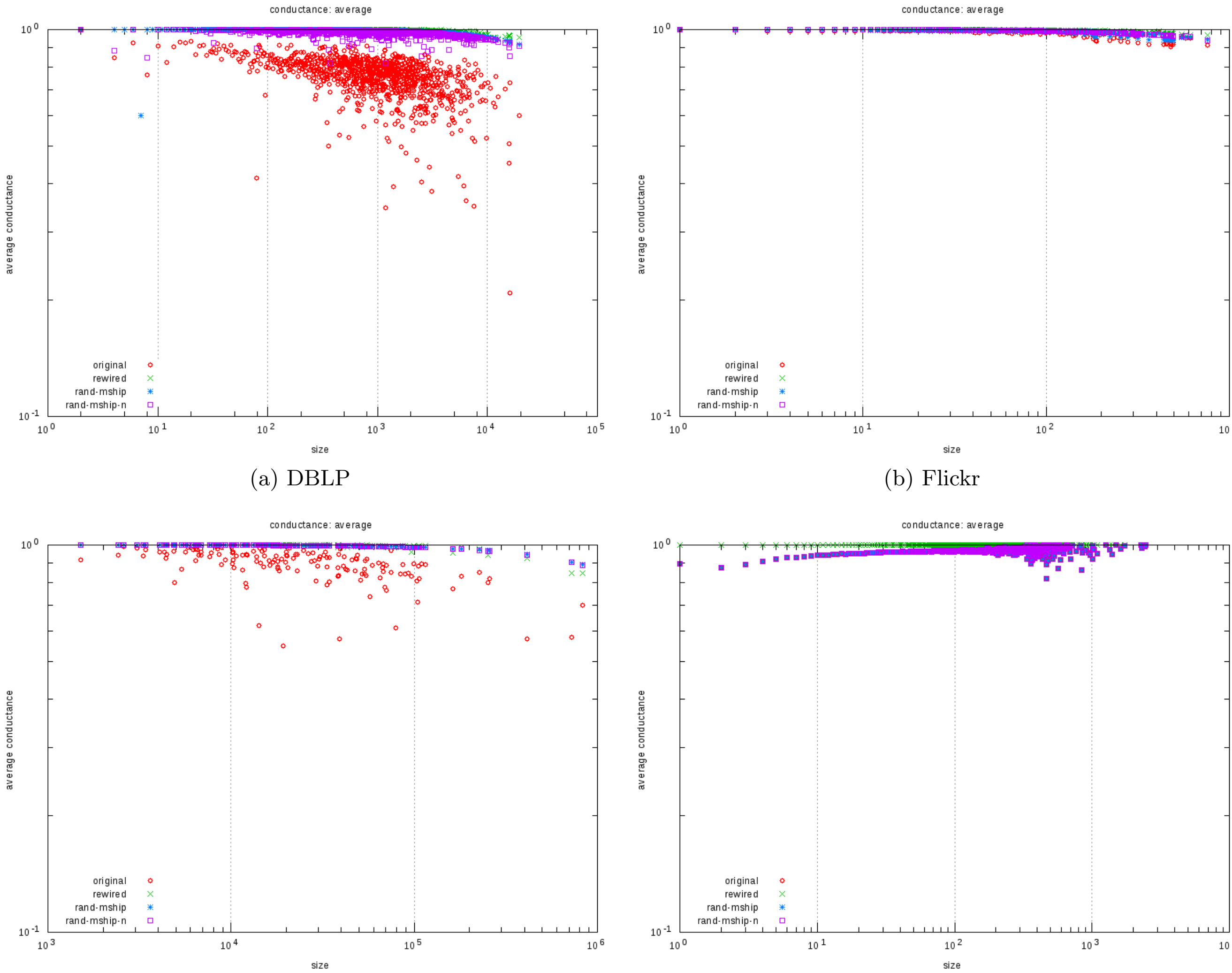
In addition, in plots of the size of the largest connected component in the community as a fraction of the community size, we found that real communities have a much higher fraction of their nodes in the largest connected component as compared to random communities. An illustrative plot is shown in Figure 3 for the LiveJournal dataset; the observed trend is similar for the other 3 networks, but the separation was not as significant.

