Web Mining
(網路探勘)

Social Network Analysis
(社會網路分析)

1011WM07
TLMXM1A
Wed 8,9 (15:10-17:00) U705

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/

2012-11-07

課程大綱 (Syllabus)

週次	日期	內容（Subject/Topics）
1	101/09/12	Introduction to Web Mining (網路探勘導論)
2	101/09/19	Association Rules and Sequential Patterns (關聯規則和序列模式)
3	101/09/26	Supervised Learning (監督式學習)
4	101/10/03	Unsupervised Learning (非監督式學習)
5	101/10/10	國慶紀念日(放假一天)
6	101/10/17	Paper Reading and Discussion (論文研讀與討論)
7	101/10/24	Partially Supervised Learning (部分監督式學習)
8	101/10/31	Information Retrieval and Web Search (資訊檢索與網路搜尋)
9	101/11/07	Social Network Analysis (社會網路分析)

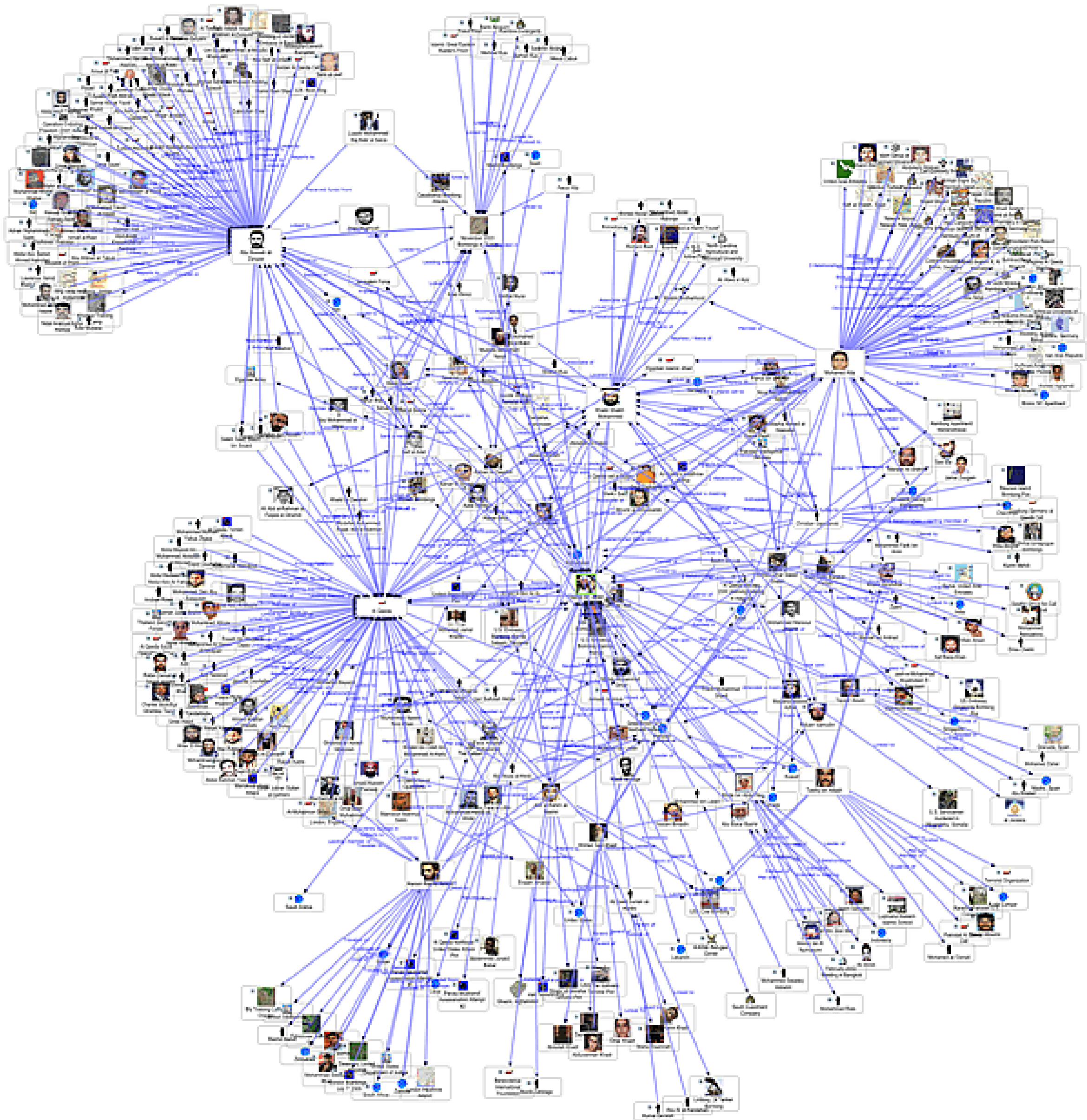
課程大綱 (Syllabus)

