Academic Network of UMSI Professors

Wei-Tzu Huang, Shuo, Yang, & Lei, Zhang
School of Information
University of Michigan
Ann Arbor, MI

Abstract—**It is common that many research articles are highly collaborated with at least two researchers. It is interesting to know which universities have a stronger collaborative relationship of UMSI and why they cooperate. This paper built an academic network based on the relationship between professors in UMSI and scholars from other organizations.**

Keywords: co-author; academic network; visualization

I. Introduction

In recent two decades, “co-authored article” is one of issues bibliometrics focusing on. A lot of studies found that the co-authored articles increased rapidly because “big science” becomes a research trend in the academia. The collaborations perform different among studies and disciplines. Ref. [1] mentioned a study adopting quantitative research methodology tended to be a co-authored article. Ref. [2] proposed that studies in psychology, management, and education had higher collaboration rate than literature and history. Collaboration in physics is the most common compared to other discipline. Many articles are highly collaborated with more than two thousand authors in specific article. Research organizations manager now may want to know whom his faculties collaborate with, and also think about how to allocate resource among faculties.

This study identified the collaboration relationship between UMSI faculties and researchers from other organizations. To get enough data of published article, we only focused on SI professors. The data was downloaded from Web of Science database, and extracted components we need. We downloaded totally 536 articles, and 142 of them were more than one author. In the analysis result, the visualizations show that the collaborations increase gradually. Some factors influence the collaborations such as geographic location.

II. Paper Evolution

In order to make project work smoothly, we change some steps from previous project plan. First, we wanted to download the data through API for Web of Science. When applying for an API account, the database asked for a regular IP address with authentication to access. However, the IP address changed each time we log into the UM network system. Hence, we changed to download data by ourselves. Second, we planned to identify each author SI faculties collaborated with. In fact, the database didn’t have complete author authority control. An author may have more than one form of name presenting in the database. For example, some authors showed their name on some of their articles with middle name initial, but some of articles didn’t. The program would identify these names as different authors. Therefore, we changed to present on university level, and to show universities only in the visualizations. The other adjustment was we would like to collect all the articles published by professors. In fact, the Web of Science database doesn’t index all journals and conference publications. Other sources like professors’ CV lack of co-authors’ institutions.

III. Data

A. Data source

Our data comes from Web of Science database published by Thomson Reuters.

B. Data collection

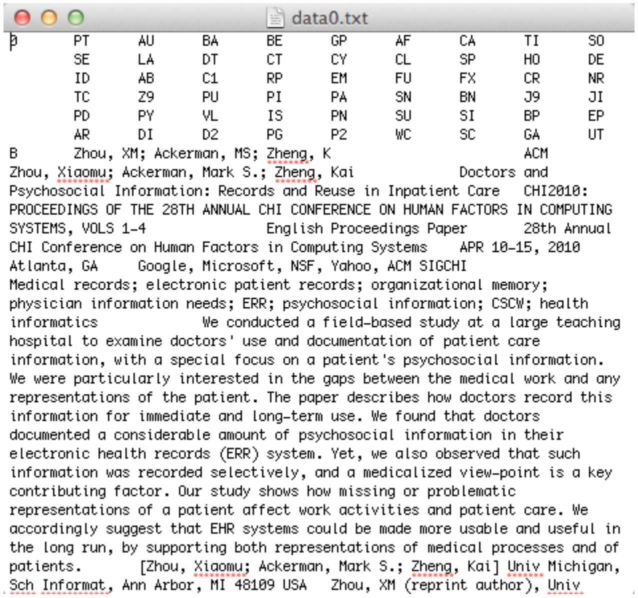
Our first step to the project is data collection. We made searching query for each professor. The queries tried to include all possible forms of the professor and exclude other authors with the same form. For example, professor Dragomir Radev’s searching query was “AU=(Radev, DR OR Radev, D NOT RADEV DD) AND SU=(COMPUTER SCIENCE OR ENGINEERING OR INFORMATION SCIENCE LIBRARY SCIENCE OR LINGUISTICS) AND OG=(UNIV MICHIGAN OR COLUMBIA UNIV) ”. In the author field, we included “Radev, DR”, which R is his middle name initial. Then, after we drag “Radev, D” into query, we needed to exclude “Radev, DD”, which was contained in “Radev, D” but not the professor we wanted to find. We set the subject and organization limitations to exclude persons in other area with the same name.

