Implementation of Network Community Profile using Local Spectral algorithm and its application in Community Networking

Vaibhav VPrakash

Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India
vaibhavvp1990@gmail.com

Abstract

The problem that is addressed here and being investigated is to empirically review the paper entitled "Empirical comparison of algorithms for network community detection" - Jure Leskovec, Kevin J Lang and Michael W Mahoney"wherein we look at the characteristics and specific properties of various social networks used in the public and private domain. The objective of the investigation is to understand completely the network community detection using Local Spectral and Metis+MQI algorithms and to analyse how communities are created and ranked on specific metrics. Five communities have been compared using the same heuristics of the established functions in the entitled paper and an inference is drawn based on the graph generated by the same.

Keywords: *Community Networking, Community Detection, Social Networks, Network Community Profile, Local Spectral Algorithm.*

1. Introduction

Detecting clusters or communities in real-world graphs such as large social networks, web graphs and biological networks is a problem of continued dynamic practical interest that has received a great deal of attention. A "network community" is typically thought of as a group of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network. Here the distinguishing between "amongst" and "between" is critical [1].

Here, initially we have taken five networks which are going to be compared following which a small comparison has been made between LOCAL SPECTRAL and METIS+MQI algorithms. Further, a small note has been given in SNAP and the implementation of the algorithm for which specific graphs have been generated and inferred based upon the definitions and characteristics and properties given in the entitled paper.

2. The Theory of Network Community Detection

In the study of complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally. In the particular case of non-overlapping community

DOI : 10.5121/ijdps.2012.3607

finding, this implies that the network divides naturally into groups of nodes with dense connections internally and sparser connections between groups. But overlappingcommunities are also allowed. The more general definition is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same community(ies), and less likely to be connected if they do not share communities[8].

In particular the networks chosen to plot the Network Community Profile are

Soc-Epinions1 -> Social Network

Email-Enron ->Communication Network

Cit-HepPh -> Citation Network

Ca-AstroPh -> Collaboration Network

Ca-CondMat -> Collaboration Network

3. Brief Comparison of Local Spectral and Metis+MQI

LOCAL SPECTRAL	METIS + MQI
• Returns connected clusters	• Better at finding cuts with low conductance
• More compact	• Disconnected internally
• It finds clusters that have worse(higher) bounding cut conductance	• Is better at finding lower conductance even at larger scales
• It returns clusters with higher variance in the ratio of external to internal conductance	• It finds clusters that have better(lower) bounding cut conductance

4. Implementation and NCP plot comparison

Stanford **N**etwork **A**nalysis **P**ackage (**SNAP**)[12] is a general purpose network analysis and graph mining library that is easily scales to massive networks, is efficient and easily extendible. It naturally supports rich networks with complex data types associated with nodes and edges of the network. SNAP was developed by Jure Leskovec during his Phd studies at Carnegie Mellon and was built on top of a general purpose STL (Standard Template Library)-like library GLib that was developed at Jozef Stefan Institute.

The NCP plots were experimented on five networks datasets as given below

Soc-Epinions1 -> Social Network[11]

Email-Enron ->Communication Network[11]

Cit-HepPh -> Citation Network[11]

Ca-AstroPh -> Collaboration Network[11]

Ca-CondMat -> Collaboration Network[11]

