

A case of The Good, The Bad and The Ugly: a note on determining valuable originators in preferential attachment networks

Abhishek Nagaraj¹

Aseem Sood²

February 22, 2010

Abstract

Using computer simulations of preferential attachment networks we find support for the hypothesis that 'well connected' nodes (as defined by standard parameters of network centrality) are ideal candidates to be chosen as originators in a social network. We do this by successfully clustering candidate originator nodes into three distinct groups based on their corresponding diffusion characteristic and then commenting on their centrality parameters.

¹ Indian Institute of Management Calcutta. (abhishekn2010@email.iimcal.ac.in)

² Indian Institute of Management Calcutta. (aseems2010@email.iimcal.ac.in)

We are thankful to Prof. Arijit Sen of the Indian Institute of Management Calcutta under whose guidance this term paper was written.

1. INTRODUCTION

What causes an innovation or an idea to spread in a social network is not only a fundamentally interesting problem, but it is also one with substantial sociological considerations. Research in this area can be used to support theories about why certain ideas ‘catch on’, why certain technologies become popular or why a certain disease reaches epidemic proportions.

Specifically we ask, how does the originator and her position in a social network influence the nature of diffusion? For example, how can computer companies potentially engineer widespread software adoption by finding the right people to give free copies to? How can locating and quarantining the right people arrest the spread of virulent diseases? These are a few questions that motivate this article.

In this study, we use a simulation-based approach to generate our dataset. After having repeatedly tracked the diffusion of ‘innovations’ in algorithmically generated networks we find support for the hypothesis that the more ‘central’ a node is, the better a candidate it is for origination. More specifically we find that ‘good’ originators have favourable centrality scores when measured using standard parameters of node centrality, namely ‘Eigen Vector Centrality’, ‘Average Reciprocal Distance’ (ARD) and ‘K-Step Reach’ ($k=3$ in our particular case).

2. STUDY DESIGN

The data for our study comes from a NETLOGO simulation³ we built to analyse diffusion. A screenshot of the simulation area is shown in Fig. 1.

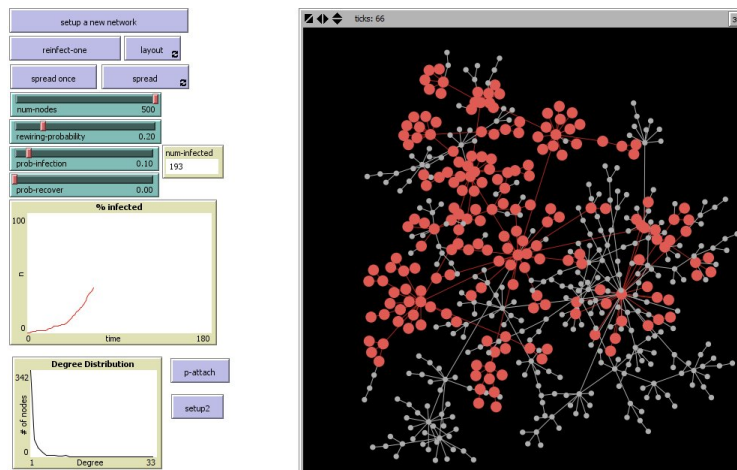


Fig. 1

³ The simulation builds upon work by Adamic and Bakshy (SmallWorldDiffusionSIS, 2008) and Wilensky, U. (NetLogo Preferential Attachment model, 2005).

The network, as represented in the rightmost area of the simulation, is composed of nodes, which represent entities, and links, which represent relationships that exist between nodes. We do not attach any significance to the nature of this relationship and posit simply that all links represent a homogenous relationship that exists among any two nodes which share a link in common. The network itself consists of 500 nodes and was constructed algorithmically using a preferential attachment scheme (Barabási 2000) which is described as follows.

The Preferential Attachment scheme supposes that when a network grows, newcomers to the network prefer expending resources on establishing relationships with already 'well connected' nodes. It uses the notion of 'degree' (i.e., the number of neighbours a node has) to define connectedness in this case. Wilensky (2005) describes the process of the network construction as follows:

The model starts with two nodes connected by an edge.