C. Data limitation

There were some limitations in our data and analysis. First, not all articles published by professors were indexed in the Web of Science database. The database only indexes journals that numbers of citation are over the threshold. Second, one article with only one author may have more than one address shown in the database. It’s because authors may have two positions in different organizations, and they provided all addresses in their articles. This kind of data still shows specific relationship between organizations.

D. Data composition

The original data crawled from database consists of 54 columns of various attributes of papers. The bibliographical data output is shown in Figure 1.

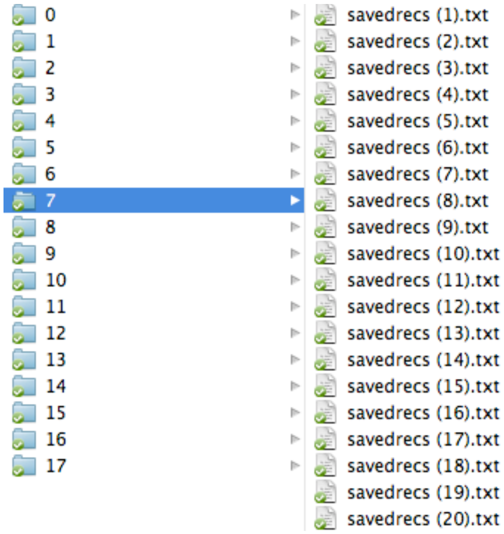


```
0        PT       AU       BA       BE       GP       AF       CA       TI       SO
         SE       LA       DT       CT       CY       CL       SP       HO       DE
         ID       AB       C1       RP       EM       FU       FX       CR       NR
         TC       Z9       PU       PI       PA       SN       BN       J9       JI
         PD       PY       VL       IS       PN       SU       SI       BP       EP
         AR       DI       D2       PG       P2       WC       SC       GA       UT
B        Zhou, XM; Ackerman, MS; Zheng, K                                 ACM
Zhou, Xiaomu; Ackerman, Mark S.; Zheng, Kai             Doctors and
Psychosocial Information: Records and Reuse in Inpatient Care    CHI2010:
PROCEEDINGS OF THE 28TH ANNUAL CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING
SYSTEMS, VOLS 1-4                English Proceedings Paper       28th Annual
CHI Conference on Human Factors in Computing Systems    APR 10-15, 2010
Atlanta, GA     Google, Microsoft, NSF, Yahoo, ACM SIGCHI
Medical records; electronic patient records; organizational memory;
physician information needs; ERR; psychosocial information; CSCW; health
informatics              We conducted a field-based study at a large teaching
hospital to examine doctors' use and documentation of patient care
information, with a special focus on a patient's psychosocial information.
We were particularly interested in the gaps between the medical work and any
representations of the patient. The paper describes how doctors record this
information for immediate and long-term use. We found that doctors
documented a considerable amount of psychosocial information in their
electronic health records (ERR) system. Yet, we also observed that such
information was recorded selectively, and a medicalized view-point is a key
contributing factor. Our study shows how missing or problematic
representations of a patient affect work activities and patient care. We
accordingly suggest that EHR systems could be made more usable and useful in
the long run, by supporting both representations of medical processes and of
patients.        [Zhou, Xiaomu; Ackerman, Mark S.; Zheng, Kai] Univ Michigan,
Sch Informat, Ann Arbor, MI 48109 USA   Zhou, XM (reprint author), Univ
```

Figure 1. Bibliographical data output

E. Data process

We searched 18 SI professors using the queries showed above, downloaded each of their papers as txt files to local directory. Each professor was assigned to his/her own directory, which is shown in Figure 2.



```
0     savedrecs (1).txt
1     savedrecs (2).txt
2     savedrecs (3).txt
3     savedrecs (4).txt
4     savedrecs (5).txt
5     savedrecs (6).txt
6     savedrecs (7).txt
7     savedrecs (8).txt
8     savedrecs (9).txt
9     savedrecs (10).txt
10    savedrecs (11).txt
11    savedrecs (12).txt
12    savedrecs (13).txt
13    savedrecs (14).txt
14    savedrecs (15).txt
15    savedrecs (16).txt
16    savedrecs (17).txt
17    savedrecs (18).txt
      savedrecs (19).txt
      savedrecs (20).txt
```

Figure 2. Intermediate data files

Then, we wrote the program named "project-0.py" to read data from these txt files and put them into one txt file "data0.txt".

