Social Computing and Big Data Analytics
(社群運算大數據分析)

Time: 2016/04/18 (14:00-16:00)
Place: China Airlines

Min-Yuh Day
戴敏育
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2016-04-18
戴敏育博士
(Min-Yuh Day, Ph.D.)
淡江大學資管系專任助理教授
中央研究院資訊科學研究所訪問學人
國立台灣大學資訊管理博士

Publications Co-Chairs, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013-)
Program Co-Chair, IEEE International Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2012-)
Workshop Chair, The IEEE International Conference on Information Reuse and Integration (IEEE IRI)
Outline

• Social Computing and Big Data Analytics
• Analyzing the Social Web: Social Network Analysis
Business Insights with Social Analytics
Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine

Big Data Analytics and Data Mining
Stephan Kudyba (2014),
Big Data, Mining, and Analytics:
Components of Strategic Decision Making, Auerbach Publications

Source: http://www.amazon.com/gp/product/1466568704
Architecture of Big Data Analytics

Big Data Sources
- * Internal
- * External
- * Multiple formats
- * Multiple locations
- * Multiple applications

Big Data Transformation
- Raw Data
- Middleware
- Extract Transform Load
- Data Warehouse
- Traditional Format CSV, Tables
- Transformed Data

Big Data Platforms & Tools
- Hadoop
- MapReduce
- Pig
- Hive
- Jaql
- Zookeeper
- Hbase
- Cassandra
- Oozie
- Avro
- Mahout
- Others

Big Data Analytics Applications
- Queries
- Reports
- OLAP
- Data Mining

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools
- OLAP
- Queries
- Reports
- Data Mining

Big Data Analytics Applications

Data Mining
Big Data Analytics Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Social Big Data Mining

(Hiroshi Ishikawa, 2015)

Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies

- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Analysts

- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Analyzing the Social Web: Social Network Analysis
Social Network Analysis (SNA)
Social Network Analysis

• **A social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest.

• **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior.

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Social Network Analysis (SNA)

Centrality

Prestige
Degree

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Degree

A: 2  B: 4  C: 2  D: 1  E: 1

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Density

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Density

Edges (Links): 5
Total Possible Edges: 10
Density: $\frac{5}{10} = 0.5$

Source: https://www.youtube.com/watch?v=89mxOdPFxA
Nodes (n): 10
Edges (Links): 13
Total Possible Edges: \(\frac{n \times (n-1)}{2} = \frac{10 \times 9}{2} = 45\)
Density: \(\frac{13}{45} = 0.29\)
Which Node is Most Important?
Centrality

- Important or prominent actors are those that are linked or involved with other actors extensively.
- A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts.
- The links can also be called ties. A central actor is one involved in many ties.

Social Network Analysis (SNA)

- Degree Centrality
- Betweenness Centrality
- Closeness Centrality
Degree Centrality
Social Network Analysis: Degree Centrality
Social Network Analysis: Degree Centrality

<table>
<thead>
<tr>
<th>Node</th>
<th>Score</th>
<th>Standardized Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>5/10 = 0.5</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>4/10 = 0.4</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1/10 = 0.1</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1/10 = 0.1</td>
</tr>
</tbody>
</table>
Betweenness Centrality
Betweenness centrality:

Connectivity

Number of shortest paths going through the actor
Betweenness Centrality

\[ C_B(i) = \sum_{j<k} g_{ik}(i) / g_{jk} \]

Where \( g_{jk} \) = the number of shortest paths connecting \( jk \)
\( g_{jk}(i) \) = the number that actor \( i \) is on.

Normalized Betweenness Centrality

\[ C'_B(i) = C_B(i) / \left[ (n-1)(n-2)/2 \right] \]

Number of pairs of vertices excluding the vertex itself

Source: https://www.youtube.com/watch?v=RXohUeNCJiU
Betweenness Centrality

A: Betweenness Centrality = 0

A → C: 0/1 = 0
B → D: 0/1 = 0
B → E: 0/1 = 0
C → D: 0/1 = 0
C → E: 0/1 = 0
D → E: 0/1 = 0

Total: 0
Betweenness Centrality

B: Betweenness Centrality = 5

A→C: 0/1 = 0
A→D: 1/1 = 1
A→E: 1/1 = 1
C→D: 1/1 = 1
C→E: 1/1 = 1
D→E: 1/1 = 1

Total: 5
Betweenness Centrality

C:
A → B: 0/1 = 0
A → D: 0/1 = 0
A → E: 0/1 = 0
B → D: 0/1 = 0
B → E: 0/1 = 0
D → E: 0/1 = 0

Total: 0

C: Betweenness Centrality = 0
Betweenness Centrality

A: 0  
B: 5  
C: 0  
D: 0  
E: 0
Which Node is Most Important?
Which Node is Most Important?
Betweenness Centrality

\[ C_B(i) = \sum_{j<k} \frac{g_{ik}(i)}{g_{jk}} \]
Betweenness Centrality

A: Betweenness Centrality = 0

A: 
B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0
Closeness
Centrality
Social Network Analysis: Closeness Centrality

