Data Mining
資料探勘

社會網路分析、意見分析
(Social Network Analysis, Opinion Mining)

1002DM06
MI4
Thu. 9,10 (16:10-18:00) B513

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail. tku.edu.tw/myday/
2012-05-10

課程大綱（Syllabus)

週次　日期　內容（Subject/Topics）

1　101/02/16　資料探勘導論 (Introduction to Data Mining)

2　101/02/23　關連分析 (Association Analysis)

3　101/03/01　分類與預測 (Classification and Prediction)

4　101/03/08　分群分析 (Cluster Analysis)

5　101/03/15　個案分析與實作一 (分群分析)：
Banking Segmentation (Cluster Analysis – KMeans)

6　101/03/22　個案分析與實作二 (關連分析)：
Web Site Usage Associations (Association Analysis)

7　101/03/29　期中報告 (Midterm Presentation)

8　101/04/05　教學行政觀摩日 (--No Class--)

課程大綱 (Syllabus)

週次　日期　內容（Subject/Topics）

9　101/04/12　個案分析與實作三 (決策樹、模型評估)：
Enrollment Management Case Study
(Decision Tree, Model Evaluation)

10　101/04/19　期中考試週

11　101/04/26　個案分析與實作四 (迴歸分析、類神經網路)：
Credit Risk Case Study (Regression Analysis,
Artificial Neural Network)

12　101/05/03　文字探勘與網頁探勘 (Text and Web Mining)

13　101/05/10　社會網路分析、意見分析
(Social Network Analysis, Opinion Mining)

14　101/05/17　期末專題報告 (Term Project Presentation)

15　101/05/24　畢業考試週

Outline

- Social Network Analysis
- Opinion Mining

Social Network Analysis

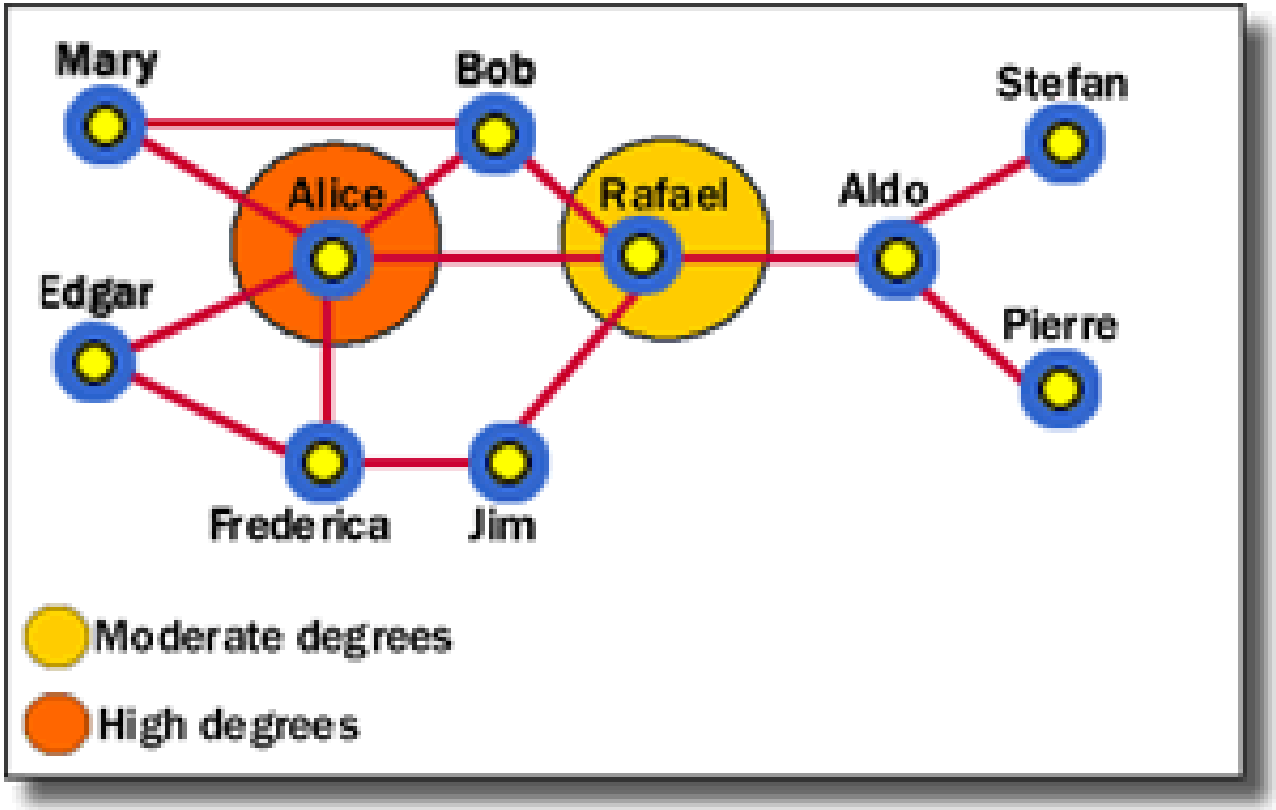
- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest
- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

Social Network Analysis

- Using Social Network Analysis, you can get answers to questions like:
  - How highly connected is an entity within a network?
  - What is an entity's overall importance in a network?
  - How central is an entity within a network?
  - How does information flow within a network?

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

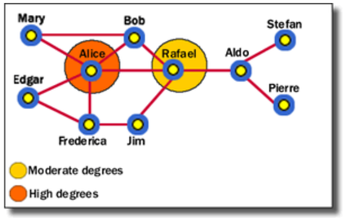
Social Network Analysis: Degree Centrality



Alice has the highest degree centrality, which means that she is quite active in the network. However, she is not necessarily the most powerful person because she is only directly connected within one degree to people in her clique—she has to go through Rafael to get to other cliques.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

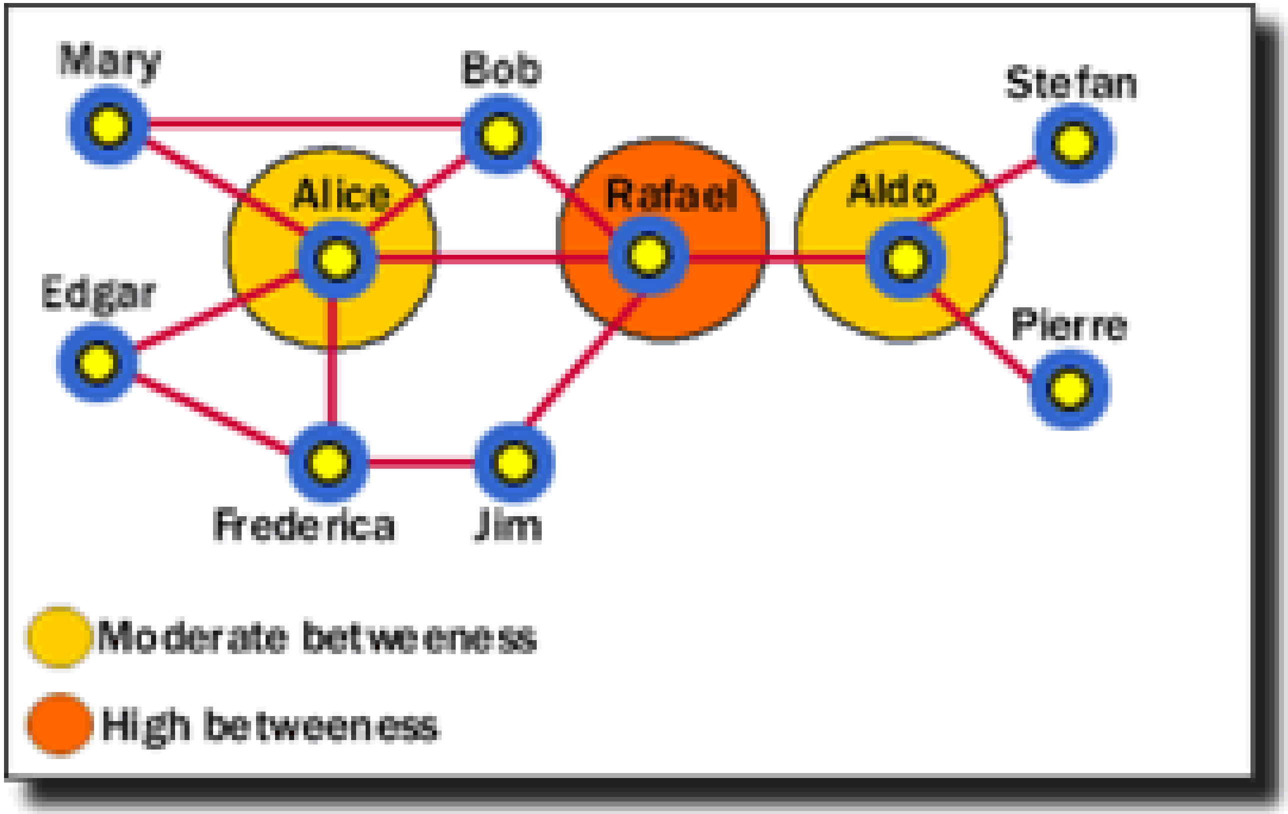
Social Network Analysis: Degree Centrality