After that, we selected all of the content from the txt file, pasted them into a spreadsheet, cut the columns we need, which are the "co-author" and the "year of publish", pasted these two columns into another txt file "data1.txt".

Next, we wrote the program named "project-1.py" to parse the data, reformat them in the way that could be read by the visualization software Gephi, write them into database, "txtTOgexf.db", see Figure 3.

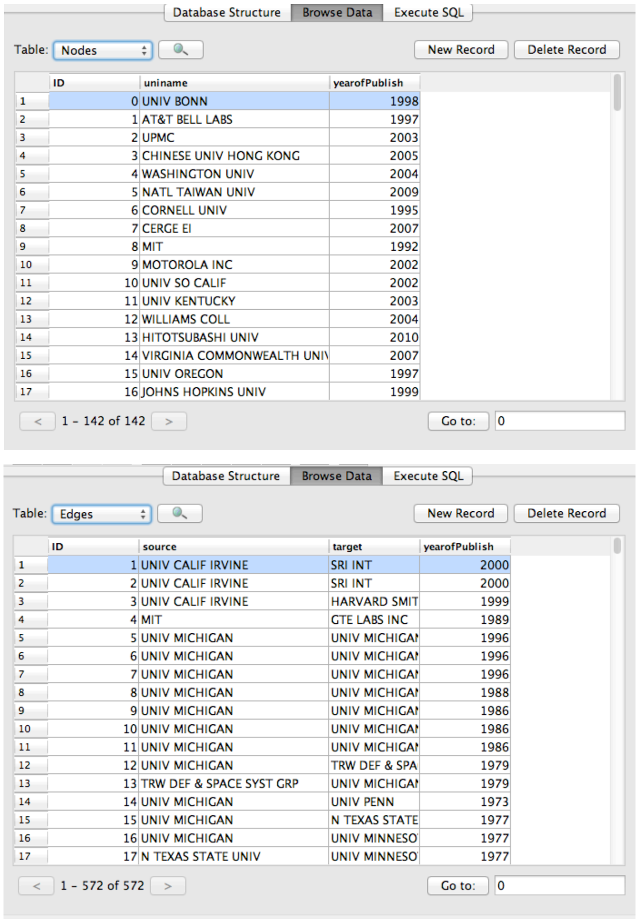


Table: Nodes

	ID	uniname	yearofPublish
1	0	UNIV BONN	1998
2	1	AT&T BELL LABS	1997
3	2	UPMC	2003
4	3	CHINESE UNIV HONG KONG	2005
5	4	WASHINGTON UNIV	2004
6	5	NATL TAIWAN UNIV	2009
7	6	CORNELL UNIV	1995
8	7	CERGE EI	2007
9	8	MIT	1992
10	9	MOTOROLA INC	2002
11	10	UNIV SO CALIF	2002
12	11	UNIV KENTUCKY	2003
13	12	WILLIAMS COLL	2004
14	13	HITOTSUBASHI UNIV	2010
15	14	VIRGINIA COMMONWEALTH UNIV	2007
16	15	UNIV OREGON	1997
17	16	JOHNS HOPKINS UNIV	1999