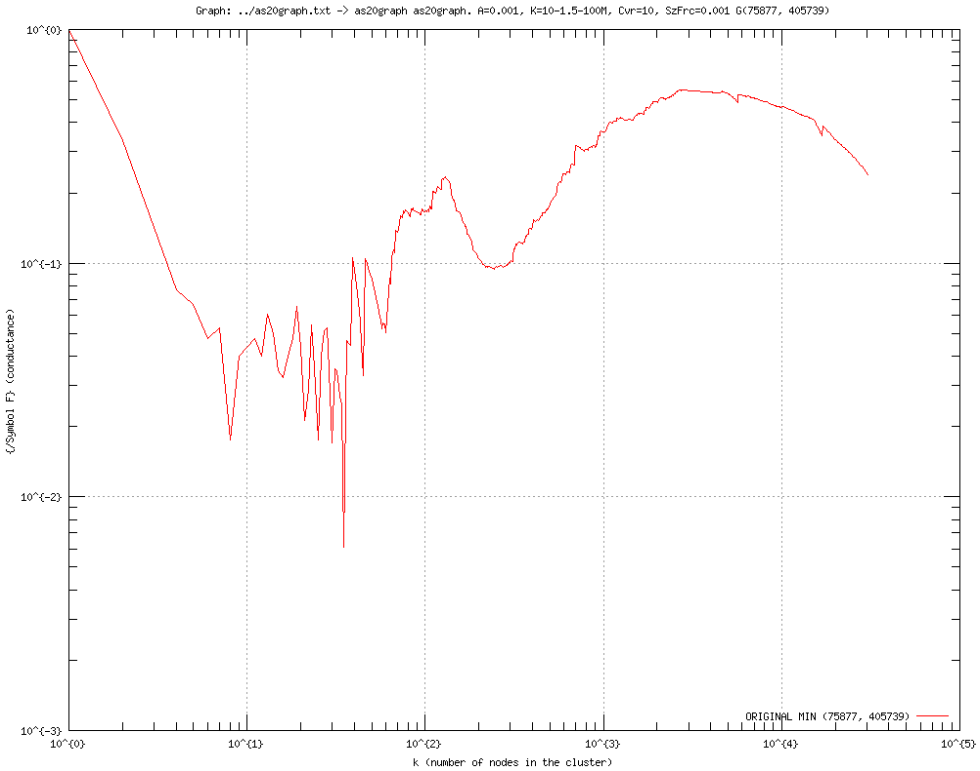

Fig (1) Soc-Epinions Social Network

Dataset statistics	
Nodes	75879
Edges	508837
Nodes in largest WCC	75877 (1.000)
Edges in largest WCC	508836 (1.000)
Nodes in largest SCC	32223 (0.425)
Edges in largest SCC	443506 (0.872)
Average clustering coefficient	0.2283
Number of triangles	1624481
Fraction of closed triangles	0.06568
Diameter (longest shortest path)	13
90-percentile effective diameter	5