週次	日期	內容（Subject/Topics）
10	101/11/14	Midterm Presentation (期中報告)
11	101/11/21	Web Crawling (網路爬行)
12	101/11/28	Structured Data Extraction (結構化資料擷取)
13	101/12/05	Information Integration (資訊整合)
14	101/12/12	Opinion Mining and Sentiment Analysis (意見探勘與情感分析)
15	101/12/19	Paper Reading and Discussion (論文研讀與討論)
16	101/12/26	Web Usage Mining (網路使用挖掘)
17	102/01/02	Project Presentation 1 (期末報告1)
18	102/01/09	Project Presentation 2 (期末報告2)

Outline

- Social Network Analysis (SNA)
  – Degree Centrality
  – Betweenness Centrality
  – Closeness Centrality
- Applications of SNA

Social Network Analysis



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

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

- Social network is the study of social entities (people in an organization, called **actors**), and their interactions and relationships.
- The interactions and relationships can be represented with a network or graph,
  - each vertex (or node) represents an actor and
  - each link represents a relationship.
- From the network, we can study the properties of its structure, and the role, position and prestige of each social actor.
- We can also find various kinds of sub-graphs, e.g., **communities** formed by groups of actors.

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

Social Network and the Web

- Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network,
  - Each page: a social actor and
  - each hyperlink: a relationship.
- Many results from social network can be adapted and extended for use in the Web context.
- Two types of social network analysis,
  - **Centrality**
  - **Prestige**

  closely related to hyperlink analysis and search on the Web

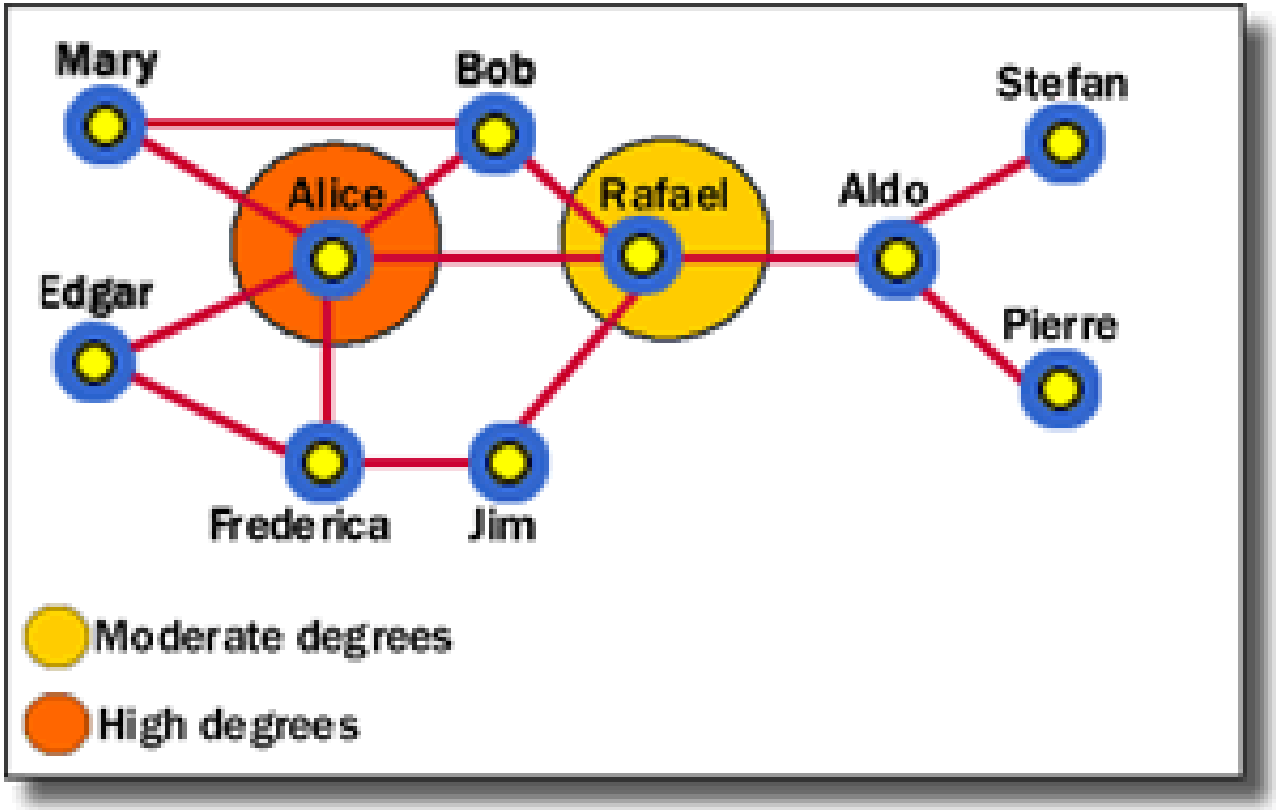
Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data"