< 1 - 142 of 142 > Go to: 0

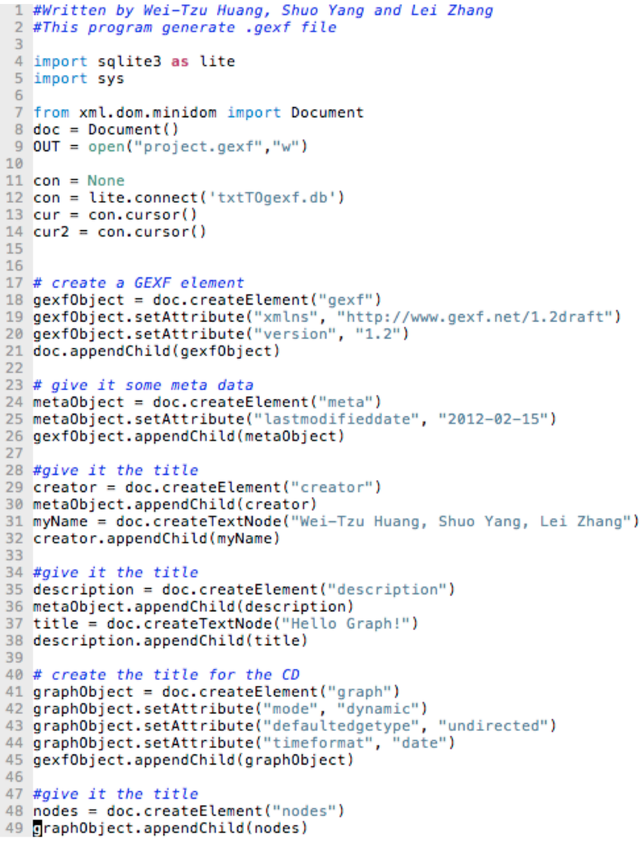
Table: Edges

	ID	source	target	yearofPublish
1	1	UNIV CALIF IRVINE	SRI INT	2000
2	2	UNIV CALIF IRVINE	SRI INT	2000
3	3	UNIV CALIF IRVINE	HARVARD SMIT	1999
4	4	MIT	GTE LABS INC	1989
5	5	UNIV MICHIGAN	UNIV MICHIGAN	1996
6	6	UNIV MICHIGAN	UNIV MICHIGAN	1996
7	7	UNIV MICHIGAN	UNIV MICHIGAN	1996
8	8	UNIV MICHIGAN	UNIV MICHIGAN	1988
9	9	UNIV MICHIGAN	UNIV MICHIGAN	1986
10	10	UNIV MICHIGAN	UNIV MICHIGAN	1986
11	11	UNIV MICHIGAN	UNIV MICHIGAN	1986
12	12	UNIV MICHIGAN	TRW DEF & SPA	1979
13	13	TRW DEF & SPACE SYST GRP	UNIV MICHIGAN	1979
14	14	UNIV MICHIGAN	UNIV PENN	1973
15	15	UNIV MICHIGAN	N TEXAS STATE	1977
16	16	UNIV MICHIGAN	UNIV MINNESO	1977
17	17	N TEXAS STATE UNIV	UNIV MINNESO	1977

< 1 - 572 of 572 > Go to: 0

Figure 3. Intermediate database

Finally, we wrote the program named "project-2.py" to retrieve data from database and generate a gexf file, which is readable by Gephi, see Figure 4.



```
#Written by Wei-Tzu Huang, Shuo Yang and Lei Zhang
#This program generate .gexf file

import sqlite3 as lite
import sys

from xml.dom.minidom import Document
doc = Document()
OUT = open("project.gexf","w")

con = None
con = lite.connect('txtTOgexf.db')
cur = con.cursor()
cur2 = con.cursor()


# create a GEXF element
gexfObject = doc.createElement("gexf")
gexfObject.setAttribute("xmlns", "http://www.gexf.net/1.2draft")
gexfObject.setAttribute("version", "1.2")
doc.appendChild(gexfObject)

# give it some meta data
metaObject = doc.createElement("meta")
metaObject.setAttribute("lastmodifieddate", "2012-02-15")
gexfObject.appendChild(metaObject)

#give it the title
creator = doc.createElement("creator")
metaObject.appendChild(creator)
myName = doc.createTextNode("Wei-Tzu Huang, Shuo Yang, Lei Zhang")
creator.appendChild(myName)

#give it the title
description = doc.createElement("description")
metaObject.appendChild(description)
title = doc.createTextNode("Hello Graph!")
description.appendChild(title)

# create the title for the CD
graphObject = doc.createElement("graph")
graphObject.setAttribute("mode", "dynamic")
graphObject.setAttribute("defaultedgetype", "undirected")
graphObject.setAttribute("timeformat", "date")
gexfObject.appendChild(graphObject)

#give it the title
nodes = doc.createElement("nodes")
graphObject.appendChild(nodes)
```

Figure 4. Part of the “project.gexf”

IV. VISUALIZATION AND OBSERVATION

Network in Figure 5 consists of 142 nodes and 340 edges. The nodes represent the organizations of people who have collaborated relationship with UMSI professors or the universities where UMSI professors have ever worked. We use Force Atlas and Noverlap layout to display the network. The size of nodes shows degree values; the color of nodes, ranging from red to yellow and then to blue, shows page rank values; the size of edges shows weight values (weight based on times of collaboration between the two professors).

