2014 Mathematical Contest in Modeling (MCM) Summary Sheet

Hereby we construct 3 sets of influence measurement to evaluate 4 models to evaluate 3 types of networks. We develop a network-split method to fasten the implementation of PageRank and HITS algorithms. Combined with time series modeling, we provide a compass for research.

In the 1st model, we build an author network connected by collaborations. The analysis presents us a high clustering network with a power law degree distribution indicating scale free. Most authors are connected through limited “connectable” people. For connection influence measure, our PageRank centrality overcomes shortcomings of the non-weighting Katz Centrality and simple centralities. We found Harary is the most important and “connectable” author.

In the 2nd model, we analyze the paper citation network. The directional nature of the paper citation inspired us to introduce a new influence measure, authority and connectivity (hub), through the HITS algorithm. To improve the accuracy, we considered the out-of-circle citation and transformed it into an initial value for the equation set. The citations among same authors and journals are also lessened. Our high hub papers accurately equal to review paper.

For the semi-bipartite paper-author combined model, we develop the third and fourth models and lastly take the time element into consideration. In two models, the paper citation network is separated from the entire network to form the source citation influence. The third model considers paper-to-author one directional effect, thus creating a tree diagram. Authors and other academic entities collect citation influence from the source. It could be computation-efficiently used level-by-level to evaluate departments, universities and nations due to network similarity. For the fourth model, we put the paper and author two-way connected to retrieve the author-author connection effect and to reveal the paper and author influence from a cognitive sense. The time series analysis shows the network science emerging in 2000s.

Then we implement our method into the film-actor network of 007 film series, though the network is one-dimensional, the results showed buried popularity of Quantum of Solace and Judi Dench. With the time series analysis, the influence measure package offers flexible but effective choices in finding collaborator, innovator and discover a new filed. The sensitivity analysis indicates parameter is strong below maximum value, and we choose the optimal value within convergent range to give a high resolution of network influence.
Splitting Networks to Locate Academic Stars

Contents

Title ............................................................................................................................................ 2
Introduction .............................................................................................................................. 2
Terminology and Definition ..................................................................................................... 3
Basic Model.................................................................................................................................. 3
Model Overview ....................................................................................................................... 3
Author Network .......................................................................................................................... 4
Paper Network .......................................................................................................................... 8
Improved Model for Complex Networks .................................................................................. 11
Scientific Influence Measure .................................................................................................... 13
Cognitive Academic Influence Measurement ............................................................................ 14
Application ................................................................................................................................. 16
Analyze a Film Series .................................................................................................................. 16
Chasing the future ...................................................................................................................... 17
Evaluating our model .................................................................................................................. 19
Sensitivity Analysis ................................................................................................................... 19
Strength and Weakness ............................................................................................................ 20
Reference ...................................................................................................................................... 20

Introduction

Scientific research is exploring unknown world with known results, therefore it requires accumulation of known knowledge in certain field. On the other hand, it depends on highly innovative thoughts to break the conventions. A non-trivial fact is individual researcher experienced increasing difficulty in achieving new finding. Only cooperative spirit could help to challenge interdisciplinary topic, and a good example is nano science. To facilitate networking and protect the intellectual property, the authorship network becomes one of best documented networks. Hence, utilizing data and modeling to evaluate influential research work becomes possible.

On 2005 Hirsch first invented H-index to evaluate the author by the significance of publications. Currently, it has been accepted as default parameter in big research search engines like Web of Science and Google Scholar. Though this method combines the quality and productivity but this do not equally link to the authors’ academic influence, especially the innovation and connection. There are following shortcomings.
1) Firstly, measuring the influence of a paper by citation number is biased. Self-citation and Matthew Effect exaggerated the number.

2) Secondly, the link between papers and authors’ influence is biased. The number of papers, the journal numbers in the field and the author contribution is not accounted for.

Hereby we want to develop a set of analysis to eradicate such imbalance. To describe the influence of node, we employed the authority and connectivity. Authority indicates the innovation and authority of the content, telling reader what is unknown. Connectivity serves as a bridge providing the relation of authority contents. On internet, the authority pages are those like YouTube, CNN, and the connecting pages (also as hub) are those like Google and AOL. In this paper, we utilized the network analysis to address the influential point.

**Terminology and Definition**

- **Connection Influence (CI)**: influence in an author network.
- **Scientific Influence (SI)**: The influence through the citation network.
- **Academic Influence (AI)**: Cognitive academic influence, the influence in researcher’s mind.
- \( x_i \): Authority Centrality (Authority); \( y_j \): Hub Centrality (Hub)

- \( a^A_i \): Authority Centrality of \( Au_i \); \( h^A_i \): Hub Centrality of \( Au_i \)
- Citing_Paper: paper that cites other papers
- Cited_Paper: paper being cited
- Outer citation is the citation a paper obtained from outside the sample network.
- Inner Citation is the citation a paper obtained from outside the sample network.
- For a certain paper node in the paper-author network, the author-paper degree \( d_i^{a-p} \), paper’s out-degree \( d_i^{out} \) and in-degree \( d_i^{in} \) is defined as below:

\[
\begin{align*}
d_i^{a-p} &= |D_i^{a-p}|, \quad D_i^{a-p} = \{A_{ij} | A_{ij} = 1, j = 1, 2, ..., m\}, i = 1, 2, ..., n \\
d_i^{out} &= |D_i^{out}|, \quad D_i^{out} = \{B_{ij} | B_{ij} = 1, j = 1, 2, ..., m\}, i = 1, 2, ..., n \\
d_i^{in} &= |D_i^{in}|, \quad D_i^{in} = \{B_{ji} | B_{ji} = 1, j = 1, 2, ..., m\}, i = 1, 2, ..., n
\end{align*}
\]

**Basic Model**

**Model Overview**

During research, researcher cooperates with others to complete a paper and papers keep citing and being cited. From the view of graph theory, the authors and papers are nodes and cooperation and citation are edges. In this paper, we draw a network model to analyze the influence. Following the instruction of the ICM contest, we developed three levels of model to evaluate the influencer in a society. First level is connection based, and in this level we only use
author as the node and analyzed the connection, coauthor-ship between the cooperators. The second level is paper based model, in which we considered the direction and the weight of the connection. The last level of the model, we utilized a hierarchical structure and two-step method to measure the influence. For three level of networks, we utilized three sets of influence parameters.

