Synthesized Influence Assessment System Based on Multivariant Network Model

February 10, 2014

Abstract

To determine the influence of academic researchers, papers and movie actors etc, we model the influence and impact between the object to a scale-free network. Inspired by the PageRank Algorithm, HITS Algorithm, we believed that the influence and impact could regarded as the stream in the network

Contents

1 Introduction

There are thousands of researchers who collaborate directly or indirectly with Erdös. In this paper, we propose a network model analyzing the complex relation between them and extend this model to analyze more data. The key problem we are going to solve is how to determine the most important node in a network which has largest influence and impact on whole network.

2 The Co-author Network of Erdös and Its Properties

To discover the properties of co-author network of Erdös, we constructed a network $G(V, E)$ from the given dataset containing 511 Erdos1 authors, where V is a set of vertices including Erdos1 authors and their collaborators, and E is a set of undirected edges indicating the collaboration between authors. Let $A \leftrightarrow B$ represent A collaborates with B, set of analyzing target P and network G can be strictly defined as follow:

$$P = \{n \mid n \leftrightarrow Erdos\} \cup \{n \mid n \leftrightarrow m, \exists m \leftrightarrow Erdos\}$$
$$V = \{n \mid n \in P\}$$
$$E = \{(s, t) \mid s \leftrightarrow t, s \in P, t \in P\}$$
$$G = (V, E)$$

Based on the definition of the model, we take two steps towards the solution of this problem:

1) Extract data and construct network model,

2) Analyze network model and clarify its properties.

2.1 Data Extraction and Model Construction

Data were extracted from the given *Erdos1.txt* file and formed into a network description in adherence to definition of the model using a simple python script combined with NetworkX library[3]. Gephi[2], a graph visualization and manipulation software, analyzes the data retrived by script and visualizes the network. See Figure 1 for the graph of Erdos1 network, layout of which is composited by Fruchterman Reingold algorithm[1]. Those do not directly collaborate with Erdös are omitted from this graph.

2.2 Network Properties Analysis

Assume Erdös node is stripped from the network, and neither the year nor times of Erdos1 authors collaborating with Erdös is taken into consideration. We analyzed

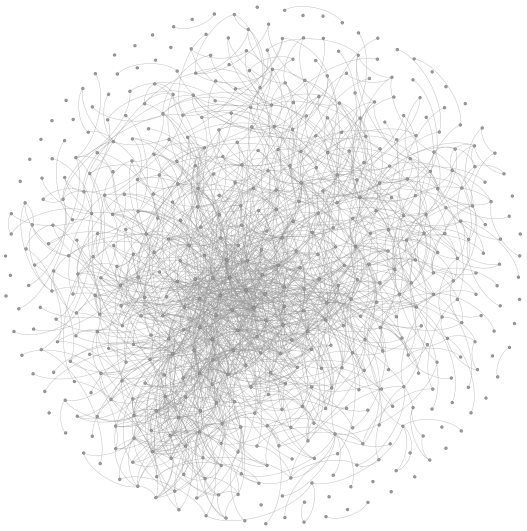

Figure 1: The graph of Erdos1 network.

the network properties in three aspects: degree centrality[8], between centrality[7] and modularity[4].

- **Degree Centrality** Degree centrality is computed by normalizing the degree of a vertex. In a graph, the degree of a vertex is defined as the number of links incident upon a node. Degree centrality measures the importance of a node to some degree. Nodes with higher degree centrality is generally considered more important than those with lower degree centrality. Degree centrality $C_d(i)$ is defined as follow:

$$C_d(i) = \frac{deg(i)}{(\|V\| - 1)},$$

  where $deg(i)$ is the degree of a node, and $\|V\|$ is the number of nodes in the network. Since $deg(i)$ of a node will not exceed the number of nodes excluding itself, it can be concluded that $0 \leq C_d(i) \leq 1$. Nodes in the network are ranked by degree centrality and sorted in descending order. Top 5 authors are listed in Table 1. Figure of degree centrality distribution is rendered by Gephi. Bigger size and deeper color in Figure 2 indicate larger degree centrality.