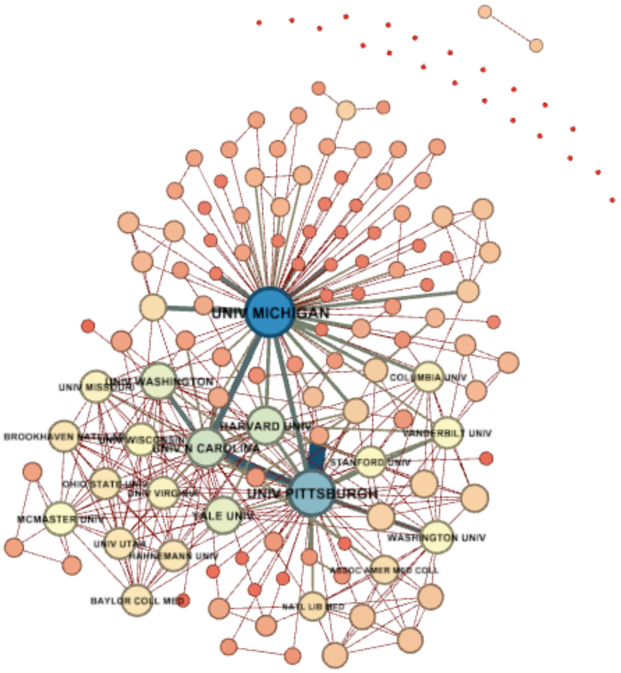


Figure 5. Academia Social Network of UMSI Professors

Because most of nodes have very similar lower degree value and page rank value comparing to UM, we use the following shape of spline to make the size and color between nodes and edges become more apparent.

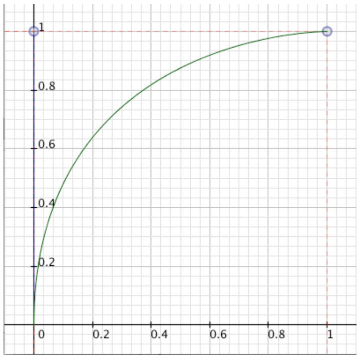


Figure 6. Shape of spline

In Figure 5, the network consists of three parts – the largest component including 129 nodes, another including 21 single nodes and the other including two nodes and an edge between them. In the second part, there are several nodes without any edge. They are the organizations where UMSI professors worked and published articles independently before. And in the third part, the two nodes are Internet2 and World Wide Web Consortium, which are weakly related to UM. Therefore, our project analysis focuses on the first part. In the largest component, University of Michigan is the center school absolutely, because the network is based on the relationship between professors in UMSI and scholars from other organizations. Besides of UM, University of Pittsburgh (U of Pitts) and University of North Carolina (UNC) play important roles in the network. These three nodes form a strong triangle that shows the frequent communication among them. We can find the collaborative relationships between UM and U of Pitts, and UM and UNC are the strongest. Both U of Pitts and UNC have Information school, and it might be the reason to the relationships among schools.

If we focus on the nodes that directly connect with UM, we can find the strongest collaborative relationship (See Figure 7).

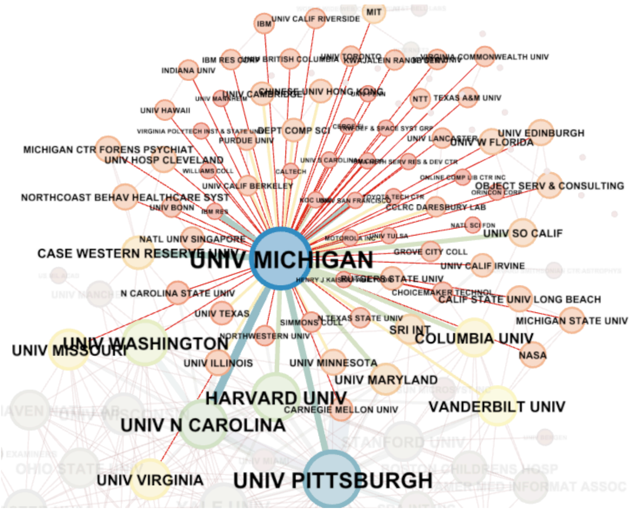


Figure 7. Partition of network centered on UM

There are more than 70 nodes connecting with University of Michigan directly in the graph. It can be interpreted that UMSI is a school that is very open to have external communication and also have high reputation to get many opportunities of collaboration with famous academic organizations.

