**Integration Recommendations in Socio- Economic Networks**

Hamid Zargariasl

Mediterranea university of Reggio Calabria, Italy

*hamid.zargariasl@unirc.it*

**Abstract-** Optimizing the process of aggregating different networks in social networks is discussed in recent years and different methods have been suggested. One of the most important approaches is selecting initial nodes to make bridges between two networks. In this paper, we want to evaluate performance of the information diffusion in the result network whose connecting nodes are chosen according to their node level centrality in primary networks. In two networks of friendship, a small group of nodes were selected according to their different centralities for linking to their counter parts in other network. After this algorithmic selection and linking, for any selected method, performance of the information diffusion in aggregated network is calculated. Results show that selecting nodes based on some kinds of centralities gives better outputs but they need to have more inter-node communication.

**Keywords**: Social Networks, Information Diffusion, Cross-Community, Network Centrality

**Introduction**

Informing all or a big fraction of nodes of a network in minimum time and inter-node communication are main goals of intervention on information diffusion process. It can be used in a lot of studies such as advertisement, media, education and so on. Changing topology, selecting initial informed nodes and their number(like [1]) are from challenges in this area.

For evaluating and optimizing the network in spreading information, the fraction of informed nodes is most important parameter in comparing different algorithms (like [2] ,[3] ). But, a main parameter that reflects the congestion of information traffic between nodes is ignored. It means that we evaluate only the informed nodes and we do not think about the costs that we have to pay for it. In this point of view, two group of parameters must be studied: Desired ones like fraction of informed nodes and unwanted numbers like number of communication between nodes. Not only in human science, but also because of interdisciplinary nature of social network approaches and using its rules in other fields like engineering [4] it is important and worth investigating. For example, though the number of informed nodes is important for a network of smart objects, but imposed inter node communication brings about more communications between nodes and consequently more energy consumption. This problem is appeared also in connecting different networks or cross-community approaches. It is desired to predict the performance of the information diffusion in aggregated networks.

For connecting two networks there are different choices. Though making one connection between any nodes of any network to any node of other network is enough, but there are other problems like vulnerability. It is also possible to connect all nodes of one network to all nodes of other network that has other practical difficulties. If we consider the networks as in the Figure 1, the essential questions are emerged as number of bridges and choosing the nodes.

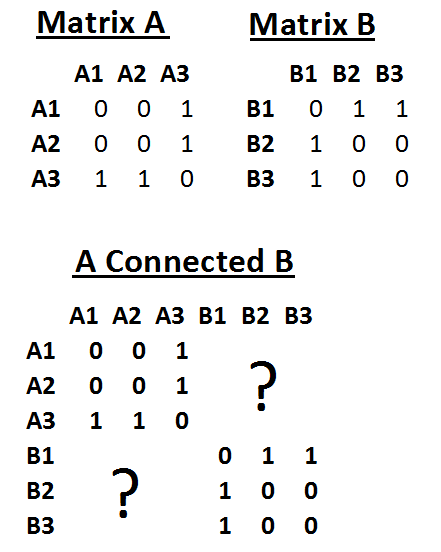


Figure 1 Problems in connecting two different networks.

We aim to find the best nodes in non connected networks for connecting them so that the information flow in the result network (aggregated) be optimized. By finding the more effective nodes, consequently the number of needed links are reduced and, in turn, costs are reduces.

Some studies (like [5]) have offered methods for maximizing the performance of the information diffusion between communities. Here, after showing benefits of these methods, we criticize them by defining new parameters.

After a brief review of used parameters for both information diffusion and network centrality, used data sets and method will be explained. The rest of paper deals with results and conclusions.

**Scaling the Fluency of Information**

We focus on parameters that are used for measuring the process of information spreading in network. For explaining, the considered metrics for evaluating the performance, a network like Figure 1 is supposed that is made by four nodes and four links. In this network the distribution of degrees are:

Number of nodes with degree one: 1 ; Number of nodes with degree two: 2 ; Number of nodes with degree three: 1

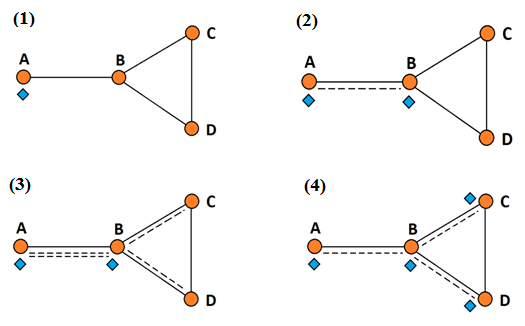


Figure 2 A sample graph with four nodes during process of information diffusion.

