Is Objective Function the Silver Bullet?

A Case Study of Community Detection Algorithms on Social Networks

Yang Yang*, Yizhou Sun†, Saurav Pandit*, Nitesh V. Chawla* and Jiawei Han†
*Department of Computer Science & Engineering
University of Notre Dame, Notre Dame, IN 46556
{yyang1, spandit, nchawla}@nd.edu
†Department of Computer Science
University of Illinois, Urbana and Champaign, IL 61801
{sun22, hanj}@uiuc.edu

***Abstract*—Community detection or cluster detection in networks is a well-studied, albeit hard, problem. Given the scale and complexity of modern day social networks, detecting "reasonable" communities is an even harder problem. Since the first use of k-means algorithm in 1960s, many community detection algorithms have been invented - most of which are developed with specific goals in mind and the idea of detecting "meaningful" communities varies widely from one algorithm to another. With the increasing number of community detection algorithms, there has been an advent of a number of evaluation measures and objective functions such as modularity and internal density. In this paper we divide methods of measurements in to two categories, according to whether they rely on ground-truth or not. Our work is aiming to answer whether these general used objective functions are well consistent with the real performance of community detection algorithms across a number of homogeneous and heterogeneous networks. Seven representative algorithms are compared under various performance metrics, and on various real world social networks.**

***Index Terms*—Social network; Community detection; Objective functions; Benchmark network; Measurements.**

I. Introduction

With the rapid proliferation of social networks, social media, blogs, and even cell phone communication networks, comes the opportunity of innovation in and application of community detection algorithms. However, the evaluation of community detection algorithms continues to be a hard problem. While a number of evaluation measures or objective functions have been proposed, each have their short-comings or assumptions. This becomes even more confounding when we also have heterogeneous information networks, as not only such networks are challenging for community detection algorithm, they are also challenging for evaluation. In this paper, we propose to compare community detection algorithms under various performance metrics, and on various "real world" social networks to explore whether current objective functions are well consistent with ground-truth of social network datasets. Another important purpose of our survey is to prove that different community detection algorithms have different performances on different social networks.

In order to conduct appropriate experiments, we divide the community detection algorithms in to two categories based on whether the social network is heterogeneous or homogeneous. Additionally, considering the heuristics or philosophy employed by community detection algorithms, some of the heuristics could be formalized in to objective functions, and then they can be optimized by maximizing or minimizing objective functions such as modularity [2, 6] and partition density [17]. However in some other algorithms, heuristics are extremely hard to be formalized in to objective functions, such as RankClus [1] and Edge Betweenness Clustering algorithm [5]. In this way we can divide each category in to two smaller categories, according to whether their heuristics can be formalized in to objective functions.

As for performance metrics, they can also be classified in to two categories according to whether their evaluations rely on ground-truth or not. Currently we are using the metrics listed in Table I, and in future more frequently used performance metrics will be included.

Besides performance metrics we discussed above, Andrea Lancichinetti et al. [16] proposed to use benchmark networks with built-in communities to evaluate the performance of community detection algorithms, this method is also involved in our comparisons.

Several datasets are chosen as our experimental subjects: Karate Club dataset [10], Mexican Political Power dataset [9], Sawmill dataset [13], Cities and Services dataset [21], and MIT Reality Mining dataset [11]. They are all small size social networks and are valuable at the startup stage of our survey, by using which we can have a more intuitive and clear view of social networks and community detection algorithms. In future work, we plan to increase the size of social networks to explore more details.

II. Related work

A. Community Detection Algorithms

The algorithms used in our survey are selected according to categories described in Section I. From Table II we can clearly see why these algorithms are selected to perform our experiments.

B. Social Network Datasets

The social networks datasets are listed in Table III.

TABLE I
PERFORMANCE METRICS OF COMMUNITY DETECTION ALGORITHMS

Performance Metrics	Based on Ground-Truth	Not Based on Ground-Truth			
	Rand Index	Internal Density	Conductance	Cut Ratio	Modularity
Comments	$(SS+DD)/(SS+SD+DS+DD)$ The first character of each variable states whether two nodes are from the same ground-truth class, and similarly the second character of each variable represents whether they are classified together by the algorithm. For example, SS is the number of pairs of nodes which are from the same ground-truth class and are also clustered together by the algorithm.	$2m_k/n_k(n_k-1)$ This is the internal density of links within the community C_k [8].	$1-l_k/(2m_k+l_k)$ This is the fraction of total edge number pointing outside the community [8].	$1-l_k/n_k(n_k-1)$ This is the fraction of all possible edges leaving the community structure [8].	$\sum_{k=1}^{K} mod_k/2M$ This is the quality of communities, where $mod_k = m_k - \frac{1}{2M}\sum D_i.D_j$, node i,j belong to same community k [8].