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.

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

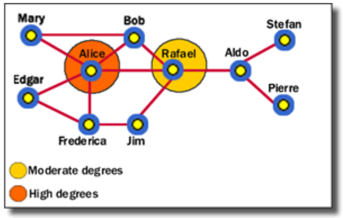
Social Network Analysis: Degree Centrality



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/

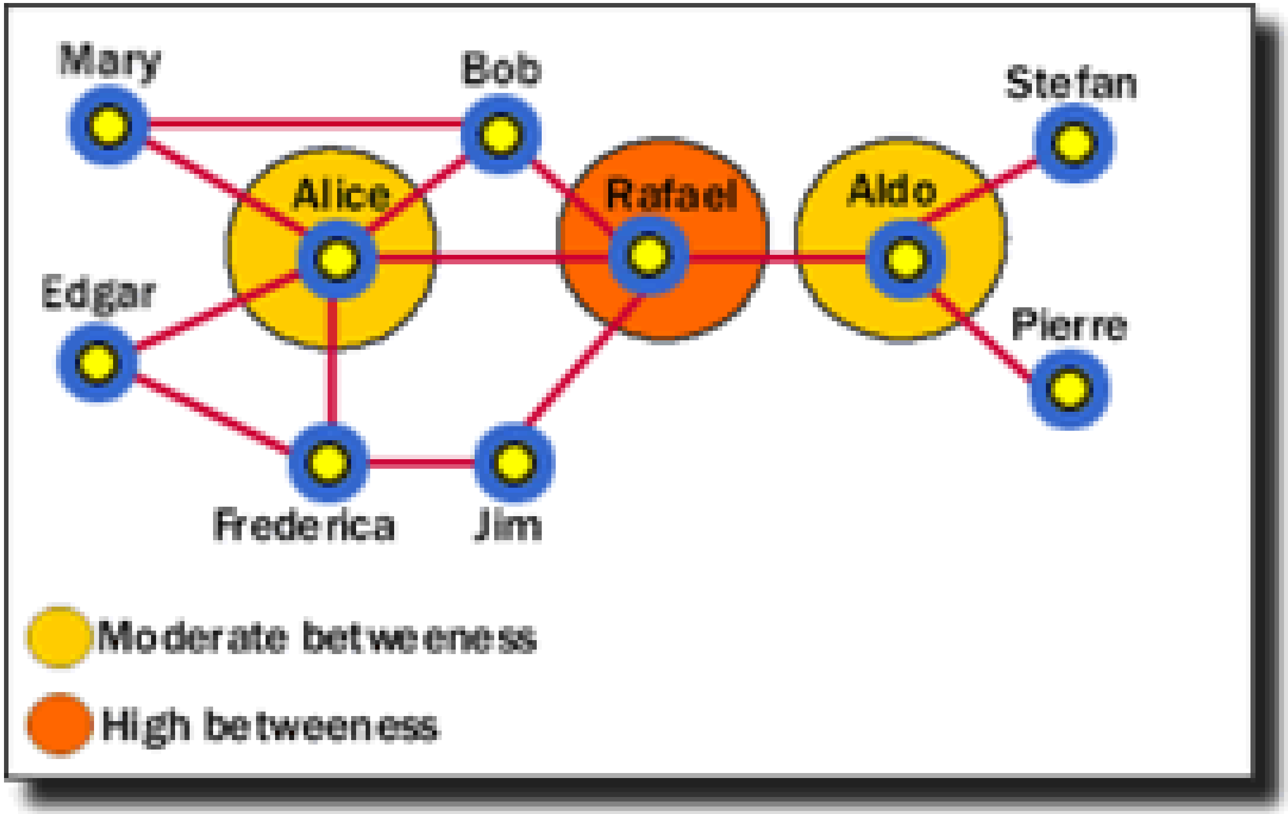
Social Network Analysis: Degree Centrality



