VIRAL MARKETING: INFLUENTIAL NODES

A Project Report

Presented to

Dr. Katerina Potika

Department of Computer Science

San Jose State University

In Partial Fulfilment

Of the Requirement for the Class

CS286C

By

Ujjawal Garg

May 2018

TABLE OF CONTENTS

Introduction

Social networks enable propagation of ideas and information through interactions between individuals. One of the most important concept is to identify a set of nodes that can propagate these ideas to the largest range on the network, and also do so in the least amount of time possible. This is achieved by ranking the nodes based on some metric. Generally, this metric is some sort of centrality measure, that defines how *central* the node is in the network. Three most common types of centrality measures are:

1. **Degree Centrality**: Rank the nodes based on their degree
2. **Closeness Centrality:** Rank the nodes based on their distances to other nodes
3. **Betweenness Centrality:** Rank the nodes based on how many shortest paths go through them.

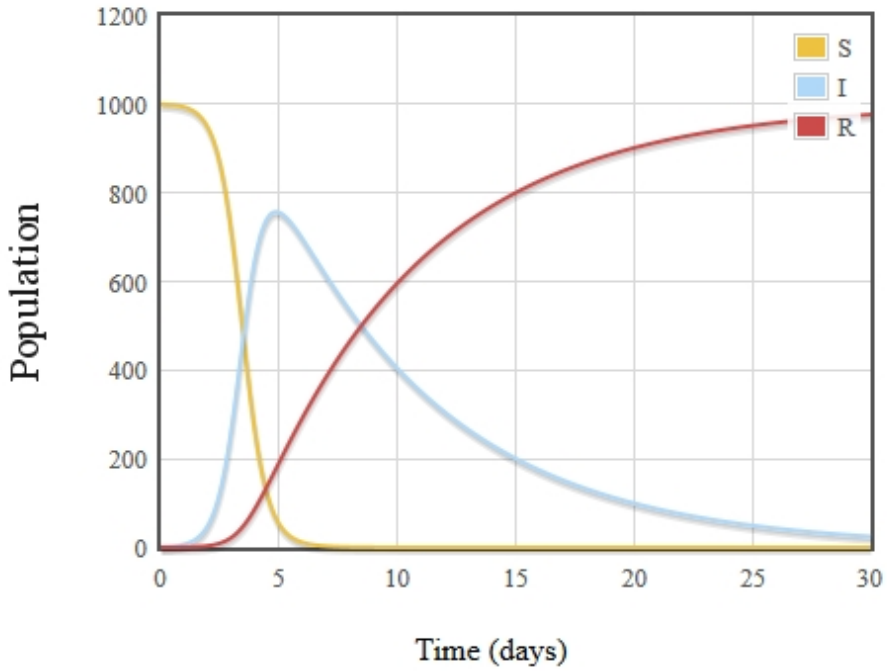
For this project, we follow the research done by Chen et. al [1]. They came up with a new centrality metric, which they called **Local Centrality**. The proposed Local centrality measure acts as a trade-off between this low-relevance and high computation cost.
In order to evaluate these methods, we need some model to determine the influence. In the paper, authors have used SIR (Susceptible-Infected-Recovered) model [2] for this purpose. I used the same model to evaluate the performance of these metrics. Section I describes the SIR model and how it can be used for evaluation. Section II introduces the **Local Centrality** measure introduced by [1]. Section III contains the results of the analysis and conclusion of this report.

I. SIR Model

In this model, each node can be in any of three states in its lifetime:

1. **Susceptible**: Not yet Infected
2. **Infected:** Will infect a randomly chosen neighbour node at each time step
3. **Recovered:** Immune to disease and will not take part in spreading infection

Initially all nodes are in Susceptible state. We select the node that we believe to be most influential and set its state to infected. Now, after each time step, for every infected node, a randomly chosen neighbour node will get infected. Infected nodes recover with probability *1/k, i.e.,* each node will infect an average of *k* nodes before becoming immune. Here *k* is the average degree of the network. Let F(t) denote the total number of infected and recovered nodes at time t. This value will increase monotonically and eventually become stable, once every node has been recovered. Fig 1. shows the graph denoting the relation between count of nodes in each state at different time-steps for 1000 nodes.