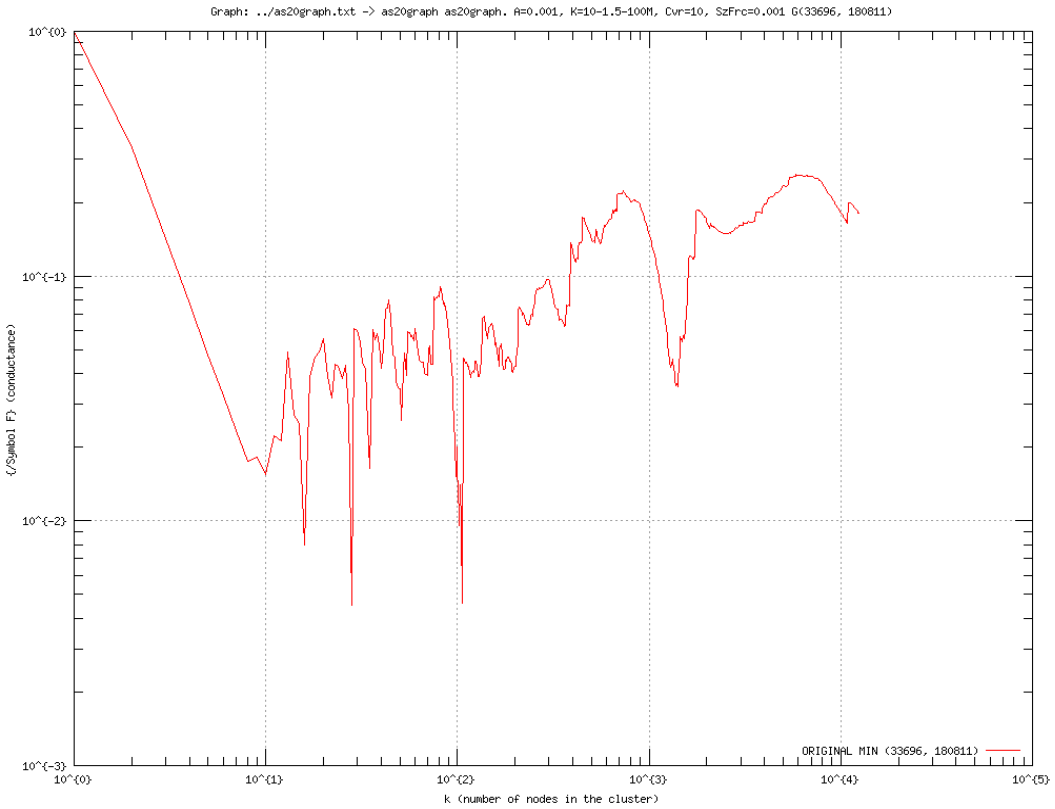

Fig (2) Enron Email network

Dataset statistics	
Nodes	36692
Edges	367662
Nodes in largest WCC	33696 (0.918)
Edges in largest WCC	361622 (0.984)
Nodes in largest SCC	33696 (0.918)
Edges in largest SCC	361622 (0.984)
Average clustering coefficient	0.4970
Number of triangles	727044
Fraction of closed triangles	0.08531
Diameter (longest shortest path)	12
90-percentile effective diameter	4.8

