The Impact of Quality Metrics on Communities Detected in Complex Networks

research.pomona.edu/complexnetworks

[Undergraduate Category]

Jennifer Nguyen (‘17), Christina Tong (‘17), Anastasia Voloshinov (‘17),
Tzu-Yi Chen (Faculty Advisor, PI)
Computer Science Department, Pomona College

1. INTRODUCTION

Networks (also known as graphs) can be used to represent real-world systems, where nodes represent entities of a system and edges represent interactions between the entities. Examples of these networks include modeling friendship between users on Facebook, membership of teams in a sports league, or neuron interactions within the human brain.

A community within a network is a set of members that are more connected to each other than to other members; communities might represent groups on Facebook or college athletic conferences in the NCAA's Division III. In many applications, we are interested in identifying communities within these networks based solely on the interactions observed.

Community quality metrics are measurements of how well communities are formed. Good metric values generally reflect networks where connections are denser within communities than between communities. The majority of commonly-used algorithms for detecting communities optimize a metric known as modularity [4][12], which compares the densities of the interactions between members in a community and between the communities themselves to a random graph with similar characteristics, such as vertex degrees. However, other quality metrics such as conductance, coverage, performance, and silhouette index [4] could also be used within those same algorithms. For this study, we examine the impact of replacing modularity with the other quality metrics in existing implementations of the Louvain [12] and Clauset-Newman-Moore (CNM) [1] community detection algorithms.

2. EXPERIMENTAL DESIGN

We implemented coverage, silhouette index, and performance in an existing implementation of the Louvain algorithm [6], and coverage in an existing implementation of the CNM algorithm [2]. We ran thirty-six networks from the Stanford Large Network Dataset Collection [3] and The University of Florida Sparse Matrix Collection [11] through these algorithm variations, computing the community networks and their associated metric values. To analyze the differences between the resulting clusterings, and, where applicable, the networks' ground truths, we are running community difference metrics, including normalized mutual information (NMI) [5], split-join distance [9], the Meila index [7], the Rand index [8], and the adjusted Rand index [10]. We also visualized the resulting communities using Cytoscape.

3. RESULTS

In our initial testing on several smaller weighted and unweighted networks with ground truths, Louvain-coverage, Louvain-performance, and CNM-coverage occasionally but inconsistently output better clusterings than the existing Louvain-modularity and CNM-modularity algorithms. That is, according to a majority of the aforementioned community difference metrics, their output clusterings were more similar to the ground truth. Though Louvain-silhouette index never did better than Louvain-modularity or CNM-modularity, it performed almost as well in select test cases.

4. FUTURE WORK

Going forward, we will implement the remaining metric conductance in the Louvain algorithm and implement conductance, performance, and silhouette index in the CNM algorithm. We will continue to run the Louvain and CNM variations on a larger test suite of both weighted and unweighted graphs, and analyze the accuracy and usefulness of the resulting communities using the community difference metrics.

5. REFERENCES

[1] A. Clauset, M.E.J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E* 70 (2004), 066111.

[2] A. Clauset, M.E.J. Newman, and C. Moore, "Structure and Strangeness," http://www.cs.unm.edu/~aaron/research/fastmodularity.htm.

[3] A. Krevland J. Leskovec, "SNAP Datasets: Stanford Large Network Dataset Collection," http://snap.stanford.edu/data.

[4] H. Almeida, D. Guedes, W. Meira and Mla .J. Zaki, "Is there a best quality metric for graph clusters?," *Proc. European conference on Machine learning and knowledge discovery in databases*, 2011.

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

[6] L. Waltman and N.J. Van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," http://www.ludowaltman.nl/slm/.

[7] M. Meilă. "Comparing clusterings by the variation of information". *Learning theory and kernel machines*, 173-187., Springer Berlin Heidelberg, 2003.

[8] M.W. Rand. "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical association*, 66(336), 846-850, 1971.

[9] S. Dongen. "Performance criteria for graph clustering and Markov cluster experiments". *Technical Report INS-R0012*, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, May 2000.

[10] S. Wagner, & D. Wagner. "Comparing clusterings: an overview". *Karlsruhe: Universität Karlsruhe, Fakultät für Informatik*, 2007.

[11] T.A. Davis and Y.Hu, "The University of Florida Sparse Matrix Collection," *ACM Transactions on Mathematical Software* 38 (2011): 1-25.

[12] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 8, 2008.