II. Local Centrality

Degree centrality is very simple and easiest to compute, but is of little relevance as it considers only very limited information. Consider the network show in Fig. 1.

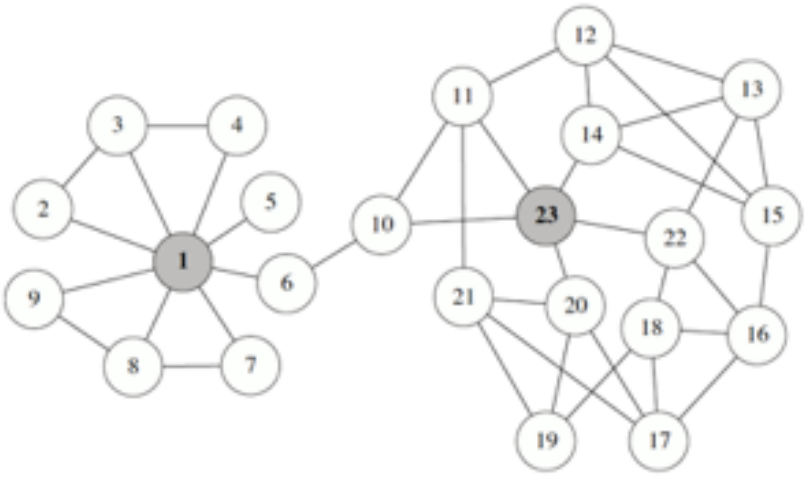


Fig. 1. An example network consisted of 23 nodes and 40 edges. Although node 23 has lower degree than node 1, its influence may be even higher.

Although node 1 has highest degree centrality, the disease, if it origins at node 1, may not spread the fastest or the most broadly since all neighbours of this node have very low degree. Node 23 on the other hand has faster spreading and larger influence.

Betweenness and Closeness centrality provide much better ranking results as they consider global overview of the network. However, they have very high computation cost, $O(n^3)$ and $O(n^2\langle k\rangle)$ respectively.

The local centrality CL(v) of node v is defined as:

$$Q(u) = \sum_{w \in \Gamma(u)} N(w),$$

$$C_L(v) = \sum_{u \in \Gamma(v)} Q(u),$$

Here Γ (u) is the set of the nearest neighbours of node u and N(w) is the number of the nearest and the next nearest neighbours of node w. Since to calculate N(w) requires traversing node w′s neighbourhood within two steps, the computational complexity of local centrality is $O(n\langle k\rangle^2)$ which grows linearly with the size of a sparse network. Fig 2, 3 and 4 shows the n-values, q-values and l-values of the nodes in a sample network.