- Degree centrality is simply the number of direct relationships that an entity has.
- An entity with high degree centrality:
  - Is generally an active player in the network.
  - Is often a connector or hub in the network.
  - s not necessarily the most connected entity in the network (an entity may have a large number of relationships, the majority of which point to low-level entities).
  - May be in an advantaged position in the network.
  - May have alternative avenues to satisfy organizational needs, and consequently may be less dependent on other individuals.
  - Can often be identified as third parties or deal makers.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

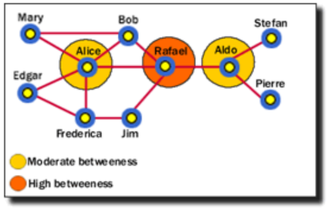
Social Network Analysis: Betweenness Centrality



Rafael has the highest betweenness because he is between Alice and Aldo, who are between other entities. Alice and Aldo have a slightly lower betweenness because they are essentially only between their own cliques. Therefore, although Alice has a higher degree centrality, Rafael has more importance in the network in certain respects.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

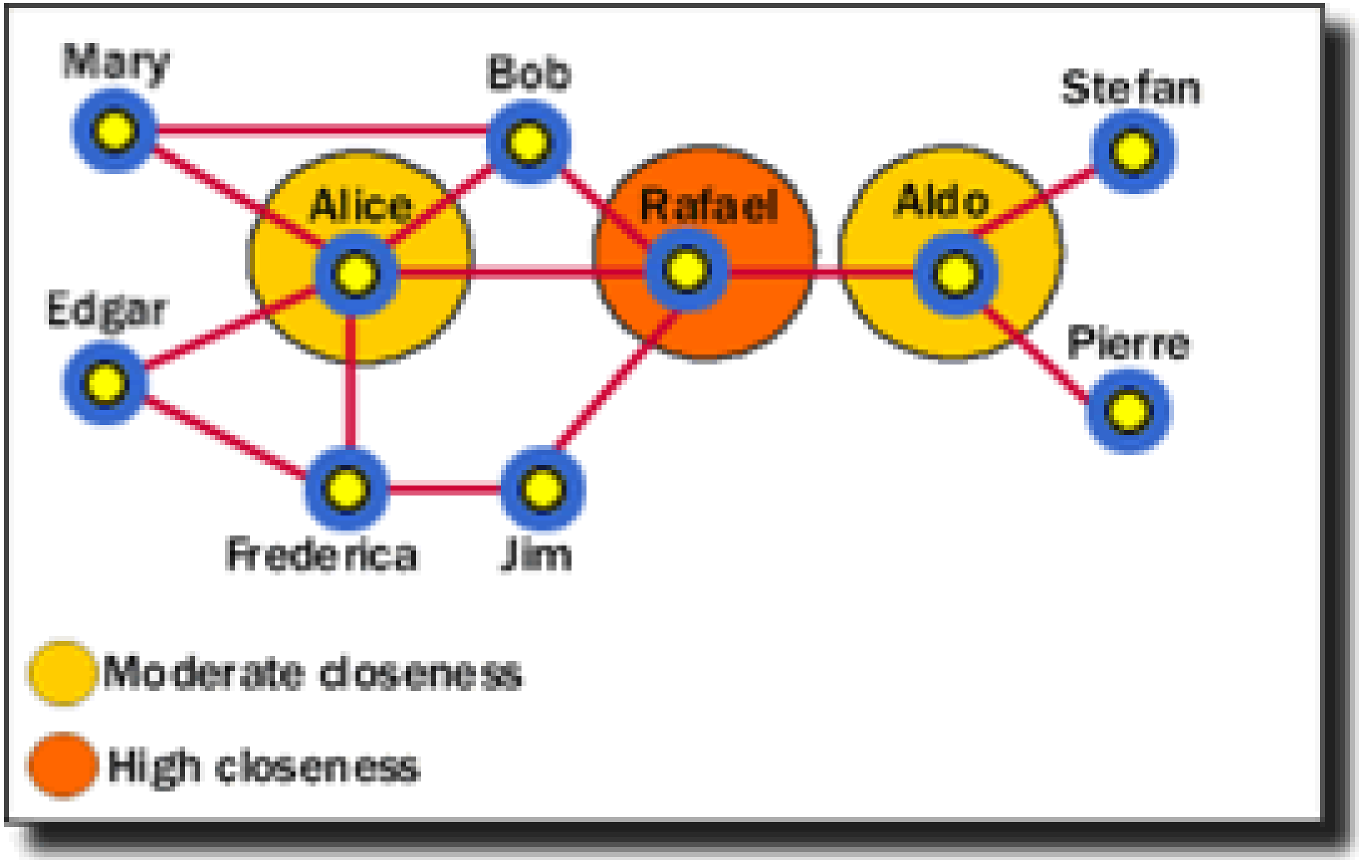
Social Network Analysis: Betweenness Centrality



- Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network.
- An entity with a high betweenness centrality generally:
  - Holds a favored or powerful position in the network.
  - Represents a single point of failure—take the single betweenness spanner out of a network and you sever ties between cliques.
  - Has a greater amount of influence over what happens in a network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

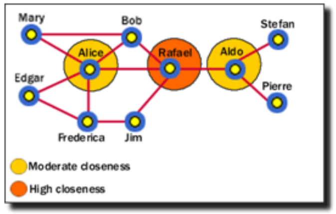
Social Network Analysis: Closeness Centrality



Rafael has the highest closeness centrality because he can reach more entities through shorter paths. As such, Rafael's placement allows him to connect to entities in his own clique, and to entities that span cliques.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