- **Betweenness Centrality** Betweenness centrality is another approach to reflect the importance of a node in a network. Betweenness centrality of a node relates to the number of the shortest paths between any two nodes in a graph passing through it; that's to say, the times of a node acting as an essential bridge between two randomly chosen nodes in a graph. Between centrality $C_b(i)$ is computed by following equation:

$$C_b(i) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}},$$

Table 1: The degree centrality of Co-authors (Top 5 authors)

Name	Degree Centrality
ALON, NOGA M.	0.036795
HARARY, FRANK*	0.032195
COLBOURN, CHARLES JOSEPH	0.021668
SHELAH, SAHARON	0.021259
GRAHAM, RONALD LEWIS	0.017989

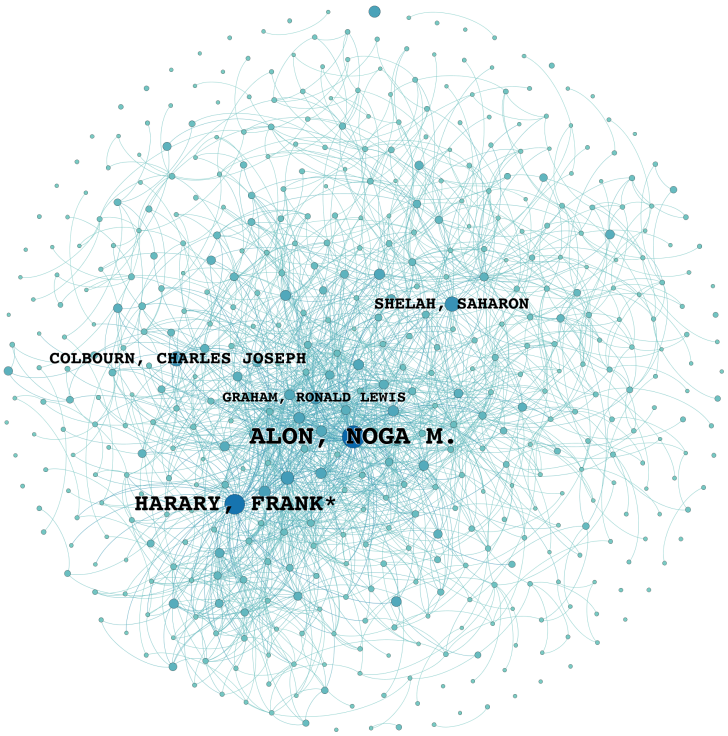

Figure 2: Degree Centrality Distribution

where σ_{st} is the number of shortest paths from s to t, $\sigma_{st}(i)$ is the number of paths passing through i. In this model, betweenness centrality of each node is computed and sorted in descending order. Top 3 authors with highest betweenness centralities is listed in Table 2. The distribution of betweenness centrality is shown in

Table 2: The betweenness centrality of Co-authors (Top 3 authors)

Name	Betweenness Centrality
ALON, NOGA M.	0.103416
HARARY, FRANK*	0.092314
GRAHAM, RONALD LEWIS	0.064862

Figure 3. The nodes with higher betweenness centrality are drawn with larger size and deeper color and usually considered to be important.

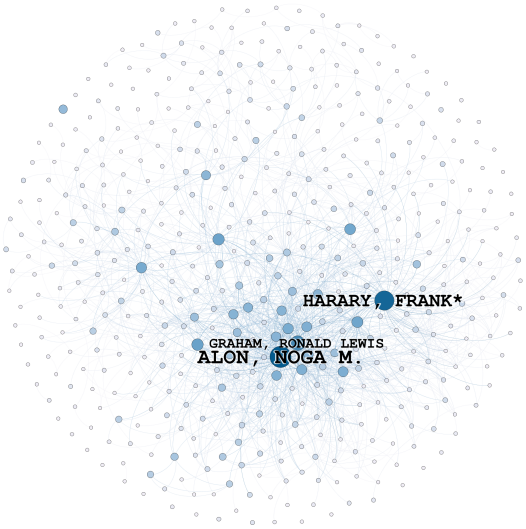

Figure 3: Betweenness Centrality Distribution

- **Modularity** Community is an critical concept about a network. A community consists a highly cohesive and low coupling group of nodes, whose properties are likely resemble to each other. As one of the most widely accepted indices of measuring communities in network, modularity describes how well a graph can be decomposed into modular communities. Here we present a graph in Figure 4, the nodes of which with same modularity are in the same group, therefore they are drawn in the same color.