- 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/

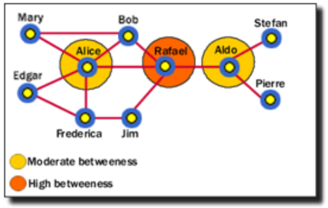
Social Network Analysis: Betweenness Centrality



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/

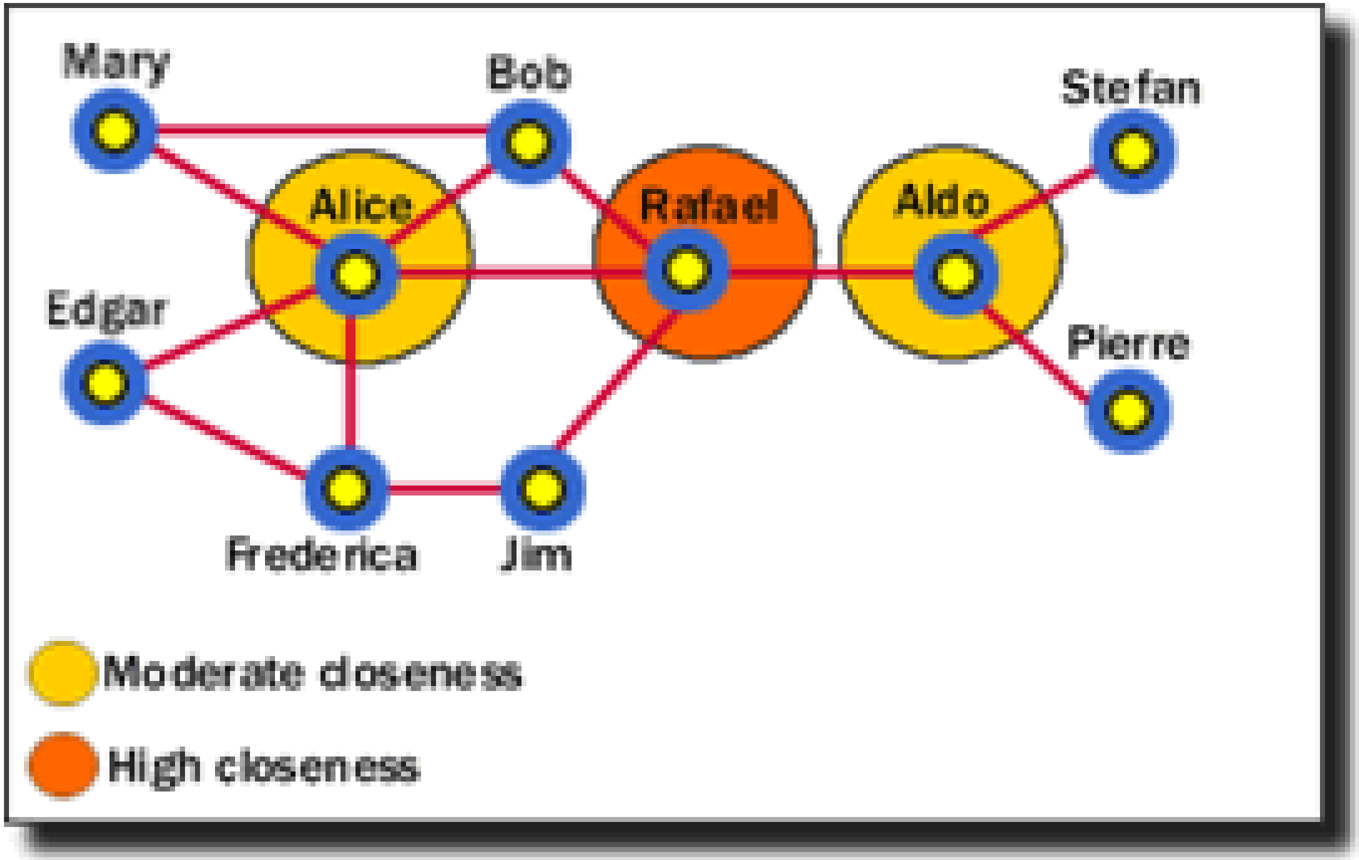
Social Network Analysis: Betweenness Centrality



- 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/

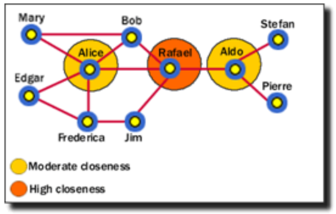
Social Network Analysis: Closeness Centrality