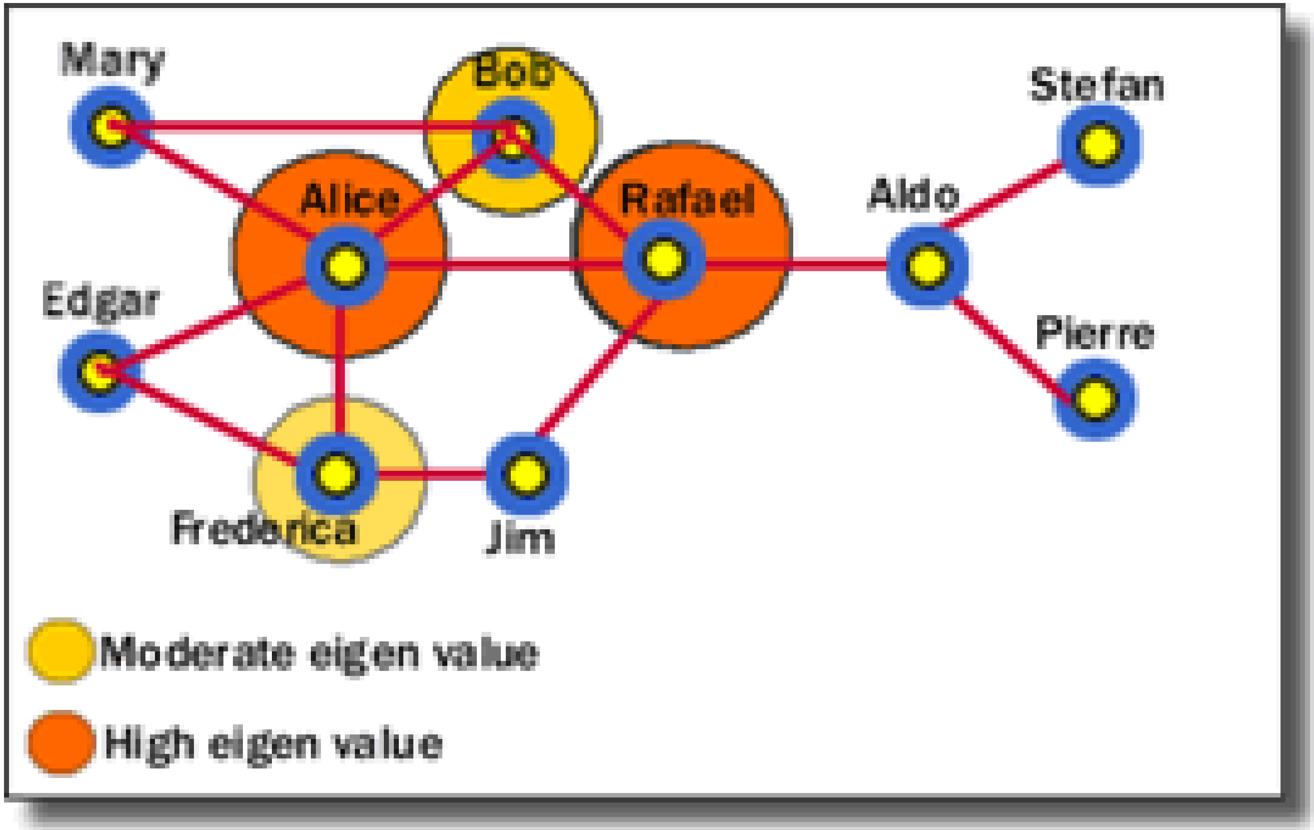
Social Network Analysis: Closeness Centrality



- Closeness centrality measures how quickly an entity can access more entities in a network.
- An entity with a high closeness centrality generally:
  - Has quick access to other entities in a network.
  - Has a short path to other entities.
  - Is close to other entities.
  - Has high visibility as to what is happening in the network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

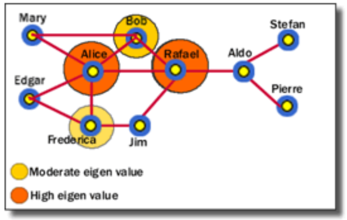
Social Network Analysis: Eigenvalue



Alice and Rafael are closer to other highly close entities in the network. Bob and Frederica are also highly close, but to a lesser value.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

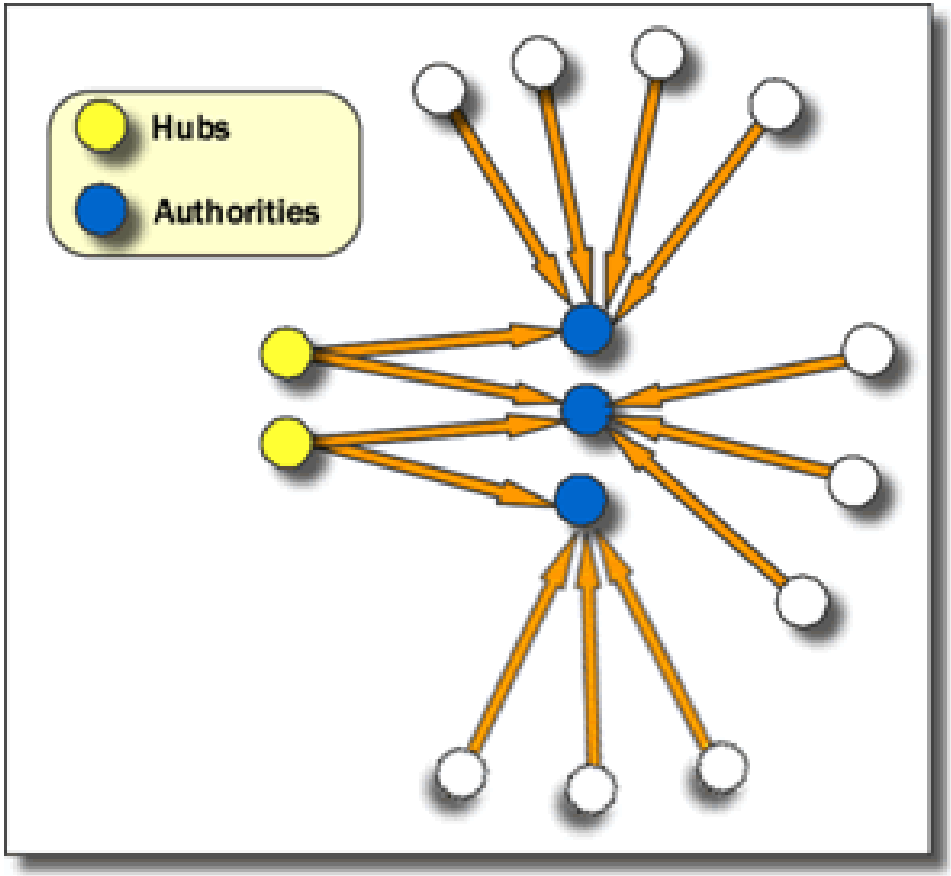
Social Network Analysis: Eigenvalue



- Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, Eigenvalue identifies the most central entities in terms of the global or overall makeup of the network.
- A high Eigenvalue generally:
  - Indicates an actor that is more central to the main pattern of distances among all entities.
  - Is a reasonable measure of one aspect of centrality in terms of positional advantage.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

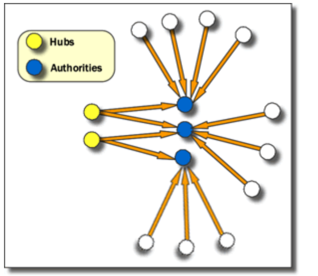
Social Network Analysis: Hub and Authority



Hubs are entities that point to a relatively large number of authorities. They are essentially the mutually reinforcing analogues to authorities. Authorities point to high hubs. Hubs point to high authorities. You cannot have one without the other.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