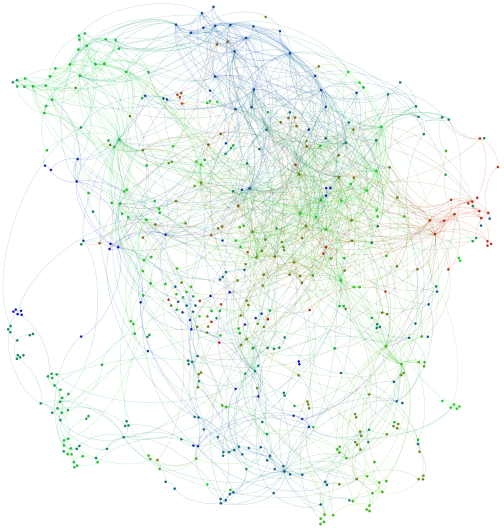

Figure 4: The community of co-author network

3 Discovery of Influential Researchers

Based on properties of Erdos1 network analyzed in previous section, we purpose several strategies in this section to measure the influence of a node in the network. On the basis that influential researchers are more likely to collaborate with other influential researchers and inclined to be aggregative, each strategy analyzes the connection between nodes and rank nodes with an importance index.

Following assumptions are adopted to simplify the model:
1) Suppose influence factor of a person is not associated with time in any form,
2) Treat multiple collaborations as one,
3) Suppose collaborators receive same benefit from a co-operation.

3.1 Estimating Reputation

A collaborative person tends to be more influential to some extent, but the accuracy of degree centrality remains as a weakness. See the graph in Figure 5, node q owns a good degree in contrast with its poor influence. The node in the center of graph can influence more nodes through connections to its neighbors, thus gains more importance. Despite of its disappointing accuracy, degree centrality is still reliable under most circumstances, so we take it into consideration as a factor representing the reputation of a person.

For each vertex in the network, degree centrality $C_d(i)$ is computed. Results of top-ranking nodes are given in the section 2.2.

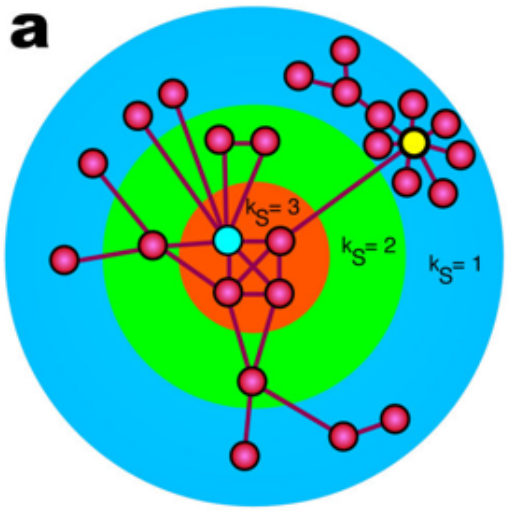

Figure 5: An example graph of incompatibility of degree centrality

3.2 Estimating the Quality of Connections

A researcher is a *hub* in a network model if he/she collaborates with a great number of influential person. Without hubs, nodes will be separated and it would be difficult for information to flow in the network. Based on definition of betweenness centrality given in section 2.2, our model considers betweenness centrality as an important contributor to the influence of a researcher.

For each vertex in the network, betweenness centrality $C_b(i)$ is computed. Results of top-ranking nodes are also given in the section 2.2.

3.3 Estimating the Significance of Works

Inspired by PageRank$^{\text{TM}}$[6] algorithm, our model assigns a value $P(i)$ for each vertex i in the network measuring the importance of their works. On the basis of a simple idea, a person that collaborates another researcher who published brilliant works are inclined to produce good papers. Let Erdös be the node with highest value of $P(i)$ and simulate the model. Similar to PageRank$^{\text{TM}}$ algorithm, our model collects value from incoming links and distributes value to outgoing links. Significance of works $P(i)$ can be computed by following equation:

$$P(i) = \sum \frac{1}{deg(B_i)}$$

where $deg(i)$ is the degree of node i, $B_i (1 \leqslant i \leqslant deg(i))$ is the adjcent node list of node i. The model repeats the iteration through network until the values converge. Results of top-ranking nodes are listed below.

Rank	Name	$P(i)$
1	ALON, NOGA M.	0.009105
2	HARARY, FRANK*	0.008801
3	SALAMON, PETER	0.007047
4	SHELAH, SAHARON	0.006528
5	COLBOURN, CHARLES JOSEPH	0.006075
6	HSU, DERBIAU FRANK	0.004817
7	TUZA, ZSOLT	0.004517
8	GRAHAM, RONALD LEWIS	0.00443
9	KOREN, ISRAEL	0.004197
10	WEST, DOUGLAS BRENT	0.004021

3.4 Comprehensive Assessment

As concluded above, it's obvious that $P(i)$ is the most significant variable, while $C_b(i)$ is less significant and $C_d(i)$ is the least. By AHP we assign important factor for each factor and uniformize three value as $S(i)$, calculating results are shown in the following table. *ALON, NOGA M.* is the most influential person in the network as he gains an outstanding value in this evaluation model.