5. CONCLUSIONS AND FUTURE WORK

In this work we managed to gain a better understanding into community structure by analysing the distributions of various features computed on real and random communities. Our observations suggest that communities may not have low conductance but have greater internal connectivity. Out of the 17 measures we computed, most others did not show any significant signal that could be used to differentiate real communities from random ones.

Another possible explanation for our observations with regards to the conductance plots is that we are overcounting the number of edges on the boundary of the network – such

Figure 1: Plots showing the average conductance versus community size. Each point in the plot represents the average conductance over all communities of the particular size given on the x-axis.



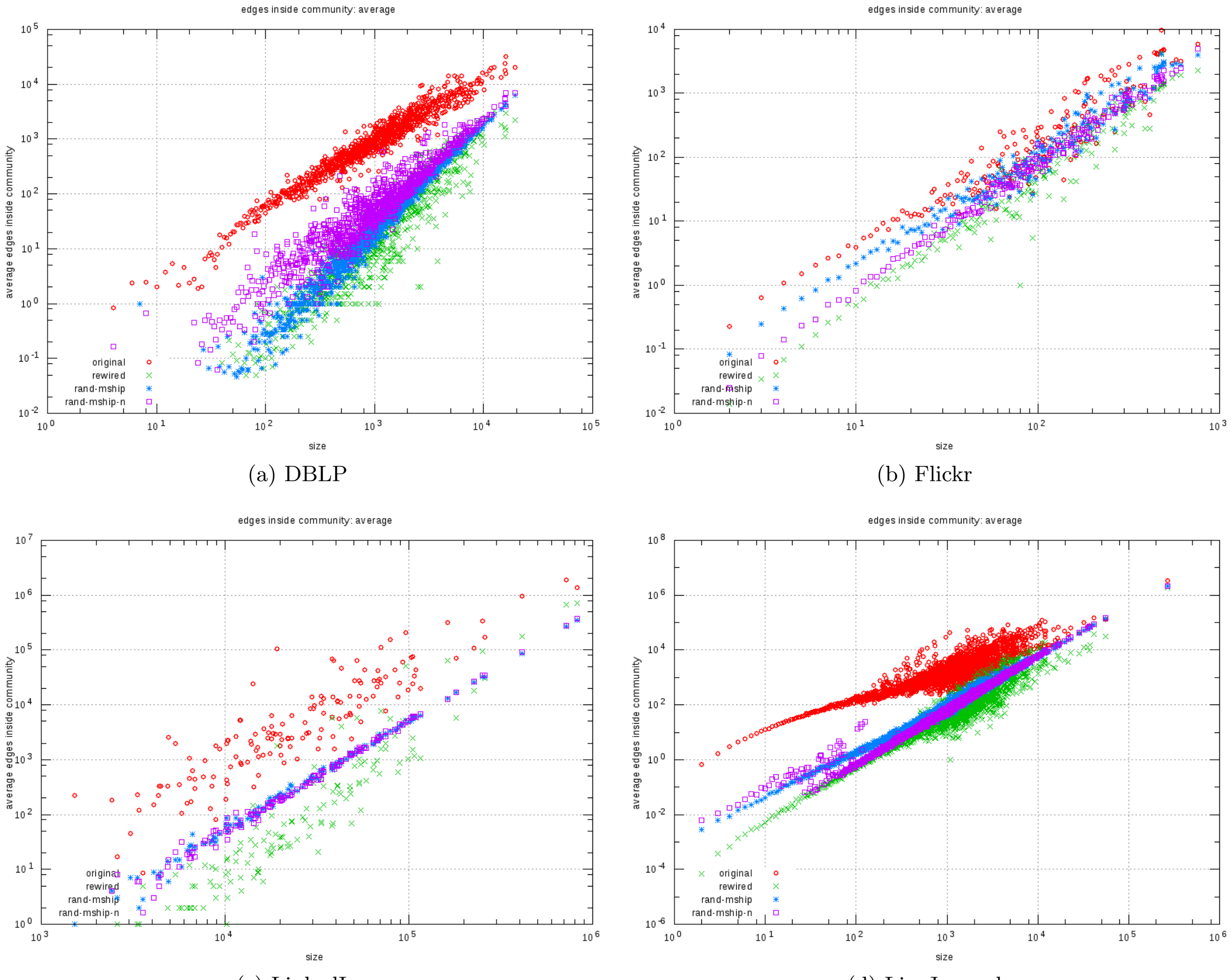
(a) DBLP

(b) Flickr

(c) LinkedIn

(d) LiveJournal

Figure 2: Plots showing the average number of edges inside the community versus community size. Each point in the plot represents the average number of edges over all communities of the particular size given on the x-axis.



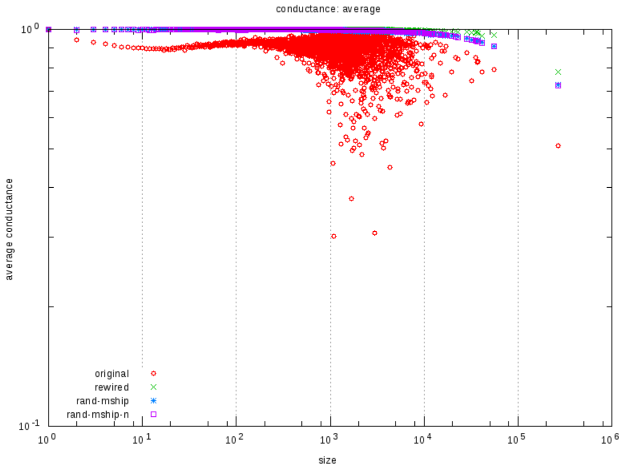
(a) DBLP

(b) Flickr

(c) LinkedIn

(d) LiveJournal

Figure 4: Plot of the average conductance versus community size in the LiveJournal dataset. Each point in the plot represents the average conductance over all communities of the particular size given on the x-axis.



edges may be between nodes that, with respect to the community in question, are not in the same community, but may be together in a different community altogether. Thus, we should not need to separate these nodes as they are, after all, in the same community. In other words, we need to take into account the fact that nodes may be part of multiple communities, and we should not count edges on the boundary between nodes that belong together in a different community. To