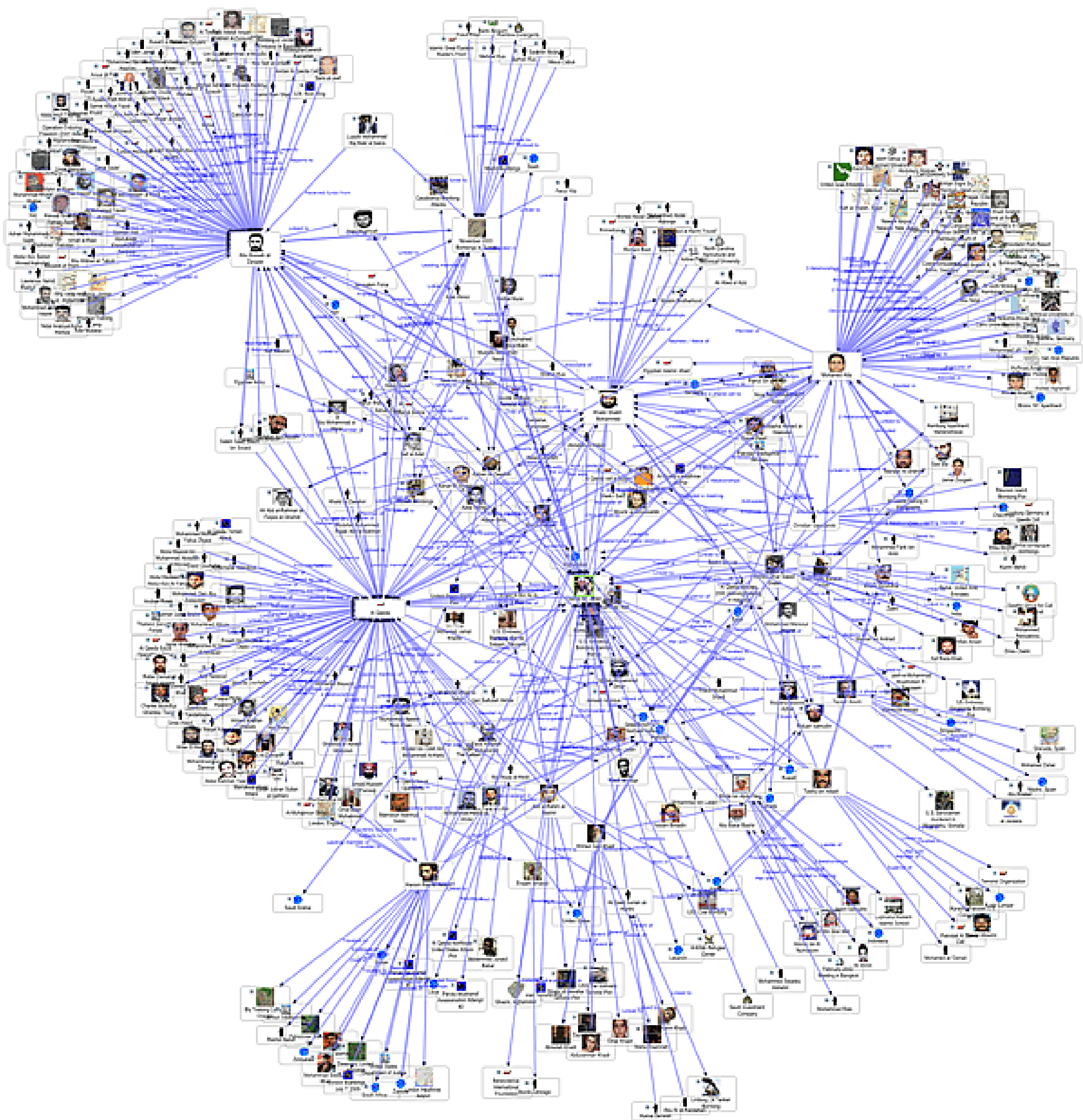
Social Network Analysis: Hub and Authority



- Entities that many other entities point to are called Authorities. In Sentinel Visualizer, relationships are directional—they point from one entity to another.
- If an entity has a high number of relationships pointing to it, it has a high authority value, and generally:
  - Is a knowledge or organizational authority within a domain.
  - Acts as definitive source of information.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis



Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis

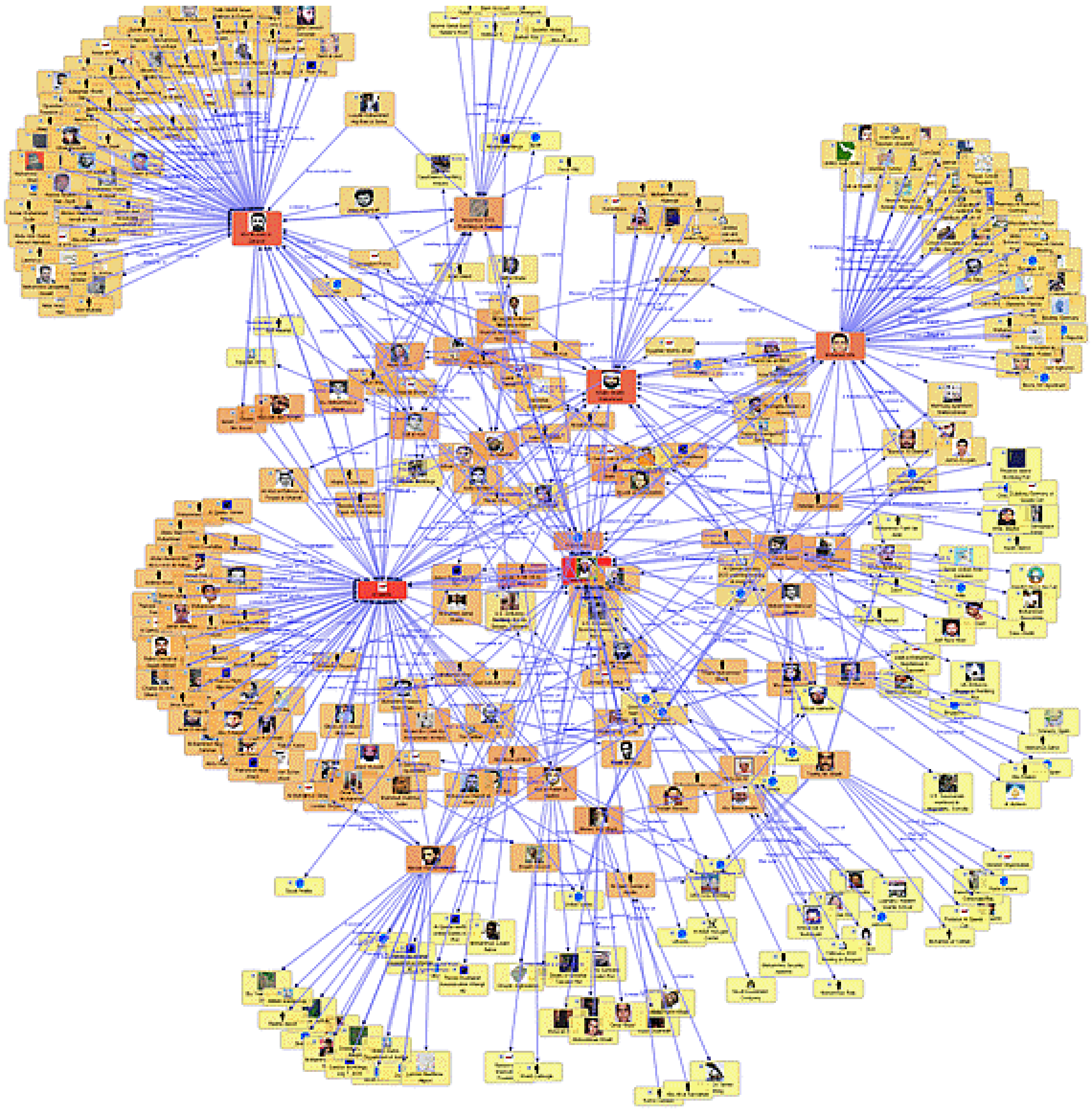
Network Metrics

Calculate   CardView   TableView   Group area   Expand groups   Collapse groups

Name	Type	Degree	Betweenness	Closeness	Eigenvalue	Hub	Authority
Osama bin Laden	Person	44	0.920492092358…	1	0.0271	0	0.011
Abdallah Al-Halabi	Person	2	0	0.654867256637…	0.0001	0	0
Abu Mussab al-Zarqawi	Person	84	0.934887847326…	0.869451697127…	0.7028	0.6572	0.1076
Al Qaeda	Terrorist Organiz…	85	1	0.962427745664…	0.0416	0.3941	0.0166
Ayman Al-Zawahiri	Person	14	0.045794908783…	0.716129032258…	0	0	0.0173
Enaam Arnaout	Person	4	0.031189325814…	0.656804733727…	0.0001	0	0
Imad Eddin Barakat Yarbas	Person	11	0.065049589038…	0.704016913319…	0.0015	0	0.0025
Khalid Shaikh Mohammed	Person	32	0.339916464724…	0.866059817945…	0.002	0	0.1528
Mohamed Atta	Person	61	0.666268740074…	0.820197044334…	0.0015	0	0.6816

Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Social Network Analysis



Source: http://www.fmsasg.com/SocialNetworkAnalysis/

Application of SNA

- Social Network Analysis of Research Collaboration in Information Reuse and Integration

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

Research Question

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

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

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

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.