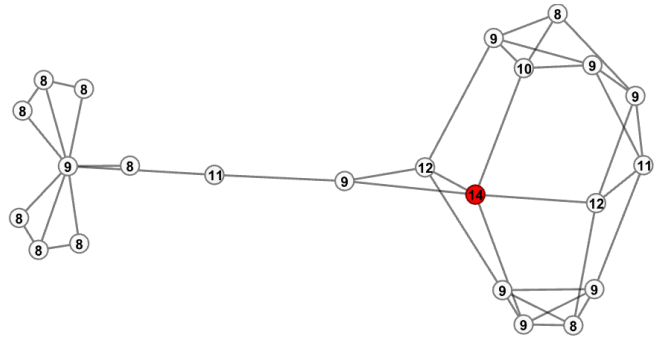


Fig 2. N-values of each node

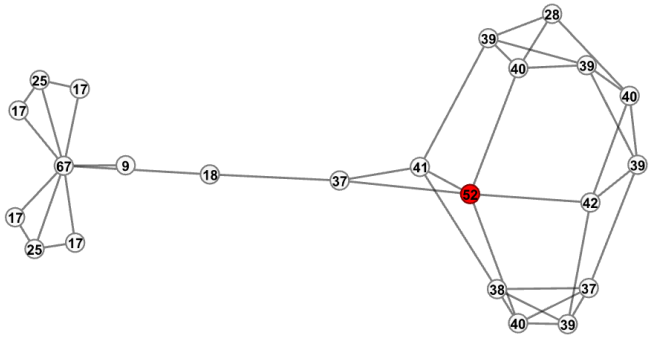


Fig 3. N-values of each node

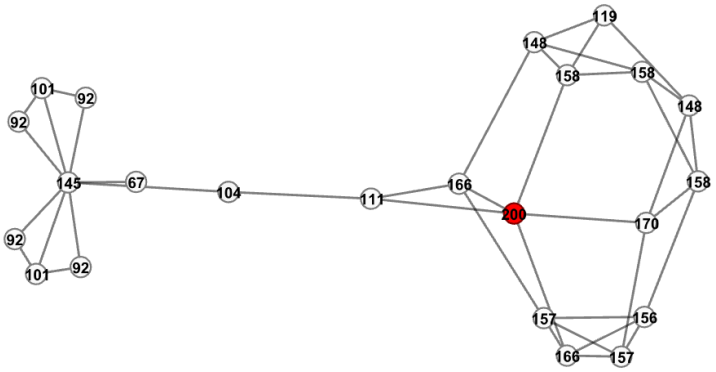


Fig 4. Final Centrality values of each node

III. Experiments

To analyse the performance of this new method, following datasets were used:

1. Netscience [3]: The network of co-authorships between scientists who are themselves publishing on the topic of networks. There are 379 nodes and 914 edges in this graph.
2. Email [4]: the network of e-mail interchanges between members of the University Rovira i Virgili (Tarragona)

Using the python-igraph and matplotlib library, graph was plotted between F(t) vs Centralities. Here the value of t is set to 10, same as in the paper. Thus the y-axis of the graph denotes the total number of nodes that have ever been infected if the corresponding node with centrality value at x-axis is selected as the seed node to spread the infection or idea. Clearly, for a perfect metric the graph would be monotonically increasing graph with values at y-axis (the influence) increasing with values at x-axis (the centrality metric). Fig 5 & Fig 6 show the results for Email and Netscience dataset respectively.

IV. Results

It is clear from these graphs that both closeness and local centralities outperform degree and betweenness metrics. There is no co-relation between the influence and betweenness or degree centrality. While it is difficult to say which is better among closeness and local centralities, the much lower computation cost of local centrality clearly makes it the winner.