At each step, a new node is added. A new node picks an existing node to connect to randomly, but with some bias. More specifically, a node's chance of being selected is directly proportional to the number of connections it already has, or its "degree." This is the mechanism which is called "preferential attachment."

Once the network has been constructed using this scheme we randomly generate an origin node for our 'infection'. The dots which are in red in Fig. 1 are the ones which have already contracted the infection. Once an origin has been chosen, the simulation runs as follows: For each infected node, the probability that each of its uninfected neighbours contracts the infection is 10%. Each such iteration counts as a 'tick' – the unit of time we use for our measurement. The simulation therefore, starts with one red dot till all the dots have been infected red.

Given a particular network and given an origin node we record the percentage number of nodes infected at each point in time. Fig. 2 is a sample graph that is generated from one such diffusion profile.

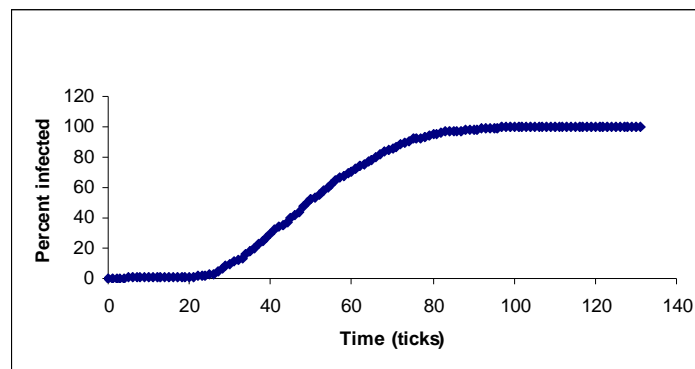


Fig. 2

We repeated this exercise for 60 randomly chosen origin nodes across three such large preferential attachment networks.

Next we use UCINET (Borgatti et al. 2002), a social network analysis tool, to calculate centrality parameters for each of our networks and the originator nodes. Combining data generated from our NETLOGO simulations and from UCINET analysis we build the dataset of 60 nodes; an extract of which is shown as Table 1.

NodeID	Degree	3Step	ARD	Eigenvector	per25	per50	per75	per90	per95
1003	0.002	0.416	0.291	0.118	33	44	56	70	82
1020	0.032	0.244	0.268	0.023	32	43	54	72	83
1075	0.012	0.28	0.264	0.028	20	30	45	60	70
1080	0.01	0.282	0.263	0.027	34	49	70	86	101
1084	0.002	0.416	0.291	0.118	19	31	46	67	76
1127	0.004	0.082	0.202	0.004	29	42	59	73	79

Table 1

In Table 1. the variables **per25**, **per50**, **per75**, **per90** and **per95** represent the number of ticks it took for the network to be infected 25%, 50%, 75%, 90% and 95% respectively when the origin of the infection was the node indicated by **NodeID**.

We use three standard centrality measures (as calculated by UCINET): Average Relative Distance, 3Step Reach and Eigenvector centrality. These measures, well developed within the field of Network Analysis, use differing approaches to measuring how ‘central’ a node is in a given network. We include their definitions and supporting notes in Appendix III.

Once the dataset of 60 nodes has been built, as shown in Table 1, we use the variables **per25**, **per50**, **per75**, **per90** and **per95** to perform k-means clustering. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Depending on the diffusion characteristic (as defined by 5 time-points along the curve shown in Fig. 2) we cluster our dataset into three distinctive groups. The means of the diffusion times in each of these three clusters is shown in Table 2.

Cluster	Distance clust-1	Distance clust-2	Distance clust-3	per25	per50	per75	per90	per95
bad	21.971	70.413	74.789	51.348	65.652	83.087	101.087	111.087
good	70.175	16.141	142.961	24.889	35.519	50.333	67.481	79.407
ugly	76.232	143.867	28.000	70.727	94.909	117.909	137.727	151.545

Table 2

Having clustered all of our nodes into the three clusters based on their realized diffusion characteristic, it now remains to be shown that the centrality measures of nodes in these three clusters define the behaviour that they demonstrate.