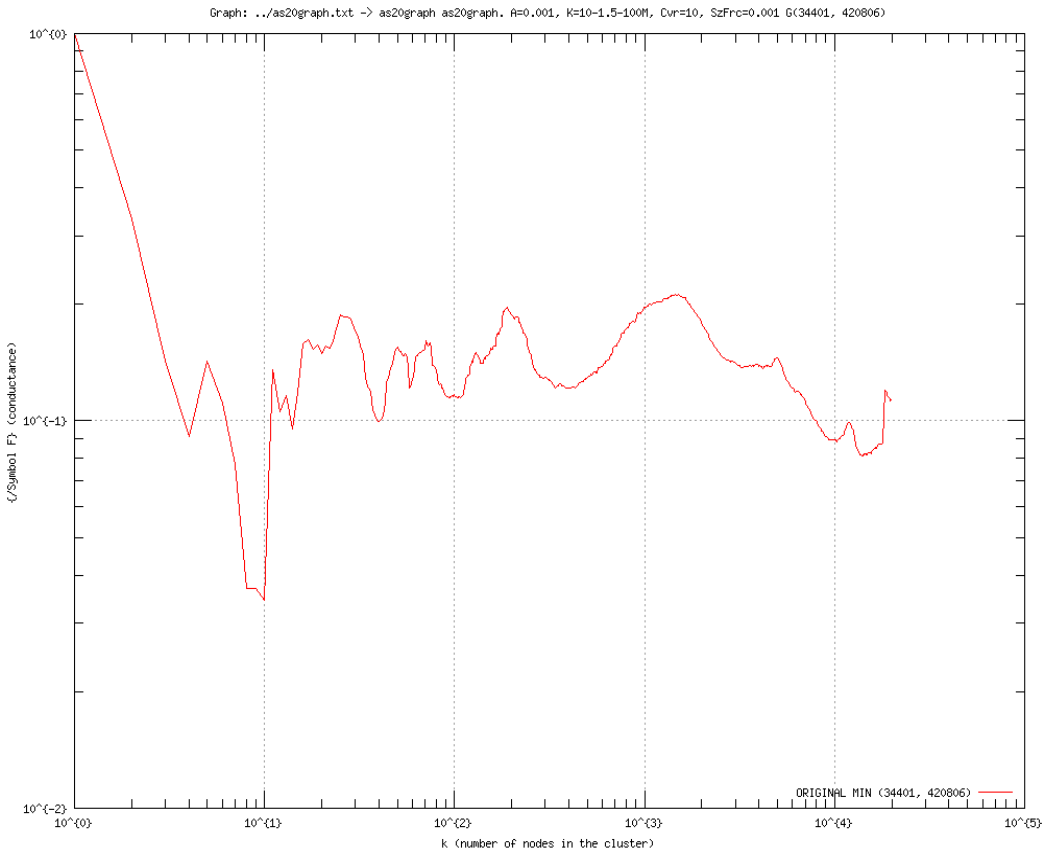

Fig (3) High energy physics citation network

Dataset statistics	
Nodes	34546
Edges	421578
Nodes in largest WCC	34401 (0.996)
Edges in largest WCC	421485 (1.000)
Nodes in largest SCC	12711 (0.368)
Edges in largest SCC	139981 (0.332)
Average clustering coefficient	0.2962
Number of triangles	1276868
Fraction of closed triangles	0.1457
Diameter (longest shortest path)	12
90-percentile effective diameter	5

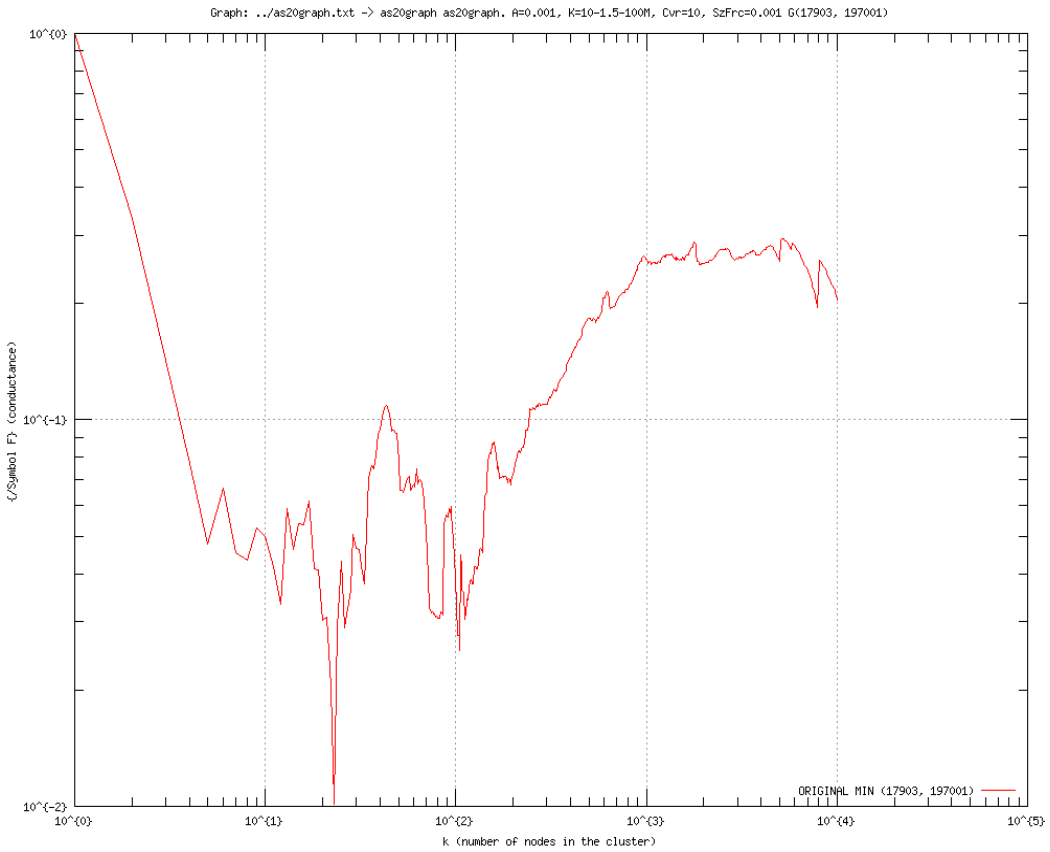

Fig (4) Astro physics collaboration network

Dataset statistics	
Nodes	18772
Edges	396160
Nodes in largest WCC	17903 (0.954)
Edges in largest WCC	394003 (0.995)
Nodes in largest SCC	17903 (0.954)
Edges in largest SCC	394003 (0.995)
Average clustering coefficient	0.6306
Number of triangles	1351441
Fraction of closed triangles	0.318
Diameter (longest shortest path)	14
90-percentile effective diameter	5.1