For all these metrics, high score indicates better quality

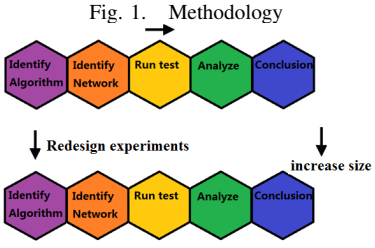
TABLE II
PERFORMANCE METRICS OF COMMUNITY DETECTION ALGORITHMS

Algorithm	RankClus [1]			LinkCommunity [17]			LineGraph [22]			K-means [18]		
	Formalization	Heter	Homo	Formalization	Heter	Homo	Formalization	Heter	Homo	Formalization	Heter	Homo
Properties	No	Yes	N/A	Yes	Yes	Yes	Depends	Yes	Yes	Yes	No	Yes
Algorithm	Walktrap [2]			SPICi [15]			Betweenness [5]			Comments		
	Formalization	Heter	Homo	Formalization	Heter	Homo	Formalization	Heter	Homo	See footnote		
Properties	Yes	N/A	Yes	Yes	N/A	Yes	No	Yes	Yes			

TABLE III
SOCIAL NETWORK DATASETS

Datasets	**Size**	**Ground-Truth Communities**	**Is Heterogeneous**
Karate Club	34 nodes	2	No
Mexican	35 nodes	2	No
Sawmill	36 nodes	3	No
Reality Mining	79 nodes	2	No
Cities&Services	101 nodes	4	Yes
Benchmark	45 nodes	3	No

Fig. 1. Methodology



C. Methodology

The methodology employed in our paper is something like "black box" approach by measuring algorithms under different objective functions, because whether these objective functions are reliable or not is unknown to us. By employing this methodology we can simplify our work of experiments and achieve more concise comparisons of these performance metrics of community detection algorithms. Currently only small sized social networks with ground-truth information are chosen, which allow us to perform the initial analyses and insights. With these analyses we can gradually increase the social networks size and see whether these insights still hold with the increment of social network size, and the type of social networks. Using this methodology we can redesign experiments with ease to explore more precise conclusions.

III. EXPERIMENTS

Our selected data sets include both heterogeneous network and homogeneous networks. Likewise, our clustering methods include algorithms designed for heterogeneous networks, homogeneous networks or both (Table II). Such that if an algorithm is designed for homogeneous network and we apply it on heterogeneous network, it may have unreasonable results. However, there is possibility that it still has high score of some objective function, in this way the biases between objective function and ground-truth information could be identified.

We apply seven selected community detection algorithms on the datasets listed in Table III. The communities detected are evaluated by performance metrics presented in Table I.

We will study these data according to two dimensions: algorithm dimension and objective function dimension. Algorithm dimension means we only study the related behaviors of one specified algorithm, while objective function dimension refers that we compare information related with specific objective function. Generally speaking the ground-truth based rand index is much more precise to differentiate the qualities of algorithms on social networks.

TABLE IV
EXPERIMENT RESULTS

Dataset	GT	RankClus						Walktrap					
		RI	ID	C	CR	M	Co	RI	ID	C	CR	M	Co
Karate	2	1	0.275	0.875	0.965	0.211	2	0.745	0.308	0.81	0.953	0.188	2
Mexican	2	0.489	0.168	0.434	0.778	0.003	2	0.536	0.197	1	1	0.018	1
Sawmill	3	0.530	0.048	0.138	0.877	0.003	3	0.560	0.307	0.845	0.981	0.178	2
Cities	4	0.668	0.988	0.168	0.018	0.004	4	0.348	0.266	1	1	0.006	1
Reality	2	0.575	1	0.987	0.988	0.099	2	0.561	1	0.968	0.969	0.100	2
Bench	3	0.718	0.208	0.625	0.937	0.107	3	1.0	0.31	0.874	0.981	0.289	3
Dataset	GT	K-means						LinkCommunity					
		RI	ID	C	CR	M	Co	RI	ID	C	CR	M	Co
Karate	2	0.941	0.168	0.503	0.897	0.057	2	0.743	0.499	0.468	0.907	0.284	8
Mexican	2	0.536	0.218	0.606	0.847	0.066	2	0.536	0.197	1	1	0.018	1
Sawmill	3	0.527	0.309	0.761	0.961	0.232	3	0.560	0.328	0.731	0.902	0.314	5
Cities	4	0.604	0.31	0.282	0.807	0.032	4	0.348	0.266	1	1	0.006	1
Reality	2	0.523	0.433	0.742	0.944	0.189	2	0.574	0.964	0.898	0.828	0.109	3
Bench	3	1	0.143	0.411	0.888	0.015	3	0.826	0.397	0.598	0.931	0.406	11
Dataset	GT	SPICi						Betweenness					
		RI	ID	C	CR	M	Co	RI	ID	C	CR	M	Co
Karate	2	0.586	0.729	0.524	0.898	0.136	5	0.913	0.21	0.63	0.933	0.199	3
Mexican	2	0.553	0.6	0.648	0.903	0.155	3	0.605	0.079	0.1	0.79	0.036	7
Sawmill	3	0.629	0.633	0.547	0.947	0.192	7	0.570	0.028	0.11	0.908	0.022	6
Cities	4	0.636	0.513	0.11	0.799	0.022	12	0.267	0	0	0.729	0.007	12
Reality	2	0.573	0.88	0.844	0.9	0.098	2	0.563	0	0.11	0.322	0.079	9
Bench	3	0.865	0.521	0.731	0.965	0.260	5	0.943	0.399	0.721	0.964	0.284	4