Rank	Name	P_i
1	ALON, NOGA M.	0.002919368
2	HARARY, FRANK*	0.00258821
3	SALAMON, PETER	0.001805353
4	SHELAH, SAHARON	0.001793
5	COLBOURN, CHARLES JOSEPH	0.001713239
6	GRAHAM, RONALD LEWIS	0.001440285
7	TUZA, ZSOLT	0.001419713
8	HSU, DERBIAU FRANK	0.00128265
9	WEST, DOUGLAS BRENT	0.001235202
10	KLEITMAN, DANIEL J.	0.001111293

4 The model of reference network

To compare the influence between the given paper, we build the network for the papers which cite the given paper.

4.1 Data Extraction

We select 5 papers from the given files:
Newman, M. The structure of scientific collaboration networks. Proc. Natl. Acad.

Sci. USA, 98: 404-409, January 2001.

Bonacich, P.F., Power and Centrality: A family of measures, Am J. Sociology. 92: 1170- 1182, 1987.

Newman, M. Scientific collaboration networks: II. Shortest paths, weighted Kleinberg, J.

Navigation in a small world. Nature, 406: 845, 2000.networks, and centrality. Physical Review E, 64:016132, 2001.

Watts, D. and Dodds, P. Influentials, Networks, and Public Opinion Formation. Journal of Consumer Research, 34: 441-458, 2007.

4.2 Model Simulation

The node represent the paper and the directed edges connected with two nodes mean that the source node cite the target node. For the huge number of papers, we build the network with the papers have a distance within 3 of the given papers.

To measure the influence between the papers, we use the parameters called PageRank$^{\mathrm{TM}}$, degree centrality hub and authority to defined the influence

$$D(i) = \alpha P(i) + \beta C_d(i) + \gamma H(i) + \delta A(i)$$

where $D(i)$ is comprehensive evaluate value, the $P(i)$ is the Pagerank Value, $C_d(i)$ is the degree centrality and $A(i)$ is authority. For academic research network, authority should be considered as the most important parameter, so we chose the factor as $\alpha = 0.2, \beta = 0.2, \gamma = 0.1, \delta = 0.5$, after calculation, The paper called *The structure of scientific collaboration networks* is the most influential paper, but in other field, the factor should change with the situation. For individual researcher, we believed that the

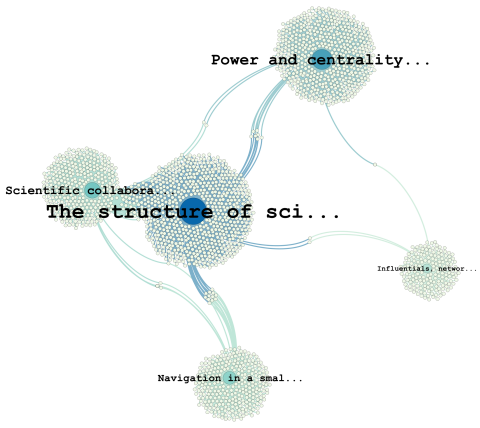

Figure 6: The reference network

PageRank$^{\mathrm{TM}}$ and hub is more important than the degree centrality and authority cause

the network for individual researcher is much more like a social network, but for specific university, department, or a journal, we have to introduce a new parameter called modularity which could measure the closeness about a group of node.

5 Analysis of Actors Collaboration Network

In this section we construct a network model of 10000 actors randomly sampled from IMDB to analyze their collaboration in movies and TV series. To identify most significant part in the network due to complex structure, the model used in this problem combines strategies described previously and new approaches.

5.1 Estimating Authority and Hub of Nodes

An actor will be more likely to cooperate with another actor with high value of hub rank H_i, which indicates he is a potential connector to great actors. Same applied to authority rank U_i. Hub rank H_i and authority rank U_i can be computed by following equation:

$$U_i = \alpha \Sigma R_i j y_j$$

$$H_i = \beta \Sigma R_i j x_j$$

where α, β, R is factors from HITS algorithm[5], y is imcoming neighbors list of i, x is outgoing neighbors list of i.

For each vertex in the network, its authority centrality U_i and hubs centrality are computed. As the network our model constructs is a undirected graph, $U_i = H_i$. We list several top-ranking actors with results below.

Rank	Name	U_i
1	Hytten, Olaf	0.001546483
2	O'Brien, William H	0.00143602
3	Blystone, Stanley	0.001413927
4	Homans, Robert	0.001104631
5	Travolta, John	0.001082538
6	Neeson, Liam	0.001082538
7	McLiam, John	0.001060445
8	Tovey, Arthur	0.001038353
9	Kennedy, John F.	0.000994194
10	Morgan, Gene (I)	0.000994168

5.2 Estimating the Significance of Works

As previously analyzed, nodes with high value of P is more likely to produce brilliant works. Thus we adopt the same algorithm as before.