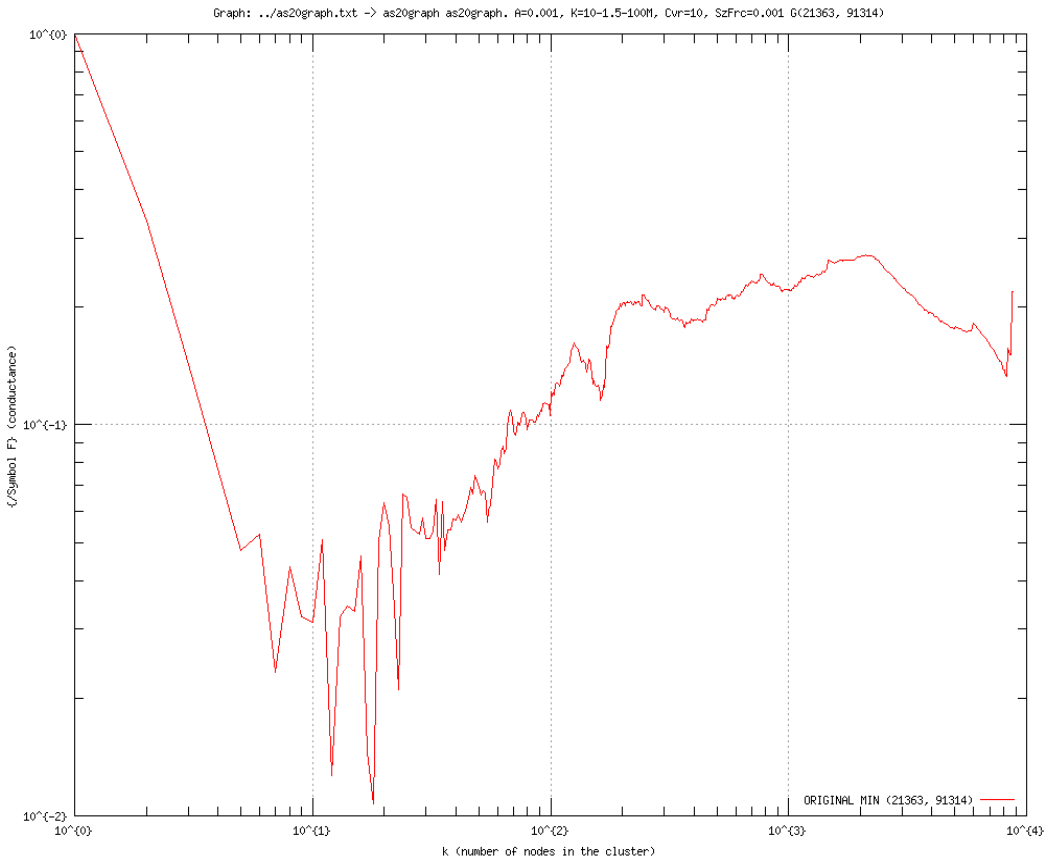

Fig (5) Condense matter collaboration network

Dataset statistics	
Nodes	23133
Edges	186936
Nodes in largest WCC	21363 (0.923)
Edges in largest WCC	182628 (0.977)
Nodes in largest SCC	21363 (0.923)
Edges in largest SCC	182628 (0.977)
Average clustering coefficient	0.6334
Number of triangles	173361
Fraction of closed triangles	0.2643
Diameter (longest shortest path)	15
90-percentile effective diameter	6.6

5. Discussion

From the graphs that are plotted namely Fig1-Fig5 one sees the data belonging to one cluster which is the Email-Enron Network, Astro-Physics Network, Condense matter collaboration Network and the Soc-Epinions Network whose characteristic of a falling conductance either plateuing with a small aberration and then the rise in conductance as the number of node increases. Normally after such a rise in all the above cases in the range of 10^3 and 10^4 finally the conductance plateaus. Therefore it is important that if we analyze each one of these networks separately.