In these tables **GT** states the number of classes of ground-truth, **RI** is the rand index score, **ID** is the internal density, **C** is the conductance, **CR** is the cut ratio, **M** represents the modularity, and **Co** is the number of communities detected by corresponding algorithms. As for these metrics, ***higher score indicates higher quality***.

If we focus on the algorithm of RankClus (originally designed for heterogeneous networks) we can see that internal density, conductance, cut ratio and modularity to some extent can reveal algorithm's performance over different datasets. For example RankClus does not perform well on the Mexican Political dataset across the different performance measures — rand index score with other social networks, internal density, modularity and conductance also suggest that the detected communities are of poor quality. RankClus has the best performance on cities and services dataset; however the related conductance, cut ratio and modularity are do not reveal the high performance. Another example is, RankClus correctly clustered all nodes in the Karate Club dataset, and however the internal density does not have the best score when comparing with other algorithms. These bring forth the critical issues with evaluation.

Another interesting observation is that Walktrap algorithm and LinkCommunity algorithm have the worst performance on the same social networks (they cluster nodes of Mexican dataset and cities dataset in to one single community). And more interesting thing is that while they have the worst performance their conductance and cut ratio scores are perfect, which gives diametrically opposed measurements. A possible reason for this is that Walktrap and LinkCommunity are both designed to optimizing some objective functions. In contrast optimization of criterions does not always lead to real qualified communities. Additionally we can see that although RankClus is designed for heterogeneous networks, it also has surprisingly high score on specific homogeneous network, this is an interesting phenomenon we need to look deep in to at our next stage of work. The behaviors of community detection algorithms vary in different social networks.

When we concentrate on a single objective function, for instance, internal density, trivially we can find that SPICi algorithm has the best internal density on Karate dataset; however RankClus has the best performance on Karate dataset. Another example is, LinkCommunity has the best modularity on Sawmill dataset while SPICi has the best performance on Sawmill dataset. Among the data in Table IV there are a lot of such examples, based on current experiments results and observations, we can see that the correlation between the real quality and objective functions that are not based on ground-truth, is not strong.

Lancichinetti et. al [16] proposed to use generated benchmark networks to measure the performance of community detection algorithms. However, in Table IV we can see that these six algorithms listed above all have very high rand index scores, which does not hold for the other datasets. There are two possible reasons, the first one is this generated benchmark network is easy to be "mined", another one is the

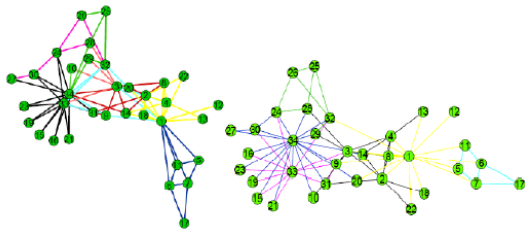


Fig. 2. LinkCommunity (Left) and Line graph (Right) clustering on the Karate Club dataset

generated benchmark networks are not good approximations of "real world" social networks. In order to know which reason contributes to this phenomenon, we need to conduct more experiments and try more benchmark networks generated under different configurations.

The quality of communities detected by algorithms is hard to evaluate, for example, in Fig. 2 we need to carefully analyze them to get the conclusion that line graph algorithm gives more reasonable partition. Performance metrics can largely help our evaluations but cannot completely define the quality.