Author Network

Network

- **Assumption:** For the network of authors presented by file *Erdos1*: Every coauthor of Prof. Erdos in *Erdos1* is represented by a node, denoted by $V_1, V_2, \ldots, V_{511}$. The collaboration between two authors, $i$ and $j$ is represented by an undirected edge $E_{ij}$. We have $V = \{V_i | i = 1, 2, \ldots, 511\}$, $E = \{E_{ij} | i, j = 1, 2, \ldots, 511\}$ and the network Erdos1 is represented by $G(V, E)$. Their collaboration is non-directional and non-weighted.

- **Network Construction:** Using C programming, we filtered the related 511 coauthors and recorded their cooperation from file. From the definition, we depict the network by following properties.

- **Properties**
  1) **Basic Properties**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Parameter/Type</th>
<th>Subject</th>
<th>Parameter/Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertex type</td>
<td>Author</td>
<td>Maximum degree</td>
<td>52 edges</td>
</tr>
<tr>
<td>Edge type</td>
<td>Collaboration</td>
<td>Volume</td>
<td>1640 edges</td>
</tr>
<tr>
<td>Format</td>
<td>Undirected</td>
<td>Diameter</td>
<td>10 edges</td>
</tr>
<tr>
<td>Edge weights</td>
<td>Unweighted</td>
<td>Largest connected component</td>
<td>474 vertices</td>
</tr>
<tr>
<td>Size</td>
<td>511 vertices</td>
<td>Mean shortest path length</td>
<td>4.19 edges</td>
</tr>
<tr>
<td>Fill</td>
<td>0.0125 edges / vertex$^2$</td>
<td>Gini coefficient</td>
<td>61.1%</td>
</tr>
<tr>
<td>Average degree (overall)</td>
<td>3.209 edges / vertex</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2) **Degree Distribution:**

The degree $d_i(g)$ of a node $i$ in graph $g$ is the number of $i$'s neighbors in $G^4$. Degree distribution is statistical distribution of the degree of coauthors in erdos1 network, which is often assumed to obey a power law.

From the power law, we found the Erdos-coauthor network is close to a scale-free network, which is social network nature. This scale-free network presents a hierarchical structure. Below we plot the Lorentz Curve of the degree distribution. Obviously, it is, also can be evidenced by a Gini coefficient 61.1%, not a well distributed network. Hence, we can conclude the Erdos-coauthor network to be a clustered network, which means a single author tends to cluster to those influential authors to construct his/her own co-work network.
3) Closeness and Betweenness Centrality Distribution

The farness of a node $i$ is defined as the sum of its distances to all other nodes, and its closeness is defined as the averaged inverse of the farness. Freeman’s betweenness quantifies the number of times a node acts as a precursor along the shortest path between two other nodes. We calculated the data by Software UCINET. The plot below shows the distribution of closeness and betweenness centrality values among all the nodes authors.

Closeness describe average connection to other authors. There are two peaks in the plot, around 0 and 2. It is obvious this network is not a well-connected. Betweenness describes the author’s ability of connecting other authors. The probability density diminish quickly after the betweenness increased. We then realize most authors are not active connectors and only a few connects most of the others, agrees with the conclusion of degree analysis that Erdos1 network is hierarchical clustered network structured in topology. This shows the “funneling” the author collaborator. This agrees with the clustering results.

4) Distance

The geodesic distance $d(i, j; g)$ is the number of links in a shortest path between $i$ and $j$. Specifically, if there is no path between $i$ and $j$ in $g$, we set $d(i, j; g) = \infty$.

Distance distribution describes the connections between most authors. Mean shortest path length obtained is close to the result in SIPRES, indicating a similar topology trend between two network structures, i.e. significant collaboration network properties.
5) **Cluster coefficient distribution:**

We defined the clustering coefficient of a node in a graph as the proportion of that node's neighbors that are connected, and proceeded to define the clustering coefficient as the corresponding measure applied to the whole network. In some case however, we may be interested in the distribution of the clustering coefficient over the nodes in the network. For instance, a network could have some highly clustered parts, and some less clustered parts, while another network could have many nodes with a similar, average clustering coefficient. Thus, we may want to consider the distribution of clustering coefficient. This distribution can be plotted as a cumulated plot. From the plot of XX, we found most clustering coefficient falls into the interval of \([0.8, 0.9]\). The data of author collaboration network in KONECT\(^{10}\) also validated that most nodes (40\%) are of high clustering.

![Figure 5 Clustering Coefficient with Comparison to Axiv Astro-Ph Network](image)

**Influence Measurement**

1) **Assumption:**
   - The influence of coauthor \(i, (i = 1, 2, \ldots, 511)\) is represented by centrality \(x_i\). An author is determined to be influential if he connects important researchers in Erdos1 or connects a large amount of researchers. Each coauthor is arbitrarily allocated with the same initial centrality 1 since only the relative rank matters.
   - Author \(i\) distribute his centrality equally to collaborator. For example, author \(i\) had collaborated with \(j\) and \(k\) for 4 and 1 times, then he distributes his centrality to \(j\) and \(k\) by the portion of \(4/5\) and \(1/5\).