3. RESULTS

First, we establish the nature of the three clusters we have obtained in Table 2. To do this we plot an estimate of the diffusion curve shown by the ‘average node’ in each cluster as defined by the average values of **per25**, **per50**, **per75**, **per90** and **per95** of each node in the cluster. This graph is as shown in Fig 3.

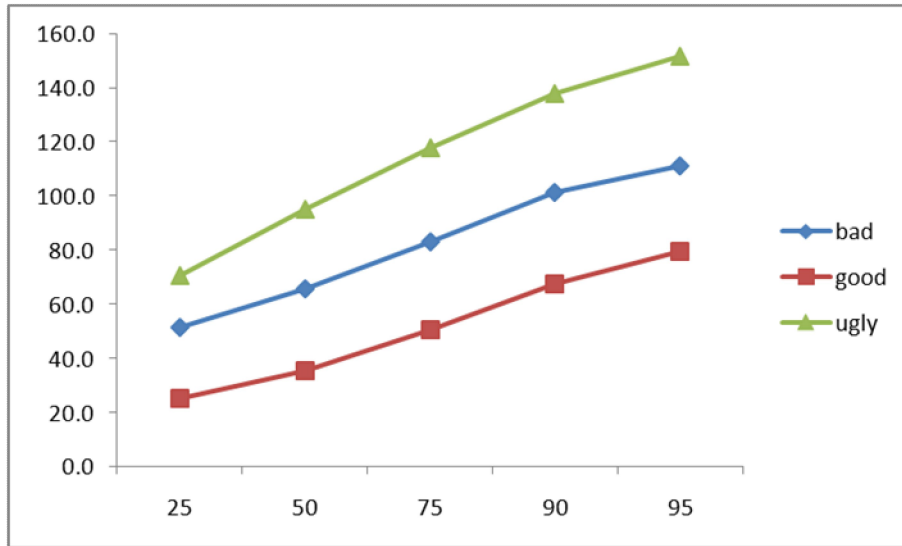


Fig. 3

As can be seen in Fig.3 each cluster represents an average diffusion characteristic that can be compared strictly with any other cluster. The cluster, which we name ‘good’, on an average has nodes which were the origins of faster diffusions than those in the cluster ‘bad’, the nodes in which, in turn, did better than those in the cluster ‘ugly’. We list diffusion graphs for nodes in each of these three clusters in Appendix II.

Having established this notion of ranking of three clusters we are in a position to look at centrality measures which define these networks. A summary of these centrality measures can be found in Table 3.

Cluster	3Step	ARD	EigenVector
good	0.176	0.202	0.031
bad	0.100	0.155	0.014
ugly	0.043	0.052	0.001

Table 3

Table 3, our main results table, shows that all the three centrality measures **3Step**, **ARD** and **Eigenvector** are highest for the cluster 'good' followed by the values for the cluster 'bad' and then the cluster 'ugly'. This finding is the support for the result we are looking for: nodes which demonstrate high centrality measures are indeed good originators if rapid diffusion is desired.

3. FINAL THOUGHTS

Barabási and Albert (1999) report that preferential attachment networks can potentially be a good model for diverse sets of real-world networks, including the world-wide web and the network of scientific citations. We therefore feel confident in saying that our results could be used to investigate what could turn out be important phenomena in a wide variety of real-world networks.

Moreover, our study can be the starting point for a variety of allied investigations into the characteristics of originator nodes. In particular, we could look at differentiating among centrality measures and establish conditions under which a few measures are more reliable than others. Another line of investigation could be to measure the impact of centrality on the speed of diffusion. A preliminary logistic regression that we perform is included in Appendix I and seems to indicate there could be significant differences in the effect of each of the three measures in promoting rapid diffusion.