IV. Conclusion and Future Work

Seven representative algorithms were compared under various performance metrics, and on various "real world" social networks. Based on our current observations of experiments results, we can conclude that performance metrics based on the ground-truth information are more reliable than objective functions that are not based on ground-truth, such as internal density and modularity. Benchmark networks are not yet confirmed to be good approximations of ground-truth method.

Our current work is based on small social networks, with observations we made in Section III, we will redesign next stage experiments to enhance our opinions or achieve more precise conclusions. For example, objective functions not based on ground-truth information are not so strong to accurately reveal the performance of algorithms on social networks we have studied. However, there is possibility that these objective functions could become reliable when the size of social network increases, this requires us to conduct more experiments on larger social networks. As for benchmark networks we need try more benchmark networks created under different properties configurations, either we can prove that benchmarks networks are not good approximations of measuring algorithms, or we can find that proper configured benchmark networks could facilitate our research on community detection algorithms comparisons.

In the future work more performance metrics are to be involved, and more algorithms and datasets will be selected to reinforce the robustness of our conclusions. With the gradual increment of dataset size the relation between social network volume and objective functions is estimated to be unfolded.

Acknowledgment

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, T. Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis," EDBT 2009, March 24-26, 2009, Saint Petersburg, Russia.

[2] P. Pons, M. Latapy, "Computing communities in large networks using random walks," Journal of Graph Algorithms and Applications, 2006.

[3] K. Steinhaeuser, N. V. Chawla, "Identifying and evaluating community structure in complex networks," Pattern Recognition Lett, 2009.

[4] K. Steinhaeuser and N. V. Chawla, "Is Modularity the Answer to Evaluating Community Structure in Networks?" International Conference on Network Science (NetSci), Norwich, UK.

[5] M. Girvan, M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences of the United States of America, Vol. 99, No. 12.

[6] Y.Sun and Y.Yu, "Survey on Clustering of Information Networks".

[7] S. Pandit, V. Kawadia, Y. Yang, N. V. Chawla, S. Sreenivasan, "Detecting Communities in Time-evolving Proximity Networks," submmitted to IEEE First International Workshop on Network Science, 2011.

[8] J. Leskovec, K. J. Lang, M. W. Mahoney, "Empirical Comparison of Algorithms for Network Community Detection," WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.

[9] J. Gil-Mendieta and S. Schmidt, "The political network in Mexico," in: Social Networks 18 (1996), 4: 355-381.

[10] W. W. Zachary, "An information flow model for conflict and fission in small groups," Journal of Anthropological Research 33, 452-473 (1977).

[11] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," 2005.

[12] J. Chen , O. R. Zaïane , R. Goebel, "Detecting Communities in Social Networks using Max-Min Modularity," International Conference on Data Mining (SDM 09).

[13] J. H. Michael and J. G. Massey, "Modeling the communication network in a sawmill," Forest Products Journal, 47 (1997), 25-30.

[14] W. de Nooy, A. Mrvar, and V. Batagelj, "Exploratory Social Network Analysis with Pajek," Cambridge: Cambridge University Press, 2004), Chapter 12.

[15] P. Jiang and M. Singh, "SPICi: a fast clustering algorithm for large biological networks," Bioinformatics (Oxford, England),Vol.26,No.8.(15 April 2010).

[16] A. Lancichinetti1, S. Fortunato, J. Kertsz, "Detecting the overlapping and hierarchical community structure in complex networks," New Journal of Physics 11, 2009.

[17] Y. Ahn, J. P. Bagrow, S. Lehmann, "Link communities reveal multiscale complexity in networks," arXiv:0903.3178v3 [physics.soc-ph], 2010

[18] I. Dhillon, Y. Guan, and B. Kulis, "A Fast Kernel-based Multilevel Algorithm for Graph Clustering," Proceedings of The 11th ACM SIGKDD, Chicago, IL, August 21-24, 2005.

[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions of pattern Analysis and Machine Intelligence," 22(8):999-905, 2000.

[20] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proceedings of the National Academy of Sciences of the United States of America, 101(9):2658-2663, 2004.

[21] World Cities and Global Firms dataset was created by P.J. Taylor and D.R.F. Walker as part of their project "World City Network: Data Matrix Construction and Analysis" and is based on primary data collected by J.V. Beaverstock, R.G. Smith and P.J. Taylor (ESRC project "The Geographical Scope of London as a World City" (R000222050)).

[22] T. S. Evans and R. Lambiotte, "Line graphs, link partitions, and overlapping communities," Phys. Rev. E, vol. 80, no. 1, p. 016105, Jul 2009.