2) **Centrality Measure:**
   - **Degree Centrality:**
     It measures how many vertexes are directly connected with author \(i\). However, degree centrality will account for the influence propagation through networks.
   - **Betweenness Centrality**
     In the author network, finding out the most connectable people will help researchers to locate collaborator, therefore it may also be an influence measure.

3) **PageRank(PR) Centrality Measure:** The centrality of author \(i\) is defined here the weighted sum of his neighbors’ centrality and his initial centrality\(^{11}\):

\[
x_i = \alpha \sum_D \frac{A_j}{D_j} x_j + x^0
\]

Since centrality is based on neighboring nodes, we have to calculate the centrality by an iterative method. Based
on the aforementioned assumptions, each coauthor has its initial centrality $x_i^0 = 1$. $\alpha$ represents the weight of his neighbors’ contribution. In this case, $\alpha < 1$ (i=1,2,…,511), i.e. neighbors contribute less than initial value does. $D_j$ denotes the times of collaboration that involves $j$.

The matrix form:

$$X = (I - \alpha AD^{-1})^{-1}X_0 = D(D - \alpha A)^{-1}X_0$$

For the convergence of every component of $X$, $(D - \alpha A)^{-1}$ exists. $\alpha$ is selected.

**Validation and Conclusion**

According to centrality analysis, we ranked the following authors.

<table>
<thead>
<tr>
<th>Rank</th>
<th>No.</th>
<th>Name</th>
<th>Degree Centrality Rank</th>
<th>Betweenness Centrality Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>ALON, NOGA M.</td>
<td></td>
<td>187 HARARY, FRANK*</td>
</tr>
<tr>
<td>2</td>
<td>165</td>
<td>GRAHAM, RONALD LEWIS</td>
<td></td>
<td>438 SOS, VERA TURAN</td>
</tr>
<tr>
<td>3</td>
<td>187</td>
<td>HARARY, FRANK*</td>
<td></td>
<td>385 RUBEL, LEE ALBERT*</td>
</tr>
<tr>
<td>4</td>
<td>44</td>
<td>BOLLOBAS, BELA</td>
<td></td>
<td>449 STRAUS, ERNST GABOR*</td>
</tr>
<tr>
<td>5</td>
<td>378</td>
<td>RODL, VOITECH</td>
<td></td>
<td>355 POMERANCE, CARL BERNARD</td>
</tr>
<tr>
<td>6</td>
<td>148</td>
<td>FUREDI, ZOLTAN</td>
<td></td>
<td>148 FUREDI, ZOLTAN</td>
</tr>
<tr>
<td>7</td>
<td>479</td>
<td>TUZA, ZSOLT</td>
<td></td>
<td>10 ALON, NOGA M.</td>
</tr>
<tr>
<td>8</td>
<td>438</td>
<td>SOS, VERA TURAN</td>
<td></td>
<td>165 GRAHAM, RONALD LEWIS</td>
</tr>
<tr>
<td>9</td>
<td>440</td>
<td>SPENCER, JOEL HAROLD</td>
<td></td>
<td>44 BOLLOBAS, BELA</td>
</tr>
<tr>
<td>10</td>
<td>177</td>
<td>GYARFAS, ANDRAS</td>
<td></td>
<td>341 PACH, JANOS</td>
</tr>
</tbody>
</table>

We take $\alpha = 0.85$ as Google Page-Rank Algorithm recommended to compute PR centrality.

![PR centrality(0.85)](image)

**Figure 6** Top 10 influential authors sorted by centrality under alpha=0.85
In conclusion, we rank coauthors using PR centrality to determine the most influential authors in Erdos1. However, betweenness can serve as a complement in the way that it can describe how well an author is connected. Some of the top ranking authors are shared, like Harary. Frank, who ranks the first in both measurements. He collaborated with 44 coauthors and many of them have considerable PR centrality. However, some of them are not. For instance, Rodl. Vojtech, who ranks 6th in PR centrality, is not included in the top-ten list of betweenness because his connection is relatively limited within several authors of high PR centrality.

Taken in this sense, our model is effective in describing different type of influence. In reality, the choice of index will depend on the usage of measurement. Restricted by information, our model here still has following weakness.

1) We omit the influence from the authors outside the sample, which is important when the sample is not big enough to cover the entire circle. An alternative would be to introduce the information by measuring the connection with Erdos, who serves as a benchmark of academic influence.

2) The connection between authors is non-weighted, which indicates a bad or good collaboration will establish the identical edge. Since good collaboration produces better papers, the paper quality is a good weight index.

3) In our model, we do not considered the time of collaboration.

**Paper Network**

In real world, we could also use paper influence to define the academic influence. The listed materials offered a network of 16 paper.

![Figure 7 Clustering of The 16 Paper Network](image)

**Assumption**

1. In this model, papers are regarded as nodes, and citations are considered edges. In the sample, we have 16 node: \( P_1, P_2, \ldots, P_{16} \). Paper \( i \) receive an Outer Citation from papers outside of the sample, which is \( c_i(t) \) at the time \( t \). A vector is \( P_jP_i \), defined to denote that Paper \( i \) is cited by Paper \( j \). \( d_{out}^j \) and \( d_{in}^i \) represents the total number of vectors starting from \( P_j \) and ending at \( P_i \). We have the adjacency matrix to describe the citation of papers.
\[ A_{ij} = \begin{cases} 
1 & \text{i cited by j} \\
0 & \text{else} 
\end{cases} \]

\[ V = \{ P_i | i = 1, 2, \ldots, 16 \}, E = \{ P_j P_i | A_{ij} = 1, i, j = 1, 2, \ldots, 16 \} \text{, the network is called } G_p(E, V) \]

The influence of \( P_i \) is denoted as \( x_i \), which is affected by the initial influence and the network influence. An initial influence \( x_i^0 \) is to account for the influence contributed by the outside part within the entire network. Network influence is defined by the relative importance within the network.