We suppose that node A has a portion of the information that must be diffused in the network. Node A must know that node B has the considered data or not, then one communication between both nodes must be executed and if B has not data, it must receive the data. In this paper, we suppose that data can be transmitted without any problem and with probability equal to one. The model for information diffusion is considered as the “Cascade model”.

By sending data from A to B, one step is completed and the results are:

1st step: 1- A wants to know that other node (B) has the same data or not; 2- Sending data to B. We can derive the percentage of informed nodes by dividing informed nodes to all node number 2/4\*100=50% as in Figure 2 (2) .

2nd step: 1-Nodes B must communicate with its neighbors to proceed the information diffusion steps. Here will be two communications between nodes: from B to C and from B to D as in Figure 1 (3). We suppose that nodes A and B are aware of each other (having intended data); 2- Sending data to the neighbors those have not the information. Two transmissions of information will be needed: from B to C and from B to D.

It means that until this step, there were: two steps passed and three information transfers have been completed and 100% of network is received the initial data as in Figure 1 (4).

All mentioned calculated values are demonstrated with the following variables:

**Steps:** Passed steps from first informed node to inform all nodes

**Effective Transmissions:** All done transmissions from informed nodes to unaware nodes. This parameter is shown in tables as *Trsmt*.

**Rate[i]**: Rate of informed nodes in *i-th* step. For example, Rate[2]=0.85 shows that after completing the second step, 85 percent of nodes are informed.

Obviously, for evaluating efficiency of spreading the information, all defined parameters must be calculated from any of nodes and the average values of same parameters from all nodes must be shown as well.

**Network Centrality**

Node level centrality metrics show the importance of any node in network and can be divided to different groups [6]. It must be cited that we considered network as a symmetric network, because the context and direction of information may be in both directions. The following social network centralities are considered and calculated by UCINET [7].

**Degree**: Degree centrality of nodes in social graph

**Eigenvector:** The first eigenvector

**Bonacich Power**: Bonacich's power based centrality of nodes

**2 Step Reach**: Two Step Closeness centrality of nodes

**Between**: Betweenness centrality of nodes

**ARD**: Average Reciprocal Distance of nodes

**Data Set**

Two real networks are considered in this study:

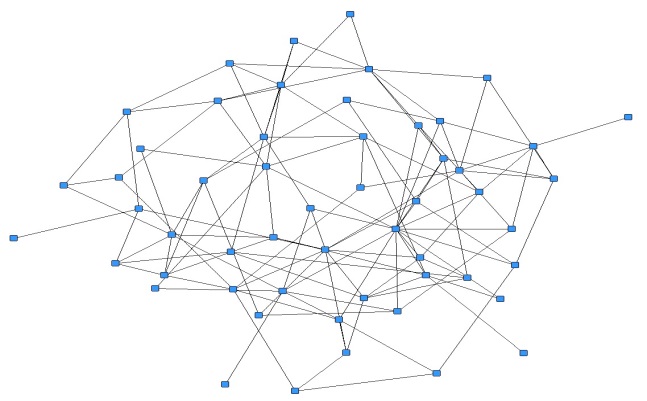
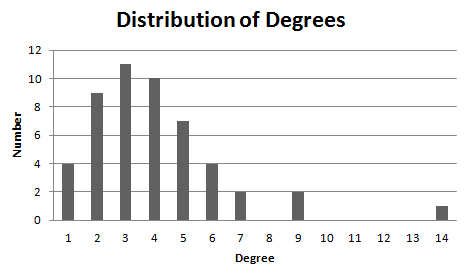
1. A friendship network by remote communication ( via telephone):

In this approach, centrality of 54 undergraduate students in 2008 (from March 17 2008 until April 20 2008) was collected from their network of remote communication.

1. A friendship co-presence network:

The contact network (Co-Presence) of 35 students in 2014 (from March 20 2014 until March 30 2014).

Graphs of considered nodes and their degree distributions are visualized by NETDRAW [8] and shown in Figure 3.