this end, we have re-computed the measures with this intution in mind, only counting boundary edges if their endpoints do not share any community memberships. Preliminary results show that this is indeed an important factor, as seen from Figure 4, a plot of the average conductance versue community size on the LiveJournal dataset. As compared to the similar plot in Figure 1(d), we observe that now there is a separation between the real and random communities, especially at smaller community sizes up to about 1000. The corresponding plot for DBLP showed that all communities had zero conductance once multiple community memberships were taken into consideration, showing that perhaps it is easy to find communities in such a graph due to the artefacts of graph construction.

In the future, we will analyse more datasets to see if our observations hold in general. We will then use these observations to design a community detection algorithm, based not on preconceived ideas of what a community looks like, but on actual insight gained from empirical studies such as this one.

6. ACKNOWLEDGEMENTS

I would like to thank Professor Leskovec for his insights and guidance throughout the course of the project.

7. REFERENCES

[1] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *In Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160. ACM Press, 2000.

[2] S. Fortunato. Community detection in graphs. *arXiv:0906.0612v1 [physics.soc-ph]*, 2009.

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

[4] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

[5] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 497–504, New York, NY, USA, 2007. ACM.

[6] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[7] 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. *CoRR*, abs/0810.1355, 2008.

[8] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. May 2004.

[9] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.