C: Closeness Centrality = 15/9 = 1.67

C→A: 1
C→B: 1
C→D: 1
C→E: 1
C→F: 2
C→G: 1
C→H: 2
C→I: 3
C→J: 3
Total=15
Social Network Analysis: Closeness Centrality

G: Closeness Centrality = 14/9 = 1.56
Social Network Analysis: Closeness Centrality

H: Closeness Centrality = $\frac{17}{9} = 1.89$
Social Network Analysis: Closeness Centrality

G: Closeness Centrality = 14/9 = 1.56

C: Closeness Centrality = 15/9 = 1.67

H: Closeness Centrality = 17/9 = 1.89
Social Network Analysis (SNA) Tools

• NetworkX
• igraph
• Gephi
• UCINet
• Pajek
Gephi
The Open Graph Viz Platform

Source: https://gephi.org/
Application of SNA

Social Network Analysis of Research Collaboration in Information Reuse and Integration

Example of SNA Data Source

IRI 2010: Las Vegas, NV, USA


Reda Alhajj, James B. D. Joshi, Mei-Ling Shyu: Message from Program Co-Chairs. 1

Stuart Harvey Rubin, Shu-Ching Chen: Forward. 1

Lotfi A. Zadeh: Precisiation of meaning - toward computation with natural language. 1-4

Reda Alhajj, Shu-Ching Chen, Gongzhu Hu, James B. D. Joshi, Gordon K. Lee, Stuart Harvey Rubin, Mei-Ling Shyu, Lotfi A. Zadeh: Panel title: Critical need for funding of basic and applied research in large-scale computing. 1

Automation, Integration and Reuse across Various Apps

László István Etesi, André Csillaghy, Lin-Ching Chang: A message-based interoperability framework with application to astrophysics. 1-6

Awny Alnusair, Tian Zhao, Eric Bodden: Effective API navigation and reuse. 7-12

Manabu Ohta, Ryohei Inoue, Atsuhiro Takasu: Empirical evaluation of active sampling for CRF-based analysis of pages. 13-18

Qunzhi Zhou, Viktor K. Prasanna: Workflow management of simulation based computation processes in transportation domain. 19-24

Source: http://www.informatik.uni-trier.de/~ley/db/conf/iri/iri2010.html
Research Question

• RQ1: What are the scientific collaboration patterns in the IRI research community?

• RQ2: Who are the prominent researchers in the IRI community?

Methodology

• Developed a simple web focused crawler program to download literature information about all IRI papers published between 2003 and 2010 from IEEE Xplore and DBLP.
  – 767 paper
  – 1599 distinct author

• Developed a program to convert the list of coauthors into the format of a network file which can be readable by social network analysis software.

• UCINet and Pajek were used in this study for the social network analysis.

Top10 prolific authors (IRI 2003-2010)

1. Stuart Harvey Rubin
2. Taghi M. Khoshgoftaar
3. Shu-Ching Chen
4. Mei-Ling Shyu
5. Mohamed E. Fayad
6. Reda Alhajj
7. Du Zhang
8. Wen-Lian Hsu
9. Jason Van Hulse
10. Min-Yuh Day

Data Analysis and Discussion

• **Closeness Centrality**
  – Collaborated widely

• **Betweenness Centrality**
  – Collaborated diversely

• **Degree Centrality**
  – Collaborated frequently

• **Visualization of Social Network Analysis**
  – Insight into the structural characteristics of research collaboration networks

### Top 20 authors with the highest *closeness* scores

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Closeness</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.024675</td>
<td>Shu-Ching Chen</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.022830</td>
<td>Stuart Harvey Rubin</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.022207</td>
<td>Mei-Ling Shyu</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.020013</td>
<td>Reda Alhajj</td>
</tr>
<tr>
<td>5</td>
<td>61</td>
<td>0.019700</td>
<td>Na Zhao</td>
</tr>
<tr>
<td>6</td>
<td>260</td>
<td>0.018936</td>
<td>Min Chen</td>
</tr>
<tr>
<td>7</td>
<td>151</td>
<td>0.018230</td>
<td>Gordon K. Lee</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>0.017962</td>
<td>Chengcui Zhang</td>
</tr>
<tr>
<td>9</td>
<td>1043</td>
<td>0.017962</td>
<td>Isai Michel Lombera</td>
</tr>
<tr>
<td>10</td>
<td>1027</td>
<td>0.017962</td>
<td>Michael Armella</td>
</tr>
<tr>
<td>11</td>
<td>443</td>
<td>0.017448</td>
<td>James B. Law</td>
</tr>
<tr>
<td>12</td>
<td>157</td>
<td>0.017082</td>
<td>Keqi Zhang</td>
</tr>
<tr>
<td>13</td>
<td>253</td>
<td>0.016731</td>
<td>Shahid Hamid</td>
</tr>
<tr>
<td>14</td>
<td>1038</td>
<td>0.016618</td>
<td>Walter Z. Tang</td>
</tr>
<tr>
<td>15</td>
<td>959</td>
<td>0.016285</td>
<td>Chengjun Zhan</td>
</tr>
<tr>
<td>16</td>
<td>957</td>
<td>0.016285</td>
<td>Lin Luo</td>
</tr>
<tr>
<td>17</td>
<td>956</td>
<td>0.016285</td>
<td>Guo Chen</td>
</tr>
<tr>
<td>18</td>
<td>955</td>
<td>0.016285</td>
<td>Xin Huang</td>
</tr>
<tr>
<td>19</td>
<td>943</td>
<td>0.016285</td>
<td>Sneh Gulati</td>
</tr>
<tr>
<td>20</td>
<td>960</td>
<td>0.016071</td>
<td>Sheng-Tun Li</td>
</tr>
</tbody>
</table>