  ****

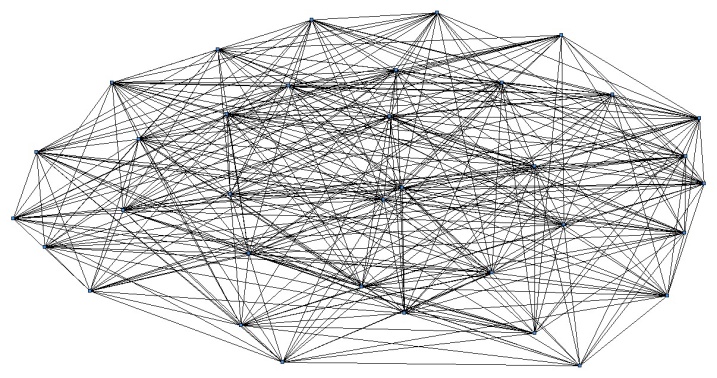
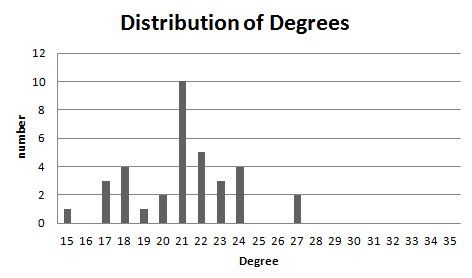
 

Figure 3The social graph of 54 (up) and 35 (down) students and their distribution of degrees

It shows that the second network is denser and minimum contacts are fifteen times in sampling period. For example there are 10 nodes out of 35 nodes that have contacts with 22 persons.

**Research design**

We prepared a software module by C++ for finding information diffusion parameters as well as some general characteristics of networks like density, links and so on.

Calculated parameters were categorized in three groups:

1. **Information Diffusion:**

Steps, Transmits (Number of inter-node data transfer), Rat[1], , Rat[2] , Rat[3] , Rat[4]

1. **Network general parameters:**

Links, Max degree, Density, distribution of degrees

1. **Social Network parameters:**

Degree, Eigenvector, Bonacich Power, 2 Step Reach, Betweenness, ARD

After calculating all mentioned parameters for both networks, four nodes (4 over 54 and 4 over 35 ) from any network were selected by different criteria. Selecting the nodes was done by centrality scores so that in any case four nodes that had higher numbers were chosen. The second network was very dense and therefore centrality of 2-step were all equal for all nodes, therefore it ignored. Also four nodes were selected randomly by random numbers. In any case two groups of nodes any including four nodes were connected to other nodes in other network. This operation made 16 links between two initial networks and made a new network comprising all nodes (Figure 4).

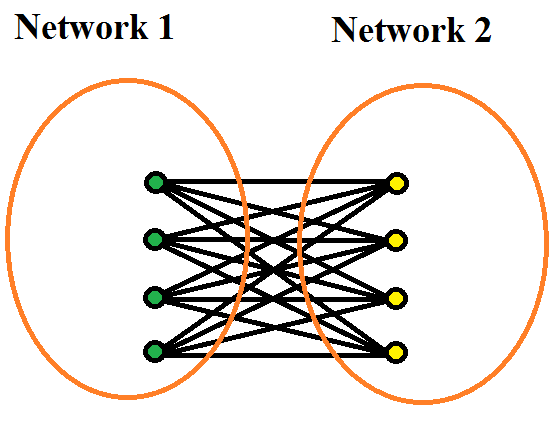


Figure 4 Bridging two separate networks

**Results**

Two networks were connected by different nodes calculated by their centrality in their network. Figure 5 shows both networks before and after connecting. In this Figure connecting nodes have higher scores for Eigen vector centrality in their network.

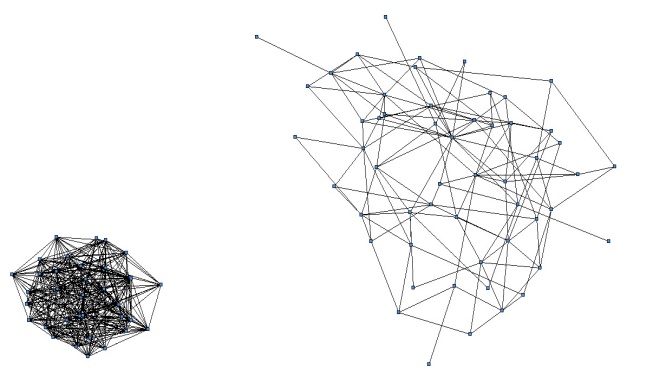
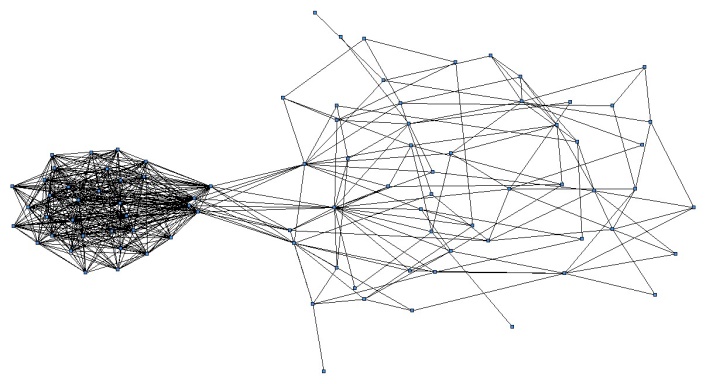
** **

Figure Network before connecting (left) and aggregated network (right) - Nodes with higher Eigen values are used

The specifications and behavior of initial networks are demonstrated separately in Table 1.