According to the thickness and color of edges connecting to UM, Case Western Reserve University, University of North Carolina, IBM RES and University of Pittsburgh are the top 4 organizations that cooperate with UM the most times. Also, compared to other nodes that are linked by red thin edges, Harvard University, University of Maryland, Columbia University and University of Southern California also have many people who have been co-authors with UMSI professors in publications. So, this phenomenon shows that these universities have many common research interests, which contributed their cooperation. However, there is some difference between these universities. For example, although IBM RES and University of Pittsburgh have similar thick edges connecting with UM, the size of the former is much smaller than the latter one and the color is also much closer to red. These phenomena mean that IBM RES has very rare cooperation with other organizations within the network. In contrast, University of Pittsburgh is an important hub that has higher degree and page rank values in the network. So, it is reasonable to conclude that organizations like University of Pittsburgh are more likely to communicate with different organizations, so that it can get more encyclopedic and abundant research resources, and then provide these resources to research cooperating with UMSI professors. On the other hand, organizations like IBM RES may offer more focused and professional resources to improve UMSI professors' research.

We use the weight of edges as criteria to filter out universities having weak cooperated relationship with UM. Besides UM, only nine organizations are included in this partition of the original network (See Figure 8). There is an interesting node, Carnegie Mellon University (CMU) in the network shown in Figure 8. The edge between CMU and U of Pitts is the strongest one in the whole network; however, there's no edge between UM and CMU. This collaborative relationship shows difference in the research interests between two schools. According to the CMU and U of Pitts' strong collaboration, we found that location may influence collaboration between schools. U of Pitts and CMU locate in the same state (Penn State) that provides them more opportunities to communicate with each other. Scholars in each university can visit the other one easily. In contrast, the longer geographical distance between UM and CMU make their cooperation less convenient. We put the nine universities on Google map to analyze their location relationship.

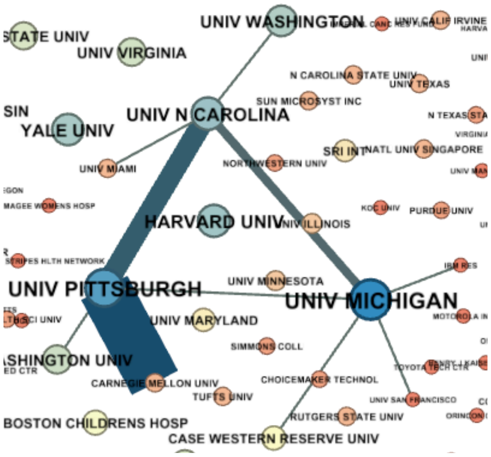


Figure 8. Network of organizations collabrated with UMSI more than three times

In Figure 8, besides UM, only nine organizations are included in this partition of the original network. Then, we put them on Google map to analyze their locations.

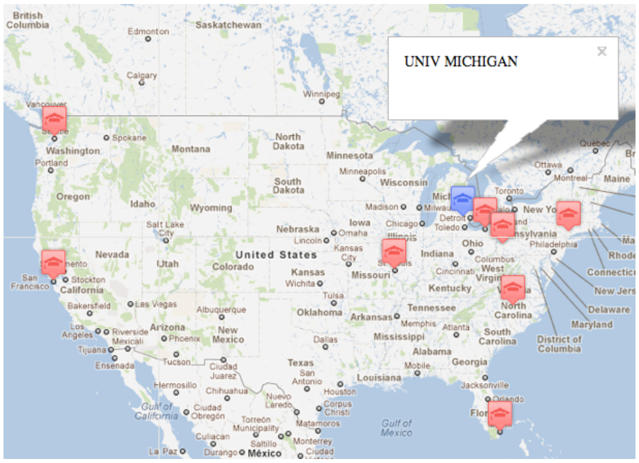


Figure 9. Location Distributions of Universities on Map

(University of Pittsburgh and Carnegie Mellon University are very close to each other, so that their icons overlap on the map.)

Although there are many famous universities in the west, most of organizations that have strong relationship with UMSI locate in the east where is closer to UM. This confirms the idea that location is an important factor to cooperation.

In Figure 10, we can find that the slope of trend line between year 1997 and 2006 is larger than 1, which means that the increasing of nodes became faster and faster during this period. This phenomenon shows that average annual amount of academic cooperation was rapidly growing. During this period, information technology developed very fast, which provided many interesting research topic to professors. However, after this period, the increasing rate becomes slower. Economic crisis appeared at that time, which may be the reason of it. Professors did not have as much funding of research as before during this time, so that the cooperation chance reduced along with it.