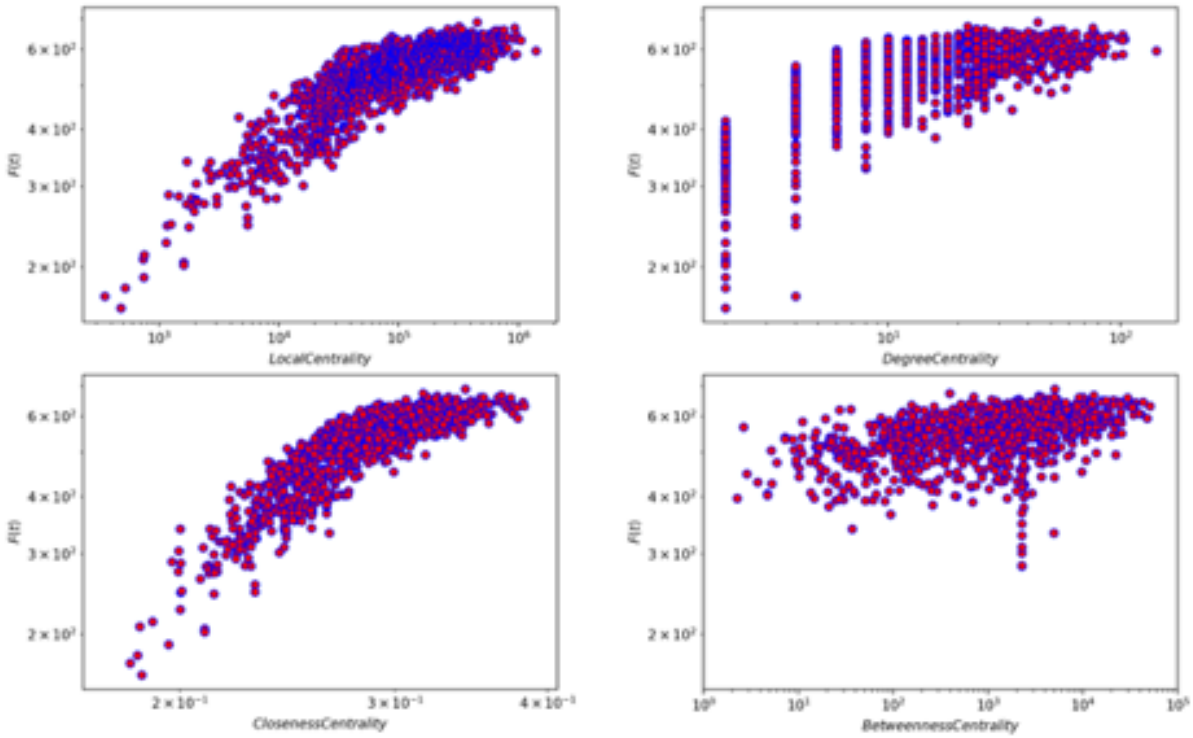


Fig 5 (Email Dataset)

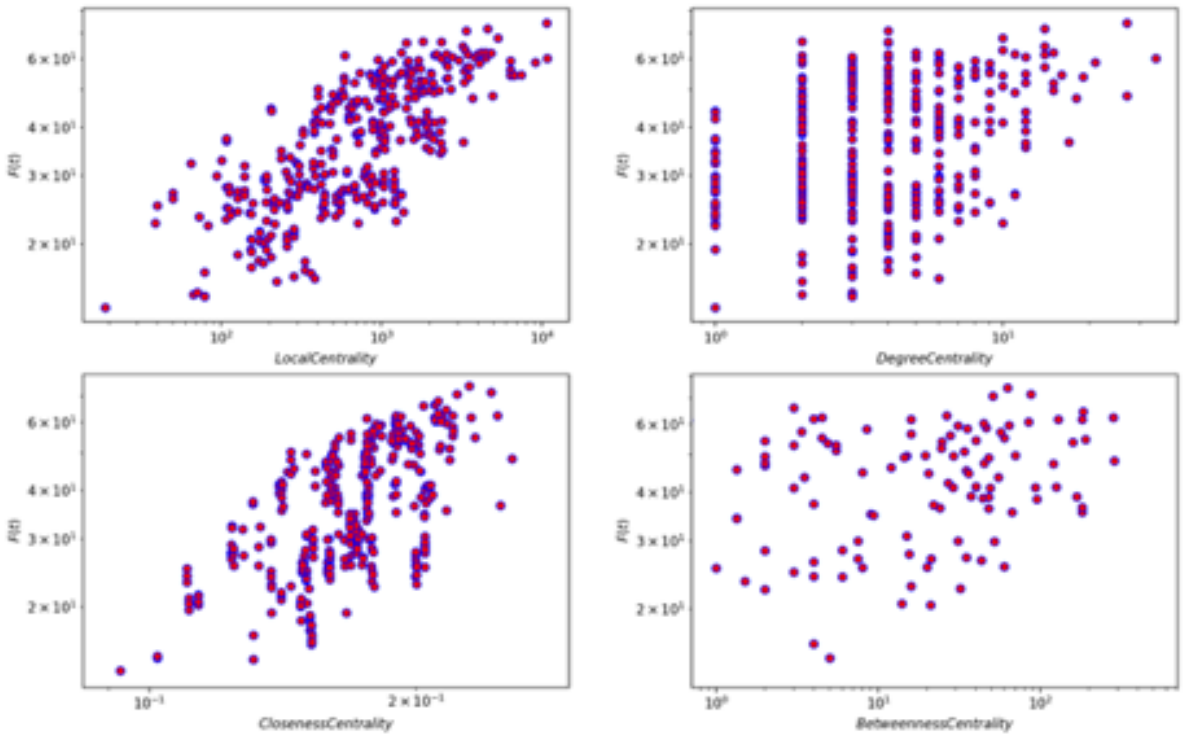


Fig 6 (Netscience Dataset)

References

[1] Ming-Sheng Shang Yi-Cheng Zhang Tao Zhou Duanbing Chen, Linyuan L. Identifying influential nodes in complex networks. 2012.

[2] R.M. Anderson, R.M. May, B. Anderson, Infectious Diseases of Humans: Dynamics and Control, Oxford University Press, USA, 1992.

[3] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices,
Preprint physics/0605087 (2006) Link: http://nrvis.com/download/data/ca/ca-netscience.zip

[4] R. Guimer., L. Danon, A. D.az-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (2003) 065103. Link: http://deim.urv.cat/~aarenas/data/xarxes/email.zip