2. For the small network sample, outside influence is identical with no weight on the quality.
3. The influence of a paper \( x_i \) could be accumulated by being cited but citing others will not do this favor. The accumulation of a Cited_Paper will be diluted according to the numbers of total papers being cited by the Citing_Paper.
4. If the upper level author-paper or journal-paper network information is available, we take into account the self-citation in authors and journals. Authors tend to cite his own paper and the previous paper of the journal he plan to submit his manuscript. Inner citation will be less weighted.

**Influence Measure**

**Centrality**

Here we use centrality again to measure the influence. According the assumptions, the centrality consists two parts: Initial centrality and Network centrality, which are expressed as below:

\[ x_i = \alpha \sum_j A_{ij} \frac{x_j}{d_j^\text{out}} + x_i^0 \]

\( \alpha \) stands for the relative importance of the initial influence and network influence. Different from original Page-rank algorithm, here \( d_j^\text{out} \) is the filtered connection that attracts our interest. To depict \( x_i^0 \), we employed the easy-to-get data \( c_i \), citation from all the papers outside the sample. To bridge the gap to real outside network effect, we consider a citation with longer path to the paper in circle will accumulate less influence to the cited paper. We adopt a function with diminishing marginal contribution of the citation to the influence:

\[ x_i^0 = \ln[c_i + 1] + 1 \]

Here, we have an exponentially diminishing influence with the time interval.

Besides, according to assumption 5, citation between same author should contribute less, which will be denoted as a author-neutralization factor: \( \mu_{ij} \). Similarly, journal-neutralization factor is denoted by \( \rho_{ij} \).

\[ \mu_{ij} = \begin{cases} 
0.6 & \text{same author} \\
1 & \text{not same} 
\end{cases} \quad \rho_{ij} = \begin{cases} 
0.8 & \text{same journal} \\
1 & \text{not same} 
\end{cases} \]

With all these factors taken into account, we have:

\[ x_i = \alpha \sum_j \rho_{ij} \mu_{ij} A_{ij} \frac{x_j}{d_j^\text{out}} + \ln[c_i + 1] \]

As we broaden our sample, more and more papers will be taken into consideration, making initial value, which
describes the outer citation, contributes less. Finally, when all published papers are included, the initial value will all be 1.

Combine all the equations into matrix form:

\[ \bar{X} = \alpha A_{\text{revised}} D^{-1} \bar{X} + \bar{X}_0 \]

Where \( A_{\text{revised}} = \rho_i H_j A_{ij}(t) \)

\[ \bar{X} = (I - \alpha A_{\text{revised}} D^{-1})^{-1} \bar{X}_0 = D^{-1} (D - \alpha A_{\text{revised}})^{-1} \bar{X}_0 \]

Again considering the convergence requirement, \((I - \alpha A_{\text{revised}} D^{-1})^{-1}\) exists and \(0 < \alpha < \lambda^{-1}_i\), where \(\lambda_i\) is the maximum eigenvalue of \(A_{\text{revised}} D^{-1}\).

**Authority and Connectivity**

The centrality computed above describes the paper influence in a comprehensive way. But there are research papers in different functions which require different measurements. One typical classification is the original paper and review paper. The original paper offers innovative findings. Though the review paper is not in an innovative status, it cites a large number of importantadvances in the field and offer prospective view. The topological difference of two papers would be the value of \(d_{i\text{out}}\). This difference is not significant in aforementioned sample. But to offer better monitoring of such paper and related authors in an emerging field, two-index measurement and HITS algorithm are introduced.

Within the same network \(G(V, E)\), Paper \(i\) contains two influence indexes:

\(x_i: \) Authority Centrality; \(y_i: \) Hub Centrality

Here authority will attract more Paper with more authority will attract higher citation from review paper. Also, if a review paper link to more author, it has a higher probability to be influential.

Therefore, we have three steps to implement the calculation:

- **Initial Centrality Allocation**: As aforementioned assumption, each node is allocated with an initial centrality, \(x_i = \ln(c_i + 1)\), \(y_i = 1\), where \(c_i\) is the outer citation number.

- **Authority Centrality Update**: Update the \(x_i\) by the sum of neighboring \(y_j\).

\[ x_i = \alpha \sum_j a_{ij} y_j + x_i^0 \]

\(\alpha\) evaluates the relative contribution of neighboring \(y_j\). We will discuss the value later.

- **Hub Centrality Update**: Update the \(y_i\) by the sum of neighboring \(x_j\)

\[ y_i = \beta \sum_j a_{ij} x_j + y_i^0 \]

\(\beta\) evaluates the relative contribution of neighboring \(x_j\).

In matrix form:

\[ \bar{X} = (I - \alpha\beta A A^T)^{-1} (\alpha A Y_0 + \bar{X}_0) \]
and $\overline{X_0} = (x_1^0, x_2^0, ..., x_n^0)$, $\overline{Y_0} = (y_1^0, y_2^0, ..., y_n^0)$. If a graph has $(I - \alpha \beta AA^T)^{-1}$, $\overline{X}$ is convergent.