Table 1 specifications of primary networks

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Node** | **MaxDeg** | **Dens** | **Steps** | **Trnsmt** | **Rate[1]** | **Rate[2]** | **Rate[3]** | **Rate[4]** |
| **Matrix A** | 54 | 15 | 0.08875 | 3.11 | 79.28 | 0.1 | 0.47 | 0.96 | 1 |
| **Matrix B** | 35 | 26 | 0.59 | 2 | 182.9 | 0.6 | 1 | 1 | 1 |

Using UCINET [] software we collected five node level different centralities for any of them. It must be mentioned that all network is considered as a 1-mode network where any link in graph represents at least one event of co presence between two nodes. The graphs are also supposed as dichotomous and symmetric graphs due to mutual probability of information transfer in any contact.

The “2 step” centrality is ignored, because all nodes of the second network (35 nodes)are saturating all network in two steps. It was the reason that all calculated values for “2 Step” centrality and Rate[2] parameters were equal to one and were equal to two for all “Steps” parameter accordingly.

Table 2 demonstrates the results of different methods of connecting two networks.

Table 2 Aggregated network by different methods of connecting

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Node** | **MaxDeg** | **Dens** | **Steps** | **Trnsmt** | **Rate[1]** | **Rate[2]** | **Rate[3]** | **Rate[4]** |
| **ARD** | 89 | 31 | 0.12615 | 2.775 | 204.303 | 0.13572 | 0.60548 | 0.934 | 1 |
| **Between** | 89 | 31 | 0.12615 | 2.787 | 199.157 | 0.13572 | 0.60737 | 0.925 | 1 |
| **Bonacich** | 89 | 31 | 0.12615 | 2.753 | 200.225 | 0.13572 | 0.61002 | 0.937 | 1 |
| **Eigen** | 89 | 31 | 0.12615 | 2.787 | 198.360 | 0.13572 | 0.60737 | 0.925 | 1 |
| **Deg** | 89 | 31 | 0.12615 | 2.730 | 202.348 | 0.13572 | 0.61179 | 0.939 | 1 |
| **Random** | 89 | 27 | 0.12615 | 2.843 | 165.685 | 0.13572 | 0.60952 | 0.925 | 0.998 |

The results show that selecting nodes by centralities has advantages than random selecting. The average required steps for reaching the data to all the network from any node are smaller But, for other parameters like Rate[1], Rate[2], it is not satisfying.

The “Trnsmt” parameter that stands for real required numbers for the information transfer is main reason for our criticism. It shows that despite gaining in other parameters of the information diffusion, there is needed more communications between nodes of the network.

These results show that tradeoff between parameters is important to be done because of opposite effects of two groups of predictors. It means that number of inter-node information transfer must be considered in these studies.

**Conclusion**

Our work emphasizes previous works for selecting more central nodes as bridges between two networks, but we showed that it makes other side effect and degrades performance. We implemented this idea in two small networks, but for getting more accurate results big data sets must be studied.

**Reference**

[1] David Kempe, Jon Kleinberg, Éva Tardos, KDD '03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 137-146

[2] Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. N. Engl. J. Med. 357, 370–379, 2007.

[3] Anastasia Katholisch et al, ON THE ROLE OF CENTRALITY IN INFORMATION DIFFUSION IN SOCIAL NETWORKS, Proceedings of the 21st European Conference on Information Systems,Utrecht University, Utrecht, Netherlands, 2013.

[4 ] Jon Kleinberg.The convergence of social and technological networks. Commun. ACM 51(11):66-72 ,2008.

[5] V´aclav Bel´ak, Samantha Lam, and Conor Hayes, Towards Maximising Cross-Community information Diffusion

[6] Freeman, L.C. "Centrality in Networks: I. Conceptual Clarification." Social Networks, 1:215-239, 1979.

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

[8] Borgatti, S.P. NetDraw: Graph Visualization Software. Harvard: Analytic Technologies, 2002.