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/

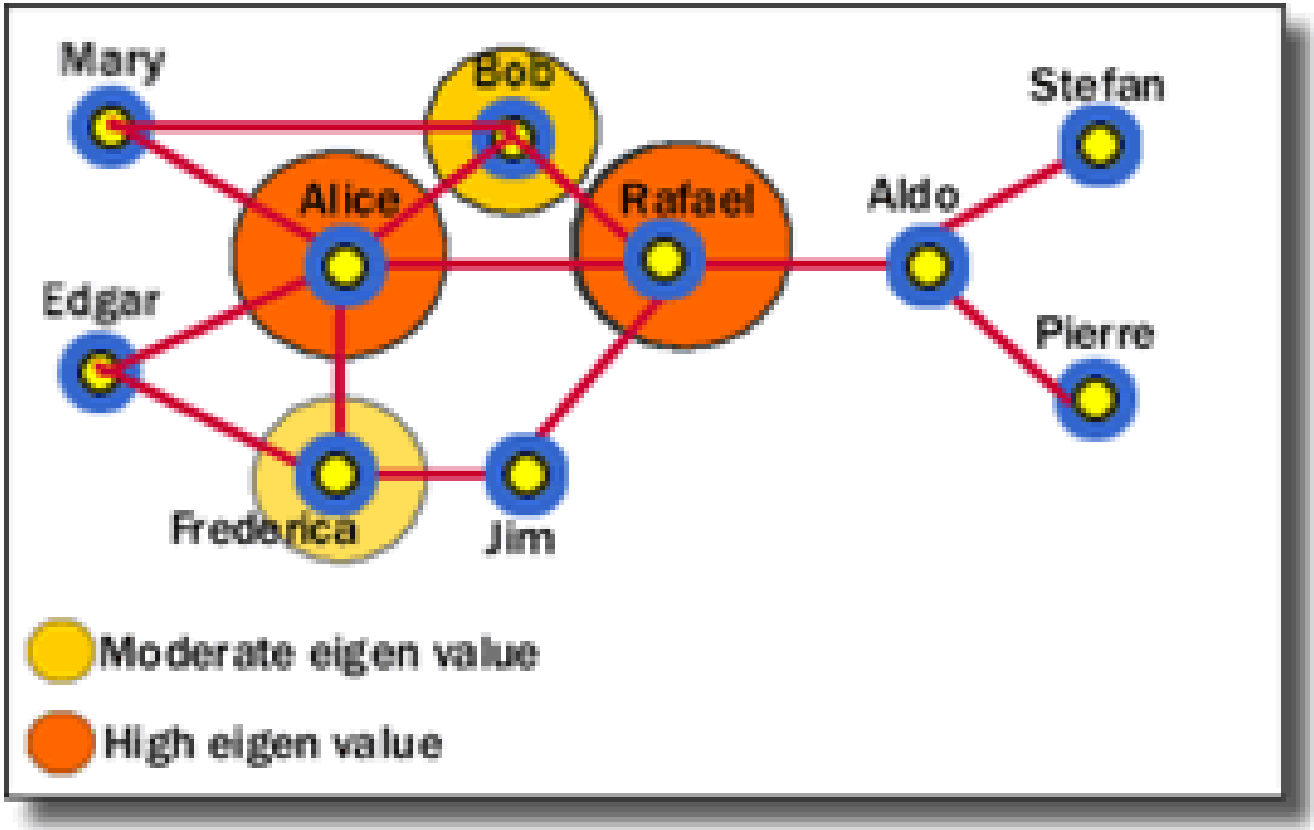
Social Network Analysis: Closeness Centrality