**Validation and Conclusion**

We adopt $\alpha$ equals to 0.85 in Page-Rank algorithm and both $\alpha$ and $\beta$ equal to 0.2 in HITS algorithm. Below are our results to the influence measure of the listed 16 papers

![Figure 8 Centralities calculated by PageRank Algorithm and HITS Algorithm](image)

Here we notice that the most influential paper would be the No.5 paper *Collective dynamics of ‘small-world’ networks*, and its influence comes from two parts, No.1 in outer citation and being cited frequently by influential works.

We can also conclude that ranking by centrality (PR centrality) in Page-Rank algorithm or authority in HITS basically will get the same results. However, hub centrality can be quite different because it indicates a paper’s review ability. For instance, paper 13 is not outstanding in terms of authority but has an extremely high hub centrality. This corresponds to the fact that it has cited most of the influential works in this network, including paper 5 and paper 6. A good validation results would be our top hub papers (No.11 and 13) are all from Review Papers.

To conclude, our model is effective in describing different aspects of influence and can be self-adjusted to satisfy outer conditions.

**Improved Model for Complex Networks**

The academic network consists of the papers, authors, departments and journals. It consists of paper-paper citation network and paper-author authorship network. The former is a directional single network and non-directional bi-partite network. It is easy to note the author is the transition node of two different networks. For author-department-university, it is follows a tree network. Only in some cases an author will belongs to two departments, we could utilize a similar treatment.
In the large network, the basic and characteristic structure would be the author-paper network. After we analyzed the network, we could level up into departments, universities and nations. In our above methods, we analyzed the separate network of papers and authors. However, the connection of author and paper will also offer information to the influence analysis.

To analyze academic influence from the above network, a key question is to define the influence we are looking for. We consider paper and citation the benchmark since they connect everything. According to our assumptions, the paper network is firstly decoupled and analyzed. Then a natural way is to form statistical measurement of author and journal influence by collecting related paper scientific influence and their relation to authors and journals. It becomes our first influence measure for the tree network.

But we note information is missing using above measure. To include such information, we retrieved the authors and journals’ effect on papers. Then we develop the academic influence measure to indicate a researcher’s viewpoint of what is influential to restrain the citation “inflation”.

Assumptions

1. There are two types of influence measure, the scientific influence and academic influence.
2. Papers are represented as nodes $P_i, i = 1, \ldots, n$ and authors as nodes $A_{ui}, i = 1, \ldots, m$. Authorship are denoted as an undirected edge. Citation between papers is denoted as a directed edge which points to the Cited_Paper.

   The adjacency matrix of paper-author network $A (n \times m)$ and paper network $B (n \times n)$ is defined below:

   $A_{ij} = \begin{cases} 1 & \text{if author } j \text{ writes paper } i \\ 0 & \text{else} \end{cases}$

   $B_{ij} = \begin{cases} 1 & \text{if } P_i \text{ cited by } P_j \\ 0 & \text{else} \end{cases}$

3. When we evaluate author’s influence, an influential paper will lend influence to its author though he is do not held good collaboration influence. The author’s authority is determined by his papers’ authority. The author’s connectivity is determined by his papers’ connectivity, i.e. hub centrality.
4. Different coauthors have different contribution for the paper therefore share the different influence. In term of the importance, first author and communication author held first and second places, and the rest are listed by the author rank appear on the paper. Then we use Contribution Factor $\eta_{ij}$ to distribute the influence.
Scientific Influence Measure

- **Mechanism**: Scientific influence measure is the influence indicator based purely on paper citation. It is based on the influence measure in the paper citation network, and paper citation influence flows into a larger network. Authors, Journals, Departments and Universities are interpreted as higher level nodes of a hierarchical network. Authors and journals build their influence from his published paper. Department and universities support authors who publish paper and collect influence from authors. Journals publish paper and get connected through paper citation to other journals. The collection of influence is weighted and single directional. Due to the comprehensive of paper citation network influence measure, it offered a good alternative for the H-index and Influence Impactor (IF).

- **Indicator**: We still adopt authority and hub centrality to indicate the innovation and connection abilities.

- **Assumptions**:
  1. Author Connection: Authors connect each other only through papers and the collaboration should be measured by paper quality. A better paper indicates stronger collaboration. Only academic connection is considered effective connection.
  2. Author-paper relationship: Consider the integrity of academic, author will not affect the citation of papers. Therefore, we could split semi-bipartite network into paper citation network and paper-author network without paper citation. However, the effects of self-citation is still considered through the author-neutralization factor $\mu_{ij}$.

- **Influence measure**
Here we take Author influence for example, which could be used for journals, universities and departments. The author collects influence by a factor of contribution:

$$a_i^A = \sum_{j=1}^{n} \eta_{ij} A^p x_j + a_i^{A(0)}$$

$\eta_{ij}$ is the contribution factor defined in the assumption 4. $a_i^{A(0)}$ represents the initial authority of author $i$, which
is the scientific citation influence. Hence, adjacent matrix between paper and author $A$ is modified into $A_I$, whose components are $\eta_{ij}A_{ij}$. We calculate the authority and connectivity through the similar way:

$$\overrightarrow{a}^A = A_I^T \overrightarrow{X} + a^{A(0)}; \quad \overrightarrow{c}^A = A_I^T \overrightarrow{Y} + c^{A(0)}$$

For journals we should only replace author by journal in the formulation. At the same time, we need to collect the adjacency matrix between journal and paper. For departments, we should replace author by departments and the paper by authors. Then we collect the adjacency matrix between departments and paper. The adjacency matrix records the affiliation relation.

After we had the adjacency matrix, we may able to locate the high innovation journal and high connection journal. After identifying paper’s journal and the author’s institutions (There may be more than one institution as discussed in Mechanism), the algorithm is identical to the statistical work without reflecting upper level’s influence on lower level.