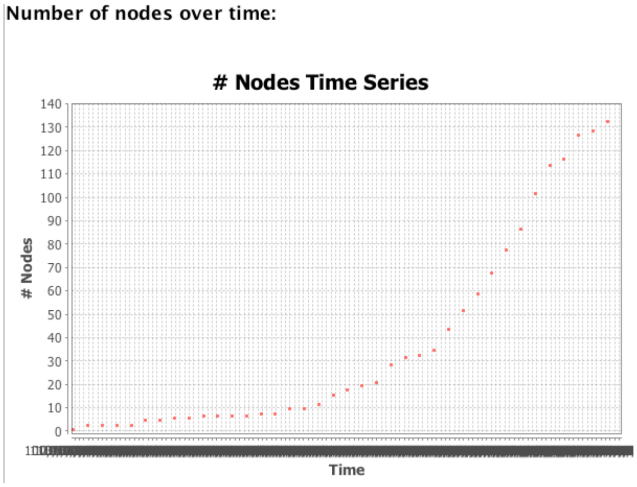


Figure 10. Timeline of Nodes

According to Figure 11, there is increasing development of edges after year 1997. This might be caused by the academic trend of "big science". The "beg science" means the collaborations are cross disciplines.

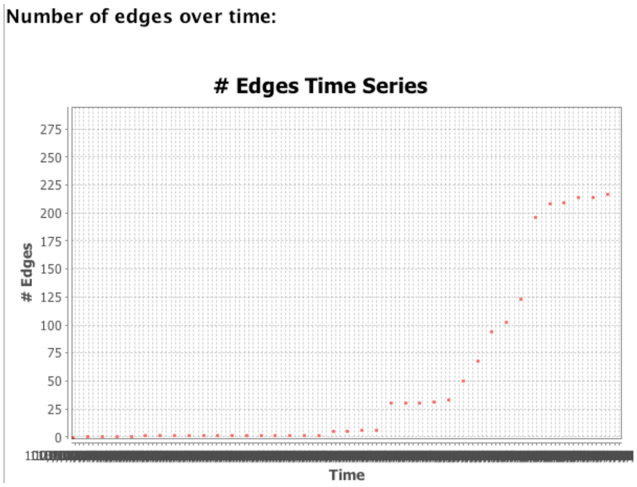


Figure 11. Timeline of Edges

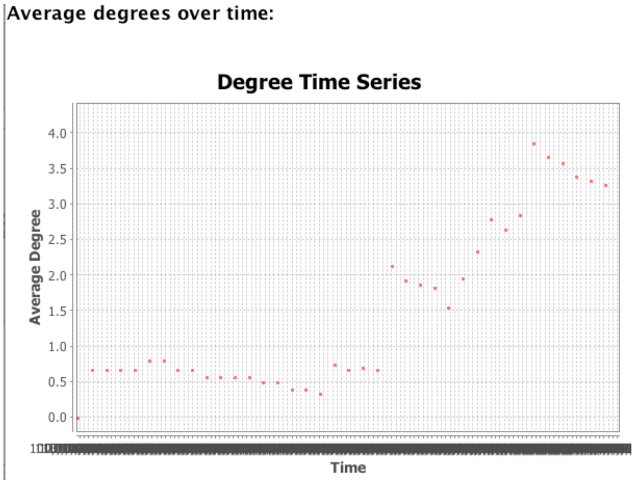


Figure 12. Timeline of Average Degree

The Figure 12 presents the average degree. Overall, the average degree increased in the past years. It means the edge strengths among schools increase larger than nodes.

V. CONCLUSION

There are two major factors that affect the co-author relationship – common research interest and geographical distance. The universities that have information schools tend to have stronger collaborative relationships with UMSI and universities tend to have stronger cooperation with the universities that are closer to them. Besides, some changes of objective conditions, such as economic environment and technology development have also significant influence on the trend of academic collaboration.

REFERENCE

[1] Moody, J., "The structure of a social science collaboration network: disciplinary cohesion from 1963-1999," American Soc Review, vol. 69, pp. 213-239, April 2004.

[2] Lariviere, V., Gingras, Y., & Archambault, E., "Canadian collaboration networks: a comparative analysis of the natural sciences, social sciences and the humamities", Scientometrics, vol. 68, pp. 519-533, July 2006.