- 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/

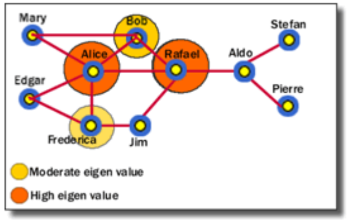
Social Network Analysis: Eigenvalue



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/

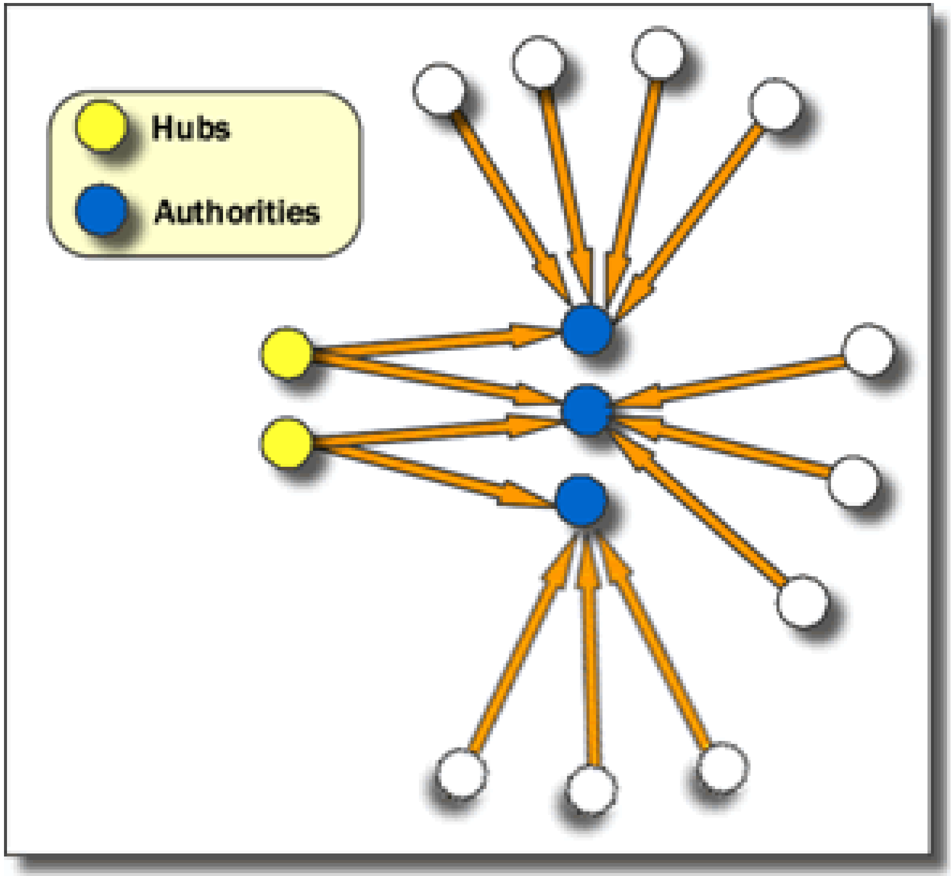
Social Network Analysis: Eigenvalue



- 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/

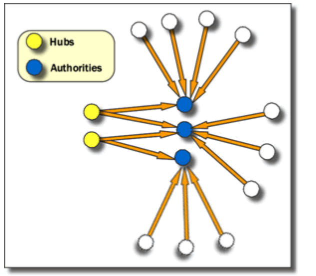
Social Network Analysis: Hub and Authority