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

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

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

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

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

Top 20 authors with the highest closeness scores

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

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

Top 20 authors with the highest betweeness scores

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

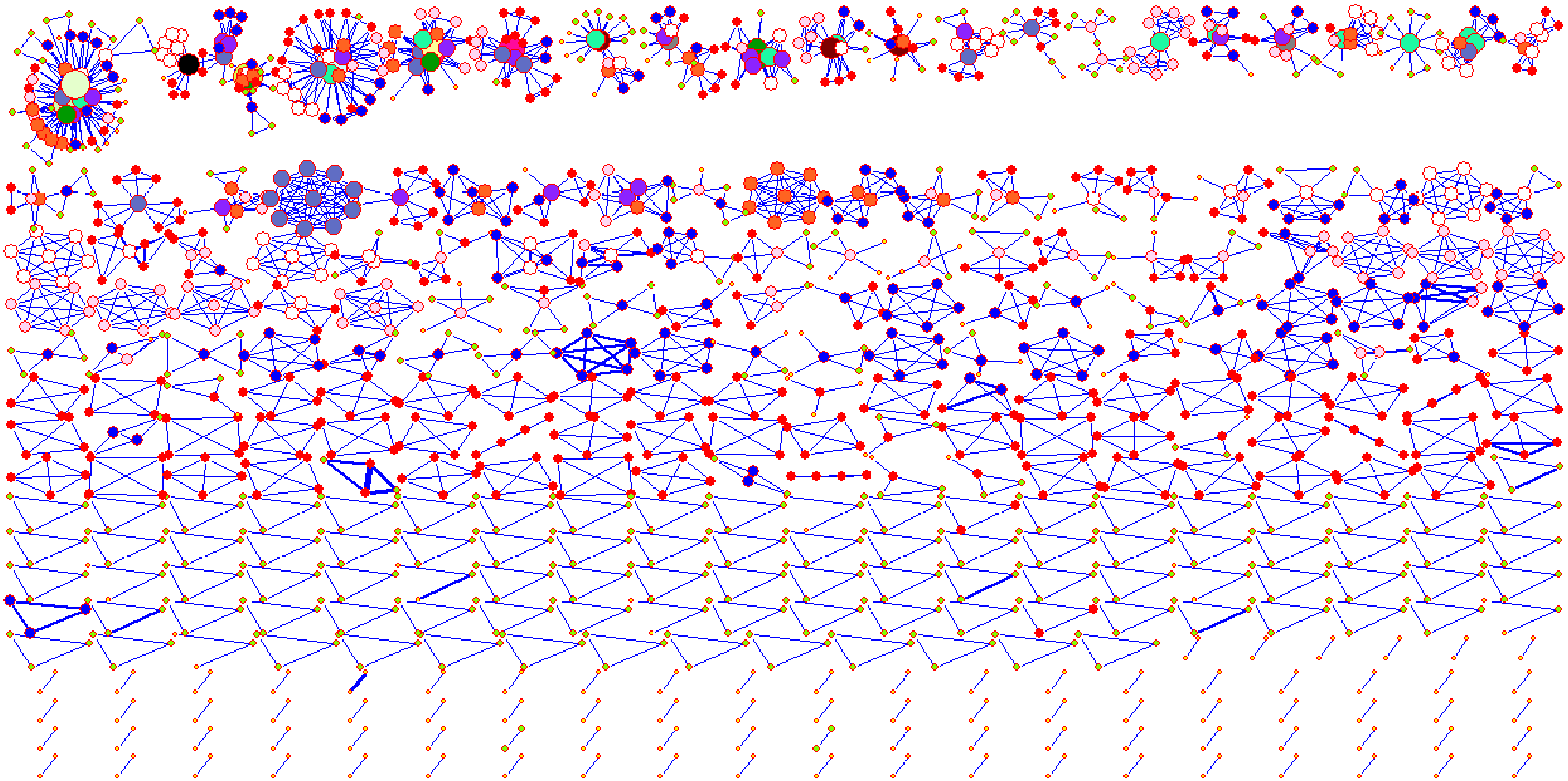
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Top 20 authors with the highest degree scores

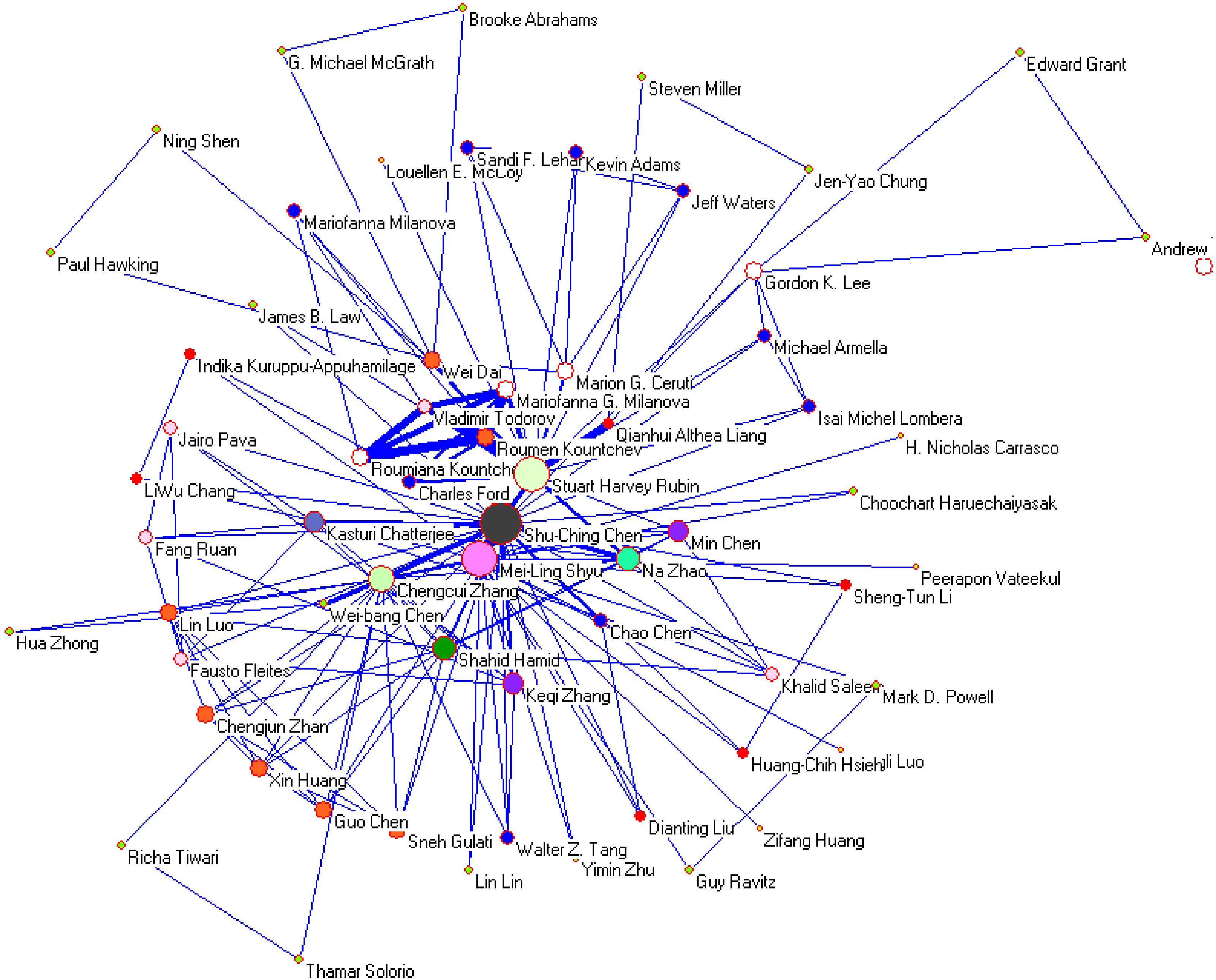
Rank	ID	Degree	Author
1	3	0.035044	Shu-Ching Chen
2	1	0.034418	Stuart Harvey Rubin
3	2	0.030663	Taghi M. Khoshgoftaar
4	6	0.028786	Reda Alhajj
5	8	0.028786	Wen-Lian Hsu
6	10	0.024406	Min-Yuh Day
7	4	0.022528	Mei-Ling Shyu
8	17	0.021277	Richard Tzong-Han Tsai
9	14	0.017522	Eduardo Santana de Almeida
10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