### Top 20 authors with the highest betweenness scores

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Betweenness</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.000752</td>
<td>Stuart Harvey Rubin</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.000741</td>
<td>Shu-Ching Chen</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.000406</td>
<td>Taghi M. Khoshgoftaar</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
<td>0.000385</td>
<td>Xingquan Zhu</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.000376</td>
<td>Mei-Ling Shyu</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.000296</td>
<td>Reda Alhajj</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>0.000256</td>
<td>Xindong Wu</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>0.000194</td>
<td>Chengcui Zhang</td>
</tr>
<tr>
<td>9</td>
<td>39</td>
<td>0.000185</td>
<td>Wei Dai</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>0.000107</td>
<td>Narayan C. Debnath</td>
</tr>
<tr>
<td>11</td>
<td>31</td>
<td>0.000094</td>
<td>Qianhui Althea Liang</td>
</tr>
<tr>
<td>12</td>
<td>151</td>
<td>0.000094</td>
<td>Gordon K. Lee</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>0.000085</td>
<td>Du Zhang</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
<td>0.000072</td>
<td>Baowen Xu</td>
</tr>
<tr>
<td>15</td>
<td>41</td>
<td>0.000067</td>
<td>Hongji Yang</td>
</tr>
<tr>
<td>16</td>
<td>270</td>
<td>0.000060</td>
<td>Zhiwei Xu</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>0.000043</td>
<td>Mohamed E. Fayad</td>
</tr>
<tr>
<td>18</td>
<td>110</td>
<td>0.000042</td>
<td>Abhijit S. Pandya</td>
</tr>
<tr>
<td>19</td>
<td>106</td>
<td>0.000042</td>
<td>Sam Hsu</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>0.000042</td>
<td>Wen-Lian Hsu</td>
</tr>
</tbody>
</table>

Top 20 authors with the highest degree scores

<table>
<thead>
<tr>
<th>Rank</th>
<th>ID</th>
<th>Degree</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.035044</td>
<td>Shu-Ching Chen</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.034418</td>
<td>Stuart Harvey Rubin</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.030663</td>
<td>Taghi M. Khoshgoftaar</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.028786</td>
<td>Reda Alhajj</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.028786</td>
<td>Wen-Lian Hsu</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>0.024406</td>
<td>Min-Yuh Day</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.022528</td>
<td>Mei-Ling Shyu</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>0.021277</td>
<td>Richard Tzong-Han Tsai</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>0.017522</td>
<td>Eduardo Santana de Almeida</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>0.017522</td>
<td>Roumen Kountchev</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>0.016896</td>
<td>Hong-Jie Dai</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>0.015645</td>
<td>Narayan C. Debnath</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>0.015019</td>
<td>Jason Van Hulse</td>
</tr>
<tr>
<td>14</td>
<td>25</td>
<td>0.013767</td>
<td>Roumian Kountcheva</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>0.013141</td>
<td>Silvio Romero de Lemos Meira</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>0.013141</td>
<td>Vladimir Todorov</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>0.013141</td>
<td>Mariofanna G. Milanova</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>0.013141</td>
<td>Mohamed E. Fayad</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>0.012516</td>
<td>Chengcui Zhang</td>
</tr>
<tr>
<td>20</td>
<td>18</td>
<td>0.011890</td>
<td>Waleed W. Smari</td>
</tr>
</tbody>
</table>

Visualization of IRI (IEEE IRI 2003-2010)
co-authorship network (global view)

Visualization of Social Network Analysis

Visualization of Social Network Analysis

Visualization of Social Network Analysis

Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"
Summary

• Social Computing and Big Data Analytics
• Analyzing the Social Web: Social Network Analysis
References

• Jiawei Han and Micheline Kamber (2011),
• Jennifer Golbeck (2013),
  Analyzing the Social Web, Morgan Kaufmann
• Stephan Kudyba (2014),
  Big Data, Mining, and Analytics: Components of Strategic
  Decision Making, Auerbach Publications
• Hiroshi Ishikawa (2015),
  Social Big Data Mining, CRC Press
Q & A

Social Computing and Big Data Analytics
(社群運算大數據分析)

Time: 2016/04/18 (14:00-16:00)
Place: China Airlines

Min-Yuh Day
戴敏育
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2016-04-18