Fig (1) SocEpinions Social Network

In the Epinions social network the conductance score reaches a minimum at 10^-2.5 at (10^1.5) nodes which forms the best community structure in this cluster and reaches a maximum value at 10^-0.3 at (10^3.3) which results in highest value of f(k) and loses its community structure.

Fig (2) Enron Email Network

In the Enron email network the conductance score reaches a minimum at two places viz 10^-2.5 at (10^1.4) nodes and 10^-2.5 at (10^2) nodes. At these two places the cluster is most community like. Then gradually it increases and f(k) becomes less community like.

Fig (4) and Fig (5)

In the Astro-Physics and Condense matter collaboration networks the conductance score reaches a minimum (10^-2) at 10^1.5 nodes which means that at this value the quality of the community is maximum (since the problem is NP hard only max and min values can be deduced) and it is most community like and as the number of nodes increases the network becomes less community like. The quality of the community goes to a low at (10^3) nodes at 10^-0.6.

However in the case of High Energy Citation Network the pattern is somewhat different. After the initial fall of the conductance there is a very steep rise and then the plataeuing starts at 10^1 nodes. One can generalize that all the five examples attempted to stimulate the comparison of the algorithm in different networks. Therefore the overall pattern is one that of falling conductance, reaching a minimum and then a rise of conductance reaching a plateau pattern. This is amazingly unique when we look at the heterogeneity of the network which stimulates the logistics of the network.

6. Inference

Hence we infer from the discussion that given a cluster of the above size, the best community like structure results from 10^1 nodes to 10^2 nodes as the conductance score is minimum for the above values.

Also at 10^3 nodes in the cluster the conductance score is the maximum which means that the community loses its structure as the nodes increase from 10^3 onwards as gradually the conductance score starts increasing.

7. Acknowledgement

I sincerely would like to thank Prof. Y. Narahari, Chairman, Department of Computer Science and Automation, Indian Institute of Science, Bengaluru for his guidance and direction which helped me to analyse and understand Social Network architectures, evolution and progression of different types of nodes in large real world graph networks.

8. References

[1]Jure Leskovec, Kevin J lang and Michael W Mahoney,(2010) "Empirical comparison of algorithms for network community detection", Proceedings of the 19th International Conference on the World Wide Web, P – 631 – 640.

[2] Reid Anderson, Fan chung and Kevin Lang(2006) "Local graph partitioning using page rank vectors", Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, P – 48 – 13 self.

[3] Bruce Hendrickson & Robert Leland(1995) "A multilevel algorithm for partitioning graphs", Supercomputing '95 Proceedings of the 1995 ACM/IEEE conference on Supercomputing (CDROM), Article No. 28.

[4] G.Gallo, M.Grigoriadis(1989) " A fast parametric maximum flow algorithm and application" - SIAM journal, Vol. 18, No. 1, pp. 30-55.

[5] G.Karypis and V.Kumar(1998) "A fast and high quality multilevel scheme for partitioning irregular graphs"- SIAM journal, Volume 20 Issue 1, Aug. 1998
Pages 359 – 392

[6] S. Arora, S. Rao, and U. Vazirani(2004) Expander flows, geometric embeddings and graph partitioning. In STOC '04: Proceedings of the 36th annual ACM Symposium on Theory of Computing, pages 222–231

[7] A. Clauset(2005). Finding local community structure in networks.Physical Review E, 72:026132.

[8] Definition of Community Detection from Wikipediahttp://en.wikipedia.org/wiki/Community_structure

[9]J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney(2008). Statistical properties of community structure in large social and information networks. In WWW '08: Proceedings of the 17th International Conference on World Wide Web, pages 695–704.

[10] Supporting website http://snap.stanford.edu/snap/index.html

[11] Supporting website http://snap.stanford.edu/data/index.html

[12] Supporting website http://snap.stanford.edu/