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

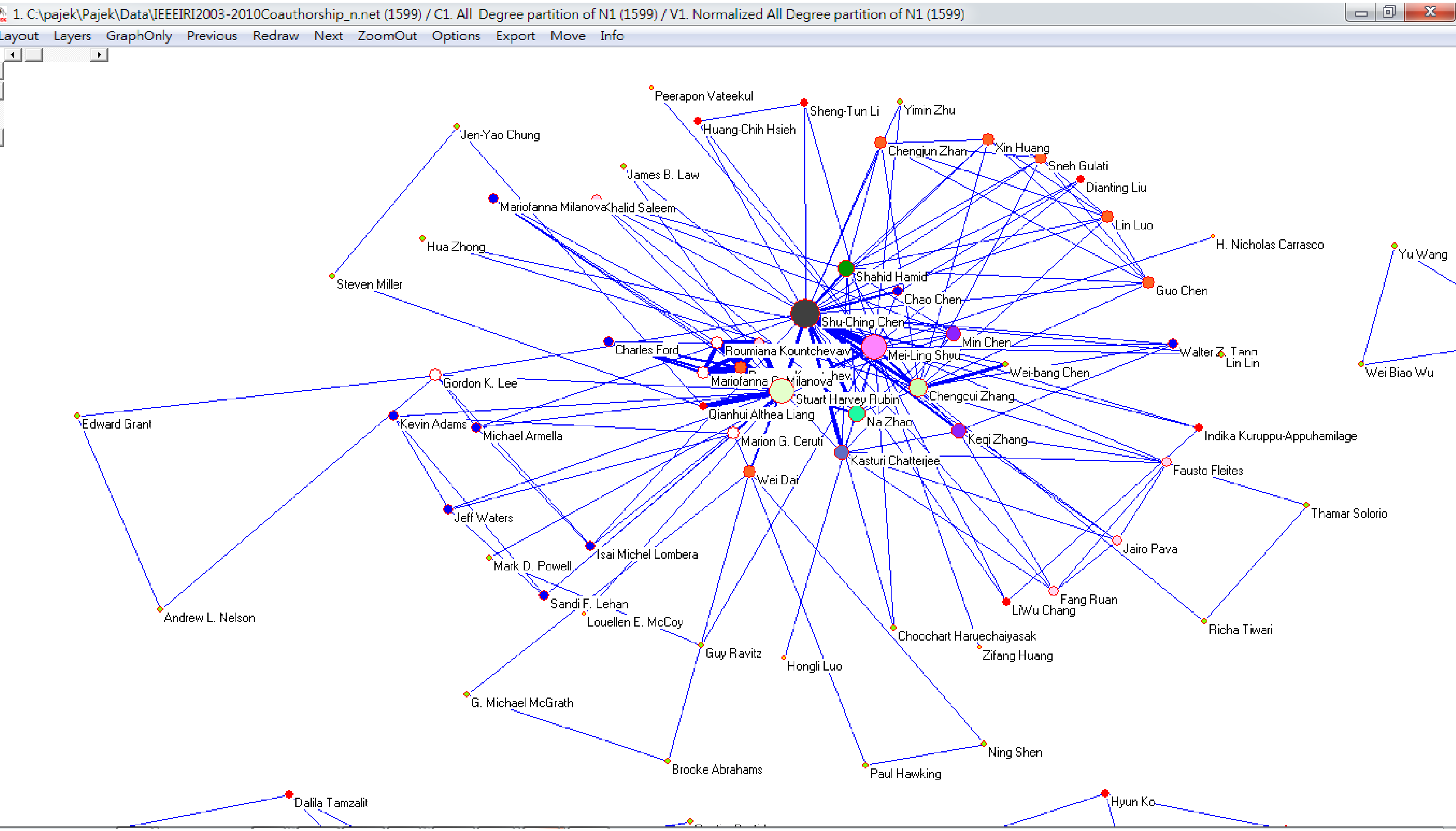
Visualization of IRI (IEEE IRI 2003-2010) co-authorship network (global view)



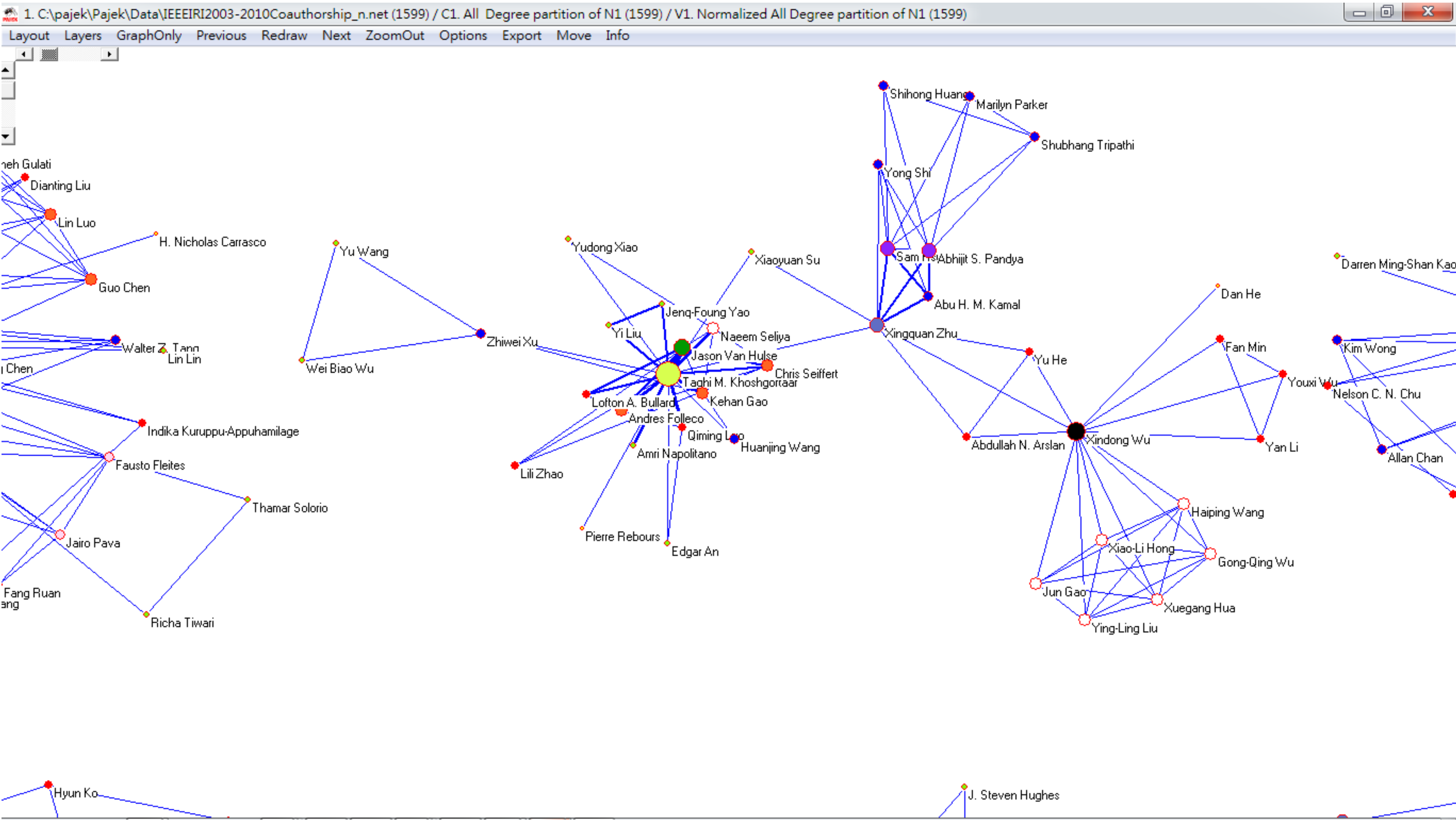
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"