Rank	Name	U_i
1	Travolta, John	0.000898182
2	Hytten, Olaf	0.000856221
3	Blystone, Stanley	0.000838392
4	Kennedy, John F.	0.000799519
5	O'Brien, William H.	0.00077497
6	Neeson, Liam	0.000730302
7	Homans, Robert	0.000684222
8	O'Brien, Conan	0.000656872
9	Voight, Jon 39	0.000637321
10	Tovey, Arthur	0.000620693

5.3 Comprehensive Assessment

As concluded above, significance of works is the most important variable, while authority and hub ranks is less important, and connecting great actors or actress is the least. We uniformize each measure value and use AHP to assign importance index to them. *Hytten, Olaf* is the most influential person in this network as he gains an outstanding value in this evaluation model.

Rank	Name	U_i
1	Hytten, Olaf	0.00566163
2	Blystone, Stanley	0.005343031
3	O'Brien, William H.	0.00519695
4	Travolta, John	0.004859621
5	Kennedy, John F.	0.004386892
6	Neeson, Liam	0.004355983
7	Homans, Robert	0.004261928
8	Tovey, Arthur	0.003938785
9	O'Brien, Conan	0.00369384
10	Voight, Jon	0.003679371

6 Model Assessment

The influence are different in many fields, but the common thing is that the influence have to be sent from the source and receive by the target, which means that all those

processes can be seemed as a great network. The influence can flow through the edge like the stream. The node just like the pool that can store the reputation as water and be the source of the stream, but the fact is that the reputation will not reduce as the node influence other nodes, the reputation will only goes away as time goes by.

The real situation about influence is complex. Cause the influence changes with time, but the network can not tell the difference made by time. We assume that the network will not be changed by the time, and analyze the network to find the most influential node. The result will be a little different from the fact cause we simply ignore the factor about time but it can be used to give the guidance suggestion. In the previous work about the influence of actors, we examine the result with the imDB database and found that the result we get is not completely the same as the real rank but the difference can be controlled in a small range.

We believed that the model can be used to assess the influence between individuals, organizations and nations and give the advise to make decision.

For instance, at individual level, to know who is the best co-author for us to cooperate to boost our mathematical influence as rapidly as possible, we could get the cite data about this field and delete the nodes which have relatively small degree to reduce the number of node in the network. After data extraction, apply the network model to the given data and calculate the parameters: authority, degree centrality, betweenness centrality, community and Pagerank value. Then we can change the factors of those parameters to satisfy our requirement: To increase the reputation in mathematical field by making new breaking through, we have to cooperate with the author who has greater authority cause he has a greater possibility to make new theory.

7 Weakness and Strength

Our model fits for large datasets. We collected tons of data in past four days, and simplify them while keeping useful information, ensuring our analysis is more accurate. The model can solve problems in a more comprehensive way as we combine several different properties of the data.

But There still exists two major problems. First of all, we do not take time into consideration, which means that our model is static and loses its flexibility. Older nodes in a network model have advantages over other nodes, and generally the influence will be impaired as time goes by. Secondly, auxiliary information can not be collected to evaluate co-operations between two nodes due to complex data processing and large memory requirements.

References

[1] Fruchterman T M J, Reingold E M. *Graph drawing by forcedirected placement*[J]. Software: Practice and experience, 1991, 21(11): 1129-1164.

[2] Bastian M, Heymann S, Jacomy M. *Gephi: an open source software for exploring and manipulating networks*[C]//ICWSM. 2009.

[3] Hagberg A, Swart P, S Chult D. *Exploring network structure, dynamics, and function using NetworkX*[R]. Los Alamos National Laboratory (LANL), 2008.

[4] Newman, M. E. J. (2006). *Modularity and community structure in networks*. Proceedings of the National Academy of Sciences of the United States of America 103 (23): 85778696.

[5] Christopher D. Manning, Prabhakar Raghavan Hinrich Schtze (2008). *Introduction to Information Retrieval*. Cambridge University Press. Retrieved 2008-11-09.

[6] Page L, Brin S, Motwani R, et al. *The PageRank citation ranking: bringing order to the web*[J]. 1999.

[7] Freeman, Linton (1977). *A set of measures of centrality based on betweenness*. Sociometry 40: 3541.

[8] Newman, M.E.J. 2010. *Networks: An Introduction*. Oxford, UK: Oxford University Press.