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

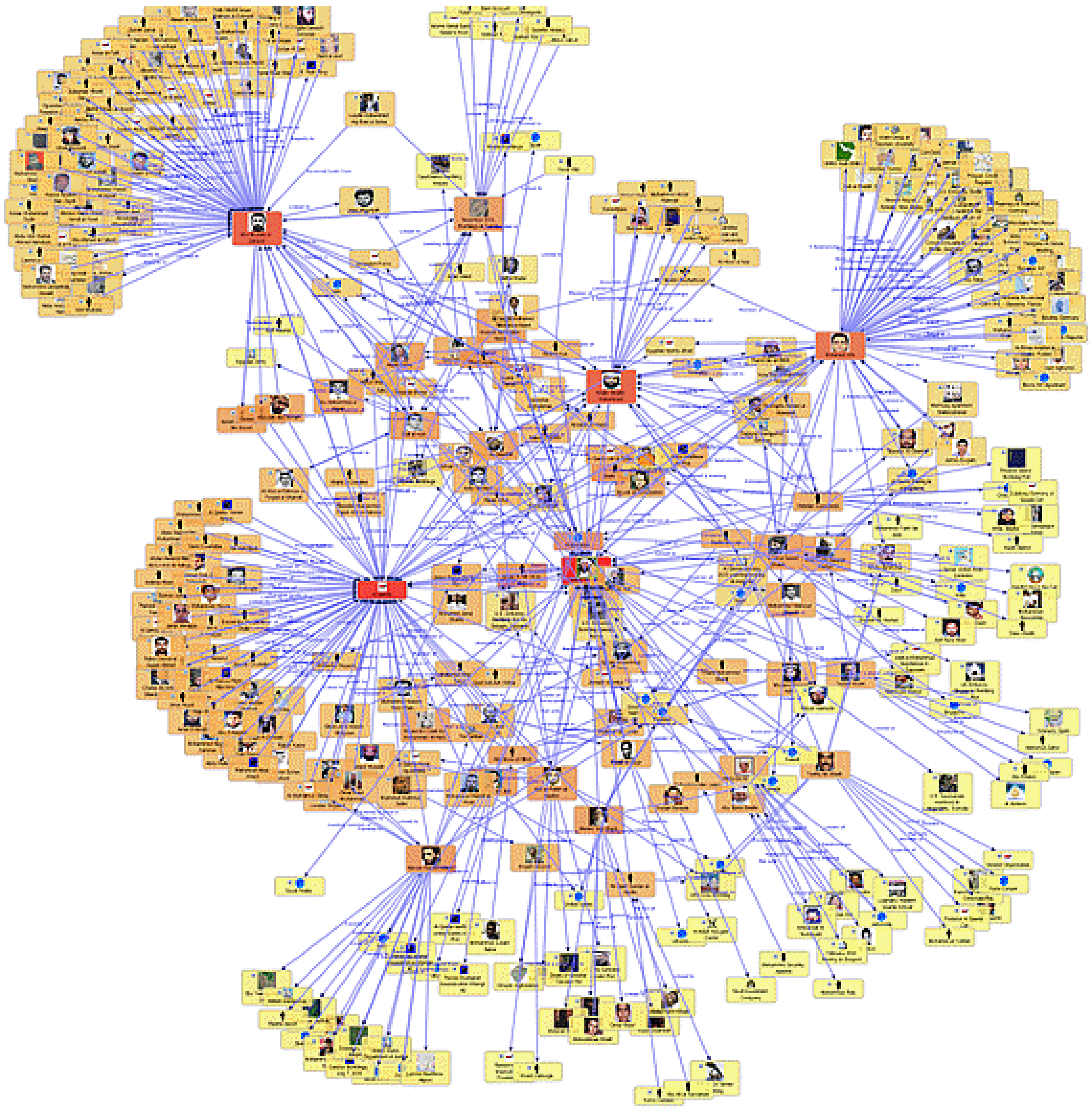
Network Metrics

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/

Degree Centrality

Central actors are the most active actors that have most links or ties with other actors. Let the total number of actors in the network be n.

Undirected graph: In an undirected graph, the **degree centrality** of an actor i (denoted by $C_D(i)$) is simply the node degree (the number of edges) of the actor node, denoted by $d(i)$, normalized with the maximum degree, n-1.

$$C_D(i) = \frac{d(i)}{n-1} \qquad (1)$$

Directed graph: In this case, we need to distinguish **in-links** of actor i (links pointing to i), and **out-links** (links pointing out from i). The degree centrality is defined based on only the out-degree (the number of out-links or edges), $d_o(i)$.

$$C'_D(i) = \frac{d_o(i)}{n-1} \qquad (2)$$

Source: Bing Liu (2011) , Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

Closeness Centrality

This view of centrality is based on the closeness or distance. The basic idea is that an actor x_i is central if it can easily interact with all other actors. That is, its distance to all other actors is short. Thus, we can use the shortest distance to compute this measure. Let the shortest distance from actor i to actor j be $d(i, j)$.

Undirected graph: The closeness centrality $C_C(i)$ of actor i is defined as

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i, j)} \qquad (3)$$

The value of this measure also ranges between 0 and 1 as n-1 is the minimum value of the denominator, which is the sum of shortest distances from i to all other actors. Note that this equation is only meaningful for a connected graph.

Directed graph: The same equation can be used for a directed graph. The distance computation needs to consider directions of links or edges.

Source: Bing Liu (2011) , Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

Betweenness Centrality

- If two non-adjacent actors *j* and *k* want to interact and actor *i* is on the path between *j* and *k*, then *i* may have some control over the interactions between *j* and *k*.
- Betweenness measures this control of *i* over other pairs of actors. Thus,
  - if *i* is on the paths of many such interactions, then *i* is an important actor.

Betweenness Centrality (cont …)

- **Undirected graph**: Let p_{jk} be the number of shortest paths between actor *j* and actor *k*.
- The betweenness of an actor *i* is defined as the number of shortest paths that pass *i* ($p_{jk}(i)$) normalized by the total number of shortest paths.

$$\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}} \qquad (4)$$

Betweenness Centrality (cont …)

Note that there may be multiple shortest paths between j and k. Some passes i and some do not. If we are to ensure the value range is between 0 and 1, we can normalize it with $(n-1)(n-2)/2$, which is the maximum value of the above quantity, i.e., the number of pairs of actors not including i. The final betweenness of i is defined as

$$C_B(i) = \frac{2\sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)} \qquad (5)$$

Unlike the closeness measure, the betweenness can be computed even if the graph is not connected.