**Cognitive Academic Influence Measurement**

- **Assumption:**
  1. Author’s academic influence is not equal to citation influence. The academic influence will take effect in obtaining funding and leading research topic. Since it works through conference and peer-review, it will highly depends on other researchers’ opinion towards him. **Matthew Effect** tells that the renowned authors’ papers usually are overvalued. Also once people find a researcher made a less-wise move, they will exaggerate their disbelief for other papers. In a cognitive sense, the paper influence should be balanced according to the author. All the above information is not presented by the citation therefore the author influence basing on the citation is biased.
  2. For researchers new to the field, the author connection would be also important. If we reconsider our assumption on author connection, we will find the connection is not totally weighted by paper quality
  3. Author’s academic influence notation:

    Authority: author: $a^A_{i^A}$, paper: $a^A_{i^P}$, $\overrightarrow{a}^A = (a^A_{i^A}, a^A_{i^P})^T$, $a^{A(0)} = (\overrightarrow{0}, \overrightarrow{X})^T$.

    Hub centrality: author: $h^A_{i^A}$, paper: $h^A_{i^P}$, $\overrightarrow{h}^A = (h^A_{i^A}, h^A_{i^P})^T$, $h^{A(0)} = (\overrightarrow{0}, \overrightarrow{Y})^T$.

- **Algorithm:**

  Since the author-paper-paper-author network has semi-bipartie properties, then we experienced difficulties in algorithm implement. Considering the huge data size for the author-paper network, if we analyze the network effect within one matrix, it will increase complexity exponentially, like 4 times for bipartite, 9 times for tripartite... We want to simplify the calculation by two-step algorithm. The algorithm will breakdown the network into two components, and the connection missed here will be compensated later through.

  **Step 1:** Scientific influence is calculated from citation network and serves as the initial value for academic influence paper-author network.

  **Step 2:** Author $i$ authority should be updated:
\[ a_i^{A\rightarrow AI} = \sum_{j=1}^{n} \eta_{ij} A_{ij} a_j^{P\rightarrow AI} + 0 \]

\[ a_i^{P\rightarrow AI} = \sum_{j=1}^{n} \eta_{ji} A_{ij} a_j^{A\rightarrow AI} + x_i \]

\( \eta_{ij} \) is the contribution factor defined in the assumption 5. Initial authority centrality \( a_i^{A(0)} \) represents the authority outside our sample if needed. It can be taken as 0 if only relative influence within this network is considered.

Hence, adjacent matrix between paper and author is modified into \( A_I \), whose component is \( \eta_{ij} A_{ij} \). Rewrite this in matrix form:

\[
\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix} a^{A\rightarrow AI} = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix} a^{A\rightarrow AI} + a^{A(0)}
\]

Similarly, hub centrality

\[
\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix} h^{A\rightarrow AI} = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix} h^{A\rightarrow AI} + h^{A(0)}
\]

The distortion caused by journal and time decaying has already been considered so it will be passed to author through initial value.

**Validation and Conclusion**

The calculation results of Scientific Influence and Academic Influence are illustrated above. We can find SH Strogatz ranks down in Aca. than Sci., which can be owe to the decrease of weight of DJ Watts in this network. This decrease afterward decreases the influence of Collective dynamics of `small-world' networks which is a co-operated publication between them two. We can obtain the conclusion that academic influence measurement induced the
influence between authors by author-paper-author bond, i.e. intra-network interaction.

Besides, we can find HITS algorithm give us a pretty close result to PR algorithm in authority centrality. But we can also find hub centrality ranks significantly different with PR centrality. Though MEJ Newman is still a leader in the hub rank, R Albert shows really higher hub rank than authority, indicating high quality citation in his publication, i.e. the HITS algorithm gives us more information about this academic network which we cannot obtain by PR algorithm.

**Application**

**Analyze a Film Series**

Film stars collaborate in one film and build connections, which is in similar pattern with author collaboration. Therefore, we retrieved the actor/actress connection and film information the nearest 8 James Bond films to form an actor/actress-film network. In this analysis we used the scientific influence measure, which takes two steps.

**Assumption:**
- We evaluate the influence of 007 series to movie audience by separating film network and film-actor network.
- The collaboration is only valued by film collaboration and the strength is weighted by the success of the film.
- The influence of a film is not only indicated by its box office sales but also by the sales of next movie.
- The film transmits influence to the actor but the inverse transmission is neglected.

To calculate the success, we employed a factor of Time FeedbackFactor (TFF), which could be expressed below:

Define film series through \( f_i, f_2, f_3, \ldots, f_n \) (\( i \neq n \)), \( A_i, i+1=TFF \),

\[
f_i = \sum_{j} A_{ij} f_j + (1-TFF) f_i^0
\]

\[
f_i^0 = \log_{1.3} \left( \frac{G_i}{G_{\text{min}}} \right) + 1
\]

For the initial centrality of films \( f_0 \), we estimate the exponential relation of network influence and box office. Since we consider the absolute influence, we do not take account of world box office change and only consider currency inflation.