Finally we conclude by restating our belief in the importance of analysing originator node centrality in understanding the nature of the diffusion process. Concentrating on where an idea begins could lead to rich dividends in terms of how fast it spreads.

4. REFERENCES

Albert-László Barabási (2000) *Linked: The New Science of Networks*, Perseus Publishing, Cambridge, Massachusetts, pages 79-92.

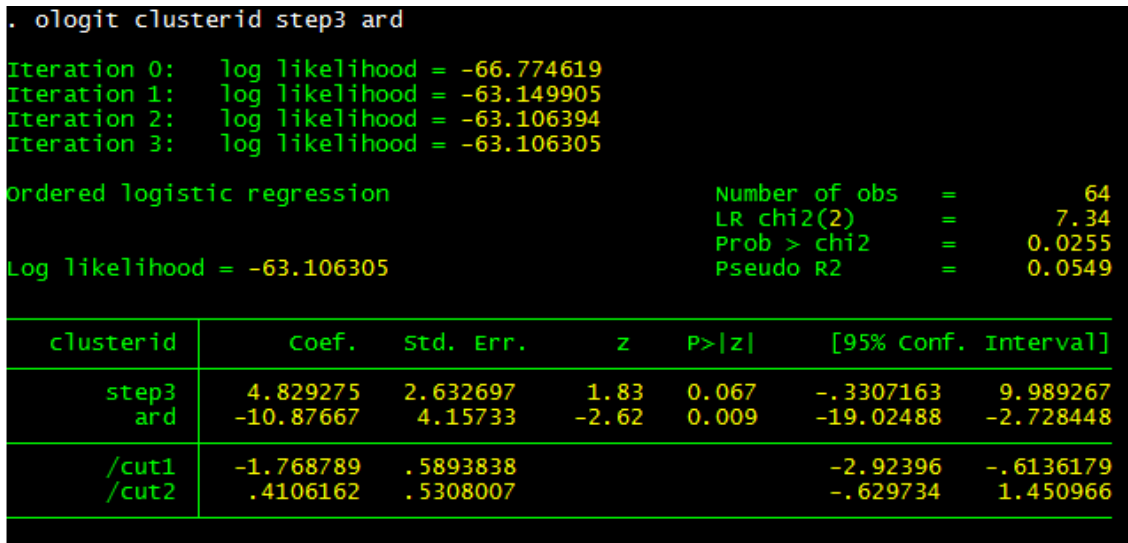
Albert-Laszlo Barabási and Reka Albert (1999), Emergence of scaling in random networks, *Science* volume 286, pp. 509

Borgatti, S.P., Everett, M.G. and Freeman, L.C. (2002) *Ucinet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Wilensky, U. (1999). *NetLogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

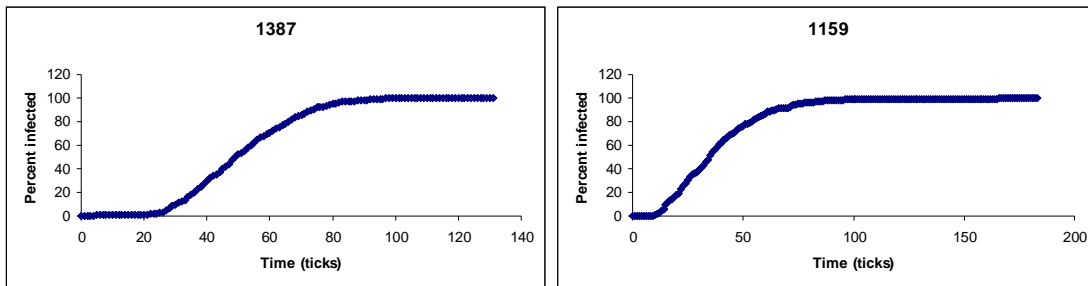
Wilensky, U. (2005). *NetLogo Preferential Attachment model*
<http://ccl.northwestern.edu/netlogo/models/PreferentialAttachment>