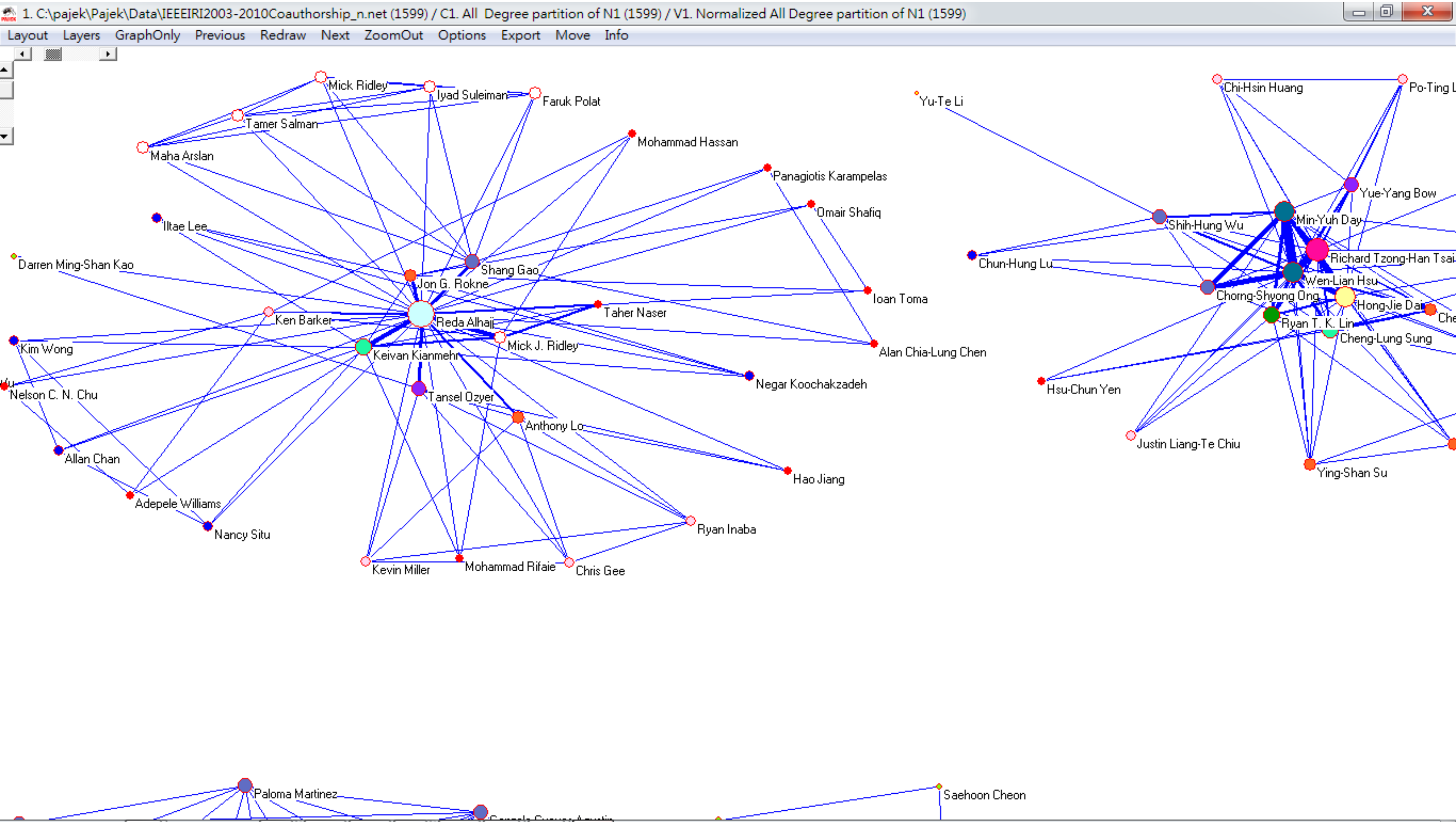
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"



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



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



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

Opinion Mining and Sentiment Analysis

- Mining opinions which indicate positive or negative sentiments
- Analyzes people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes.

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

- "(1) I bought an iPhone a few days ago.
- (2) It was such a **nice** phone.
- (3) The touch screen was really **cool**.
- (4) The voice quality was **clear** too.
- (5) However, my mother was mad with me as I did not tell her before I bought it.
- (6) She also thought the phone was too **expensive**, and wanted me to return it to the shop. … "

+Positive Opinion

-Negative Opinion

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

An aspect-based opinion summary

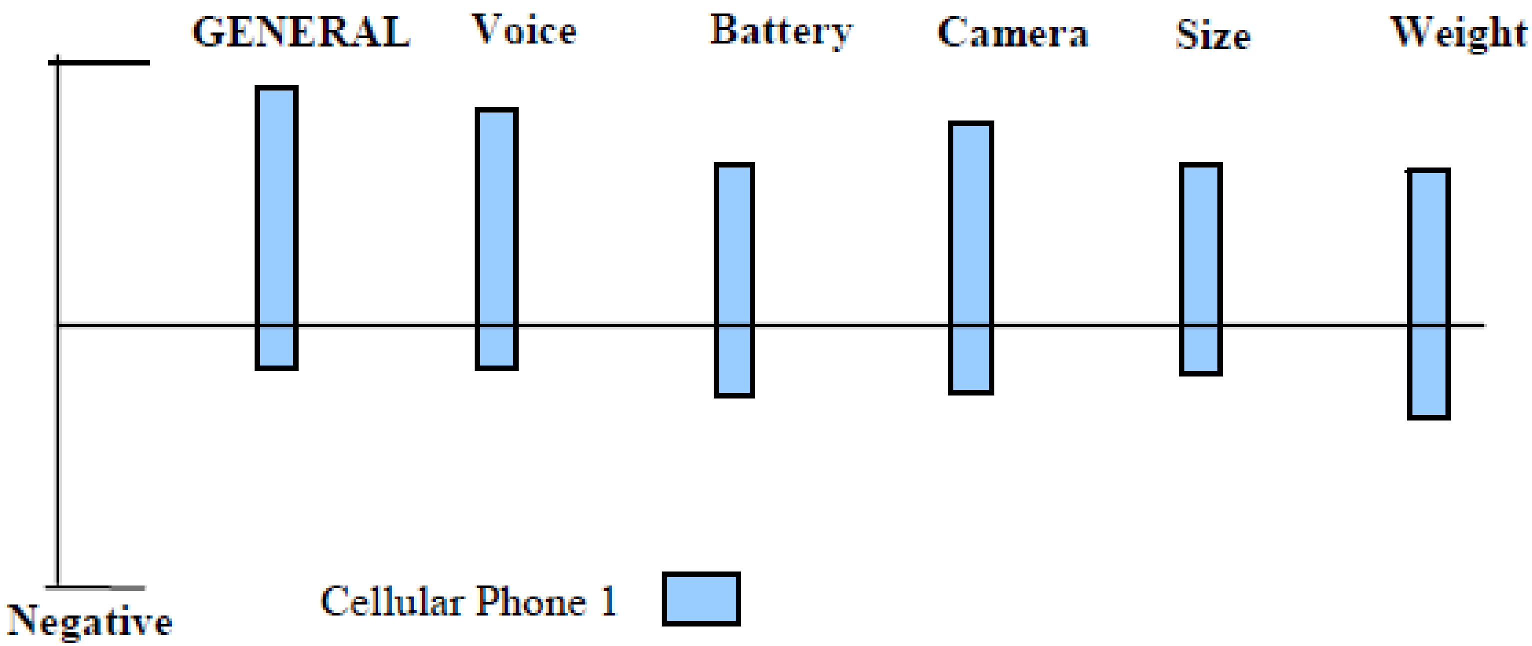
Cellular phone 1:

Aspect: **GENERAL**		
Positive:	125	<individual review sentences>
Negative:	7	<individual review sentences>
Aspect: **Voice quality**		
Positive:	120	<individual review sentences>
Negative:	8	<individual review sentences>
Aspect: **Battery**		
Positive:	80	<individual review sentences>
Negative:	12	<individual review sentences>

…

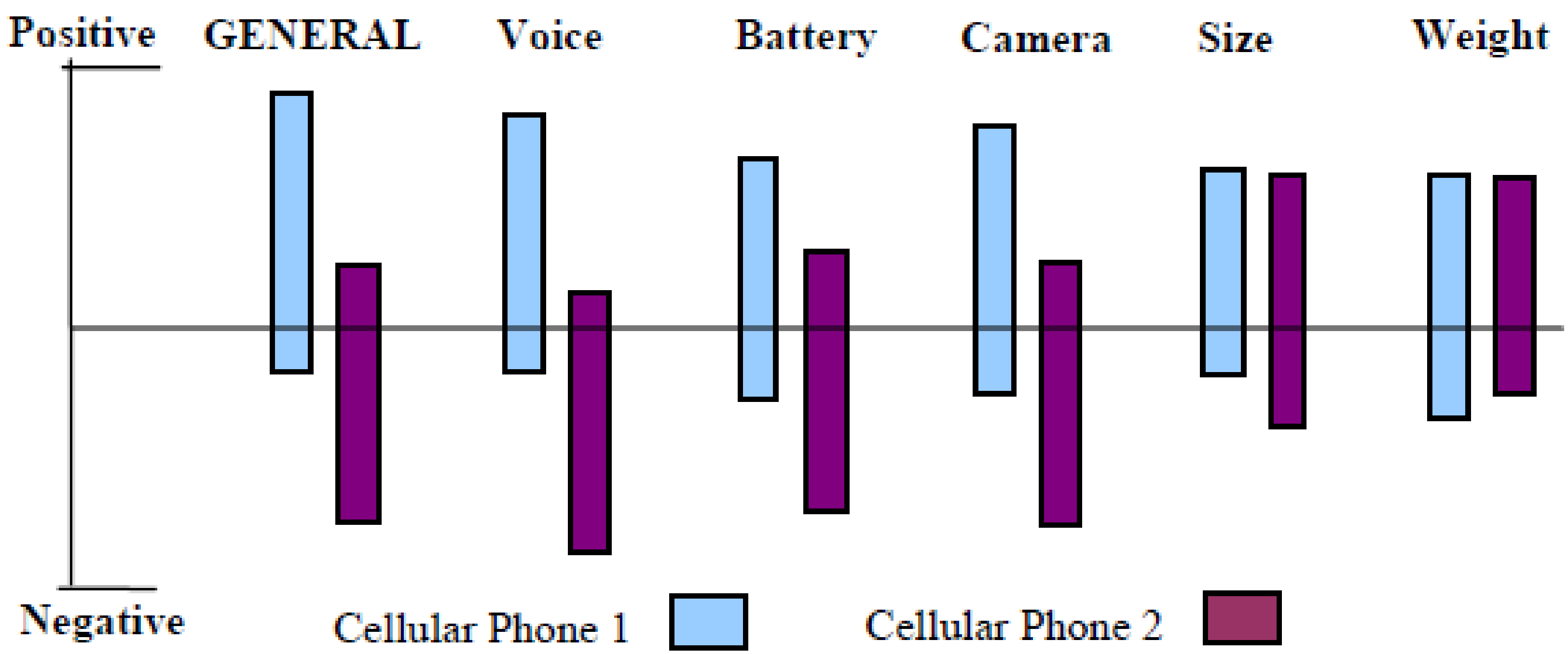
Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Visualization of aspect-based summaries of opinions



Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Visualization of aspect-based summaries of opinions



Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Classification Based on Supervised Learning

- Sentiment classification
  - Supervised learning Problem
  - Three classes
    - *Positive*
    - *Negative*
    - *Neutral*

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

- “(1) I bought an iPhone a few days ago.
- (2) It was such a nice phone.
- (3) The touch screen was really cool.
- (4) The voice quality was clear too.
- (5) However, my mother was mad with me as I did not tell her before I bought it.
- (6) She also thought the phone was too expensive, and wanted me to return it to the shop. … ”

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Opinion words in Sentiment classification

- topic-based classification
  - topic-related words are important
    - e.g., *politics, sciences, sports*
- Sentiment classification
  - topic-related words are unimportant
  - **opinion words** (also called **sentiment words)**
    - **that indicate positive or negative opinions** are important,
      e.g., *great, excellent, amazing, horrible, bad, worst*

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Features in Opinion Mining

- *Terms and their frequency*
  - *TF-IDF*
- *Part of speech (POS)*
  - Adjectives
- *Opinion words and phrases*
  - *beautiful, wonderful, good, and amazing are positive opinion words*
  - *bad, poor, and terrible are negative opinion words.*
  - opinion phrases and idioms, e.g., *cost someone an arm and a leg*
- *Rules of opinions*
- *Negations*
- *Syntactic dependency*

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Rules of opinions

Syntactic template	Example pattern
<subj> passive-verb	<subj> was satisfied
<subj> active-verb	<subj> complained
active-verb <dobj>	endorsed <dobj>
noun aux <dobj>	fact is <dobj>
passive-verb prep <np>	was worried about <np>

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Web Data Mining

Exploring Hyperlinks, Contents, and Usage Data

1. Introduction
2. Association Rules and Sequential Patterns
3. Supervised Learning
4. Unsupervised Learning
5. Partially Supervised Learning
6. Information Retrieval and Web Search
7. Social Network Analysis
8. Web Crawling
9. Structured Data Extraction: Wrapper Generation
10. Information Integration
11. Opinion Mining and Sentiment Analysis
12. Web Usage Mining

Source: http://www.cs.uic.edu/~liub/WebMiningBook.html

Summary

- Social Network Analysis
- Opinion Mining

References

- Sentinel Visualizer, http://www.fmsasg.com/SocialNetworkAnalysis/
- Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011), "Social Network Analysis of Research Collaboration in Information Reuse and Integration," The First International Workshop on Issues and Challenges in Social Computing (WICSOC 2011), August 2, 2011, in Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI 2011), Las Vegas, Nevada, USA, August 3-5, 2011, pp. 551-556.
- Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition, 2011, http://www.cs.uic.edu/~liub/WebMiningBook.html