For the actor-film connection, we have a contribution factor like author-paper model. The initial centrality for actor influence is set to 1. The expression is listed below.

\[
a_i = \sum_{j} \eta_j A_{ij} f_j + a_i^0
\]

<table>
<thead>
<tr>
<th>No.</th>
<th>Movie</th>
<th>Adjusted Sales((G_i))</th>
<th>Year</th>
<th>( f_i^0 = \log_{1.3} \left( \frac{G_i}{G_{\text{min}}} \right) + 1 )</th>
<th>rank(TSF=0.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Skyfall</td>
<td>315,702,900</td>
<td>2012</td>
<td>6.59</td>
<td>6.59</td>
</tr>
<tr>
<td>2</td>
<td>Quantum of Solace</td>
<td>195,804,500</td>
<td>2008</td>
<td>4.77</td>
<td>5.49</td>
</tr>
<tr>
<td>3</td>
<td>Die Another Day</td>
<td>230,326,700</td>
<td>2002</td>
<td>5.38</td>
<td>5.33</td>
</tr>
<tr>
<td>4</td>
<td>Casino Royale</td>
<td>212,329,500</td>
<td>2006</td>
<td>5.07</td>
<td>5.24</td>
</tr>
<tr>
<td></td>
<td>Movie Title</td>
<td>Box Office</td>
<td>Year</td>
<td>IMDB Rating</td>
<td>Audience Rating</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------</td>
<td>--------------</td>
<td>-------</td>
<td>-------------</td>
<td>----------------</td>
</tr>
<tr>
<td>5</td>
<td>Tomorrow Never Dies</td>
<td>224,708,400</td>
<td>1997</td>
<td>5.29</td>
<td>5.22</td>
</tr>
<tr>
<td>7</td>
<td>GoldenEye</td>
<td>203,772,900</td>
<td>1995</td>
<td>4.92</td>
<td>5.04</td>
</tr>
<tr>
<td>8</td>
<td>Licence to Kill</td>
<td>72,914,300</td>
<td>1989</td>
<td>1.00</td>
<td>2.62</td>
</tr>
</tbody>
</table>

From above analysis, we have noticed though *Quantum of Solace* falls behind *Die Another Day* in box office, but we should recognize its value by the success of *Skyfall*. The Quantum rank is higher than *Die Another Day*, which is coherent with the IMDB rank (6.7 and 6.1). It works the same way as to 4th and 5th films.

For actors, Pierce Brosnan and Daniel Craig get highest scores due to their starring as chief actor. The funny part is we can see that supporting roles M and Q in 007 series gain considerable influence through long term casting, with a score higher than some of the Bond’s Girls. This corresponds with most 007 fans’ recognition.

**Chasing the Future**

Models above only care about the current relationship and then accumulative influence. But if we consider time’s effect, we will have a historical view and may be able to predict the future.

Suppose we are evaluating paper \( i \)'s influence at time \( t \). The time it was published is \( t_i \). We considered time effects by weighting the citation. The more distant the citation is, the less it can contribute to the influence of \( P_j \) at time \( t \). In mathematical expression, we use \( \gamma_j(t) \) as the time factor of \( P_j \) at time \( t \) (by year):

\[
\gamma_j(t) = \begin{cases} 
  e^{-(t-t_j)} & t > t_j \\
  0 & t < t_j
\end{cases}
\]

Here, we have an exponentially diminishing influence with the time interval, which corresponds to the reality.

\[
x_j(t) = \alpha \sum_i \rho_i \mu_i \gamma_j'(t) A_{ij} \frac{x_j(t)}{d_j} + \ln[c_j(t) + 1]
\]
The adjacent matrix between papers should be revised into

\[ A(t)_{ij} = \rho_i \mu_j \gamma_{ij}(t) \]

Hence, instantaneous influence derived by Page-Rank algorithm is:

\[ \bar{X}(t) = (I - \alpha A(t)D(t)_{out}^{-1})^{-1} \bar{X}_0(t) = D(t)_{out}(D(t)_{out} - \alpha A(t))^{-1} \bar{X}_0(t) \]

Instantaneous authority and hub centrality is:

\[ \bar{X}(t) = (I - \alpha \beta A(t)A(t)^T)^{-1}(\alpha A(t)\bar{Y}_0(t) + \bar{X}_0(t)) \]

\[ \bar{Y}(t) = \beta A(t)^T \bar{X}(t) + \bar{Y}_0(t) = \beta A(t)^T(I - \alpha \beta A(t)A(t)^T)^{-1}(\alpha A(t)\bar{Y}_0(t) + \bar{X}_0(t)) + \bar{Y}_0(t) \]

By choosing observant time \( t \) as present time, we can get spot influence.
Similarly, we can compute the instantaneous authority and conductivity of author, journal, department and university.

**Trend Prediction**

In order to plot the historic development of network science, data about outer citation at 1995, 2000, 2002 and 2006 were collected from Google Scholar. Take three papers as example:

![Figure 12 Time Evolution of the Influence for the Paper](image)

We can also use this to determine which paper is presently influential by computing the instantaneous authority now. Take the top 3 (in instantaneous authority ranking) as example:

<table>
<thead>
<tr>
<th>Rank.</th>
<th>No.</th>
<th>Paper Names</th>
<th>2014</th>
<th>Accumulative</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>Emergence of scaling in random networks</td>
<td>16.70</td>
<td>39.47</td>
<td>1999</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>Collective dynamics of ’small-world’ networks</td>
<td>16.49</td>
<td>41.10</td>
<td>1998</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>On Random Graphs</td>
<td>12.94</td>
<td>27.82</td>
<td>1959</td>
</tr>
</tbody>
</table>

Paper 6’s authority exceeds Paper 5 in instantaneous authority ranking because it was more frequently cited by newly published papers.

**Decision Making**

In models above, indexes describing different types of influence are computed. For instance, authority describes the importance in and hub centrality and betweenness describes the ability to connect different parts of the network. But in terms of application, which type of index is reliable depends on the goal of our measurement. We divide the
goals into 3 types: Dabbling, Relationship Building and Innovation. Take academic circle as an example, they should refer to different indexes:

- **Dabbling**: For abecedarians of a field, like network science, a general survey is their primary concern. Hence, papers with high hub centrality will be very helpful because they tend to review more important works in this field.