Appendix I – LOGIT REGRESSION

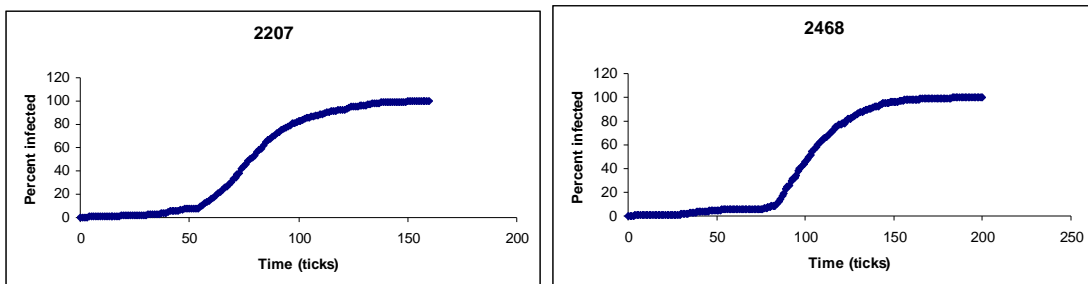


Appendix II – SAMPLE DIFFUSION CHARACTERISTICS ACROSS CLUSTERS

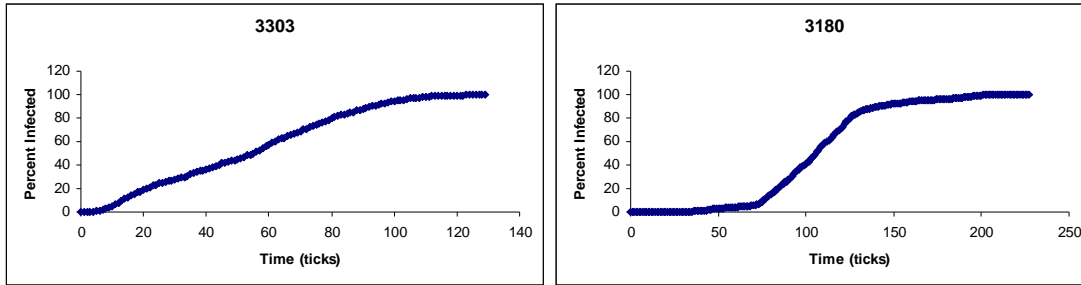
Cluster 1 – Bad



Cluster 2 – Good



Cluster 3 – Ugly



Appendix III – CENTRALITY MEASURES AND THEIR DEFINITIONS

Eigenvector Centrality

Given a network, we can define an adjacency matrix. For each edge between node i and j , $A_{ij}=1$, else $A_{ij}=0$. Now, the centrality of vertex i (denoted c_i) is given by $c_i = aSA_{ij}c_j$ where a is a parameter to ensure non-trivial solutions for centrality. These centralities, thus obtained are the elements of the corresponding eigenvector. The normalized eigenvector centrality is the scaled eigenvector centrality divided by the maximum difference possible expressed as a percentage.

Therefore the greater the value of the Eigenvector Centrality for a node, the greater is its 'connectedness'.

Average Reciprocal Distance

The far-ness of a node is measured as the sum of the length to all other nodes. The reciprocal of this summation is closeness. Now shift the reciprocal to before the summation sign so as to remove the effect of infinite distance between two nodes. The distance so obtained is normalized with respect to the maximum value of closeness, so as to obtain the ARD in percentage terms.

The Average Reciprocal Distance (ARD) is thus a measure of closeness centrality i.e. how short the path lengths between nodes are. The greater the value of ARD, the greater is the 'connectedness' of the node.

$$c_i = S1/d_j$$

where d_j is the number of edges traversed to reach node 'i' from node 'j' and c_i is the closeness parameter.

$$\text{Normalized ARD} = c_i / C_{\max} - C_{\min}$$

3-Step Reach

3 Step Reach is measures the 'connectedness' of a node within a circle of influence. For a node, the 3 Step Reach is defined as, the number of nodes reachable within 3 Steps. Therefore the greater the 3 Step Reach of a node, the greater is its circle of influence.