- **Relationship Building**: For people who already have the basic knowledge of this field, they usually hope to build academic relationships with important scholars as soon as possible. In this case, consulting scholars with high betweenness is the optimal choice because they have more connections in the network and will possibly introduce you the right person to coauthor with.

- **Innovation**: If you are a scholar who expects to get some enlightening advice on future scientific research, you probably should communicate more with authoritative scholars.

## Evaluating our model

### Sensitivity Analysis

We analyze the sensitivity of our model with modified parameters.

- **Author-Based Model**

  In author-based model, $\alpha$ is artificially determined to describe the weight of neighbor nodes’ influence against its initial influence. Below are results with different $\alpha$ from 0.85 to 0.50.

<table>
<thead>
<tr>
<th>No.</th>
<th>Author Name</th>
<th>alpha=0.85</th>
<th>alpha=0.75</th>
<th>alpha=0.70</th>
<th>alpha=0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HARARY, FRANK*</td>
<td>34.758</td>
<td>19.343</td>
<td>15.496</td>
<td>7.663</td>
</tr>
<tr>
<td>2</td>
<td>ALON, NOGA M.</td>
<td>34.669</td>
<td>17.652</td>
<td>13.626</td>
<td>6.051</td>
</tr>
<tr>
<td>3</td>
<td>GRAHAM, RONALD LEWIS</td>
<td>30.736</td>
<td>15.950</td>
<td>12.400</td>
<td>5.607</td>
</tr>
<tr>
<td>4</td>
<td>BOLLOBAS, BELA</td>
<td>30.247</td>
<td>15.806</td>
<td>12.328</td>
<td>5.633</td>
</tr>
<tr>
<td>5</td>
<td>SOS, VERA TURAN</td>
<td>29.791</td>
<td>16.187</td>
<td>12.840</td>
<td>5.906</td>
</tr>
<tr>
<td>7</td>
<td>TUZA, ZSOLT</td>
<td>27.836</td>
<td>14.569</td>
<td>11.376</td>
<td>5.224</td>
</tr>
<tr>
<td>8</td>
<td>POMERANCE, CARL BERNARD</td>
<td>27.248</td>
<td>15.360</td>
<td>12.329</td>
<td>6.130</td>
</tr>
<tr>
<td>10</td>
<td>SPENCER, JOEL HAROLD</td>
<td>24.574</td>
<td>12.843</td>
<td>10.034</td>
<td>4.670</td>
</tr>
</tbody>
</table>

Higher $\alpha$ indicates more relevance to the network’s topology rather than initial value. The PR centrality decreases and becomes less differential as $\alpha$ becomes smaller because the neighbor nodes give less contribution. With the convergent condition $0 < \alpha < \lambda^{-1}$ also taken into consideration, our choice of 0.85 is reasonable. The results are shown as below:

- **Paper-Based Model**

  In the paper-based model, we tested different $\alpha$ from 0.85 to 0.50. It turns out although absolute value may be different but the relative rank is unchanged, i.e. the model is basically insensitive to $\alpha$. 

19 of 21 Pages
Strength and Weakness

Strength

- **Multiple Influence Index**: Our model contains several indexes indicating different aspects of the network, which means we can describe the influence of a certain node more thoroughly.
- **Network Split**: Split the semi-bipartite network into two networks. Split hierarchical network with similar pattern into several bipartite networks. Easy for understanding and practical use, still capable of describing the influence transmitted from the upper level.
- **Time-dependent**: Based on basic model, time dependent model is created to describe the dynamic relative influence, making it easy to plot the trend of network science research. With sufficient data, we can also use it to make reasonable predictions.
- **Compatible with big network data**: we offered a way to compensate information from outer network to improve the accuracy of small sample influence.

Weakness

- **Compromised information from upper level**: During the process of splitting hierarchical network, influence transmitted from the upper level may not be fully expressed.
- **Sample Relied**: The accuracy of the model depends on the time of sample paper. So, unrepresentative sample may lead to biased prediction.

Appendix

Table 3 The Paper Nnumber and its Name in Our Network Analysis (Ranked by Publication Year)

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>On Random Graphs</td>
</tr>
<tr>
<td>2</td>
<td>On properties of a well-known graph, or, What is your Ramsey number?</td>
</tr>
<tr>
<td>3</td>
<td>Power and Centrality: A family of measures</td>
</tr>
<tr>
<td>4</td>
<td>Social network thresholds in the diffusion of innovations</td>
</tr>
<tr>
<td>5</td>
<td>Collective dynamics of `small-world' networks</td>
</tr>
<tr>
<td>6</td>
<td>Emergence of scaling in random networks</td>
</tr>
<tr>
<td>7</td>
<td>Models of core/periphery structures</td>
</tr>
<tr>
<td>8</td>
<td>Navigation in a small world</td>
</tr>
<tr>
<td>9</td>
<td>Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality.</td>
</tr>
<tr>
<td>10</td>
<td>The structure of scientific collaboration networks</td>
</tr>
<tr>
<td>11</td>
<td>Statistical mechanics of complex networks</td>
</tr>
<tr>
<td>12</td>
<td>Identity and search in social networks</td>
</tr>
<tr>
<td>13</td>
<td>The structure and function of complex networks</td>
</tr>
<tr>
<td>14</td>
<td>Identifying sets of key players in a network</td>
</tr>
<tr>
<td>15</td>
<td>Networks, influence, and public opinion formation</td>
</tr>
<tr>
<td>16</td>
<td>Statistical models for social networks</td>
</tr>
</tbody>
</table>
Reference