Directed graph: The same equation can be used but must be multiplied by 2 because there are now $(n-1)(n-2)$ pairs considering a path from j to k is different from a path from k to j. Likewise, p_{jk} must consider paths from both directions.

Source: Bing Liu (2011) , Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

Prestige

- Prestige is a more refined measure of prominence of an actor than centrality.
  - Distinguish: ties sent (out-links) and ties received (in-links).
- A prestigious actor is one who is object of extensive ties as a recipient.
  - To compute the prestige: we use only in-links.
- Difference between centrality and prestige:
  - centrality focuses on out-links
  - prestige focuses on in-links.
- We study three prestige measures. **Rank prestige** forms the basis of most Web page link analysis algorithms, including PageRank and HITS.

Degree prestige

Based on the definition of the prestige, it is clear that an actor is prestigious if it receives many in-links or nominations. Thus, the simplest measure of prestige of an actor i (denoted by $P_D(i)$) is its in-degree.

$$P_D(i) = \frac{d_I(i)}{n-1}, \tag{6}$$

where $d_I(i)$ is in-degree of i (the number of in-links of actor i_i) and n is the total number of actors in the network. As in the degree centrality, dividing $n - 1$ standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to or chooses actor i.

Proximity prestige

- The degree index of prestige of an actor *i* only considers the actors that are adjacent to *i*.
- The proximity prestige generalizes it by considering both the actors directly and indirectly linked to actor *i*.
  - We consider every actor *j* that can reach *i*.
- Let I_i be the set of actors that can reach actor *i*.
- The **proximity** is defined as closeness or distance of other actors to *i*.
- Let $d(j, i)$ denote the distance from actor *j* to actor *i*.

Proximity prestige (cont …)

$$\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}, \qquad (7)$$

where $|I_i|$ is the size of the set I_i. If we look at the ratio or proportion of actors who can reach i to the average distance that these actors are from i, we obtain the following, which has the value range of [0, 1]:

$$P_P(i) = \frac{|I_i|/(n-1)}{\sum_{j \in I_i} d(j,i)/|I_i|}, \qquad (8)$$

where $|I_i|/(n\text{-}1)$ is the proportion of actors that can reach actor i. In one extreme, every actor can reach actor i, which gives $|I_i|/(n\text{-}1) = 1$. The denominator is 1 if every actor is adjacent to i. Thus, $P_P(i) = 1$. On the other extreme, no actor can reach actor i. Then $|I_i| = 0$, and $P_P(i) = 0$. Each link has the unit distance.

Rank prestige

- In the previous two prestige measures, an important factor is considered,
  - the **prominence** of individual actors who do the “voting”
- In the real world, a person *i* chosen by an important person is more prestigious than chosen by a less important person.
  - For example, if a company CEO votes for a person is much more important than a worker votes for the person.
- If one’s circle of influence is full of prestigious actors, then one’s own prestige is also high.
  - Thus one’s prestige is affected by the ranks or statuses of the involved actors.

Rank prestige (cont …)

- Based on this intuition, the rank prestige $P_R(i)$ is define as a linear combination of links that point to i:

$$P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + \ldots + A_{ni}P_R(n), \tag{9}$$

where $A_{ji} = 1$ if j points to i, and 0 otherwise. This equation says that an actor's rank prestige is a function of the ranks of the actors who vote or choose the actor, which makes perfect sense.

Since we have n equations for n actors, mathematically we can write them in the matrix notation. We use $\boldsymbol{P}$ to represent the vector that contains all the rank prestige values, i.e., $\boldsymbol{P} = (P_R(1), P_R(2), \ldots, P_R(n))^T$ (T means **matrix transpose**). $\boldsymbol{P}$ is represented as a column vector. We use matrix $\boldsymbol{A}$ (where $A_{ij} = 1$ if i points to j, and 0 otherwise) to represent the adjacency matrix of the network or graph. As a notational convention, we use bold italic letters to represent matrices. We then have

$$\boldsymbol{P} = \boldsymbol{A}^T\boldsymbol{P} \tag{10}$$

This equation is precisely the characteristic equation used for finding the **eigensystem** of the matrix $\boldsymbol{A}^T$. $\boldsymbol{P}$ is an **eigenvector** of $\boldsymbol{A}^T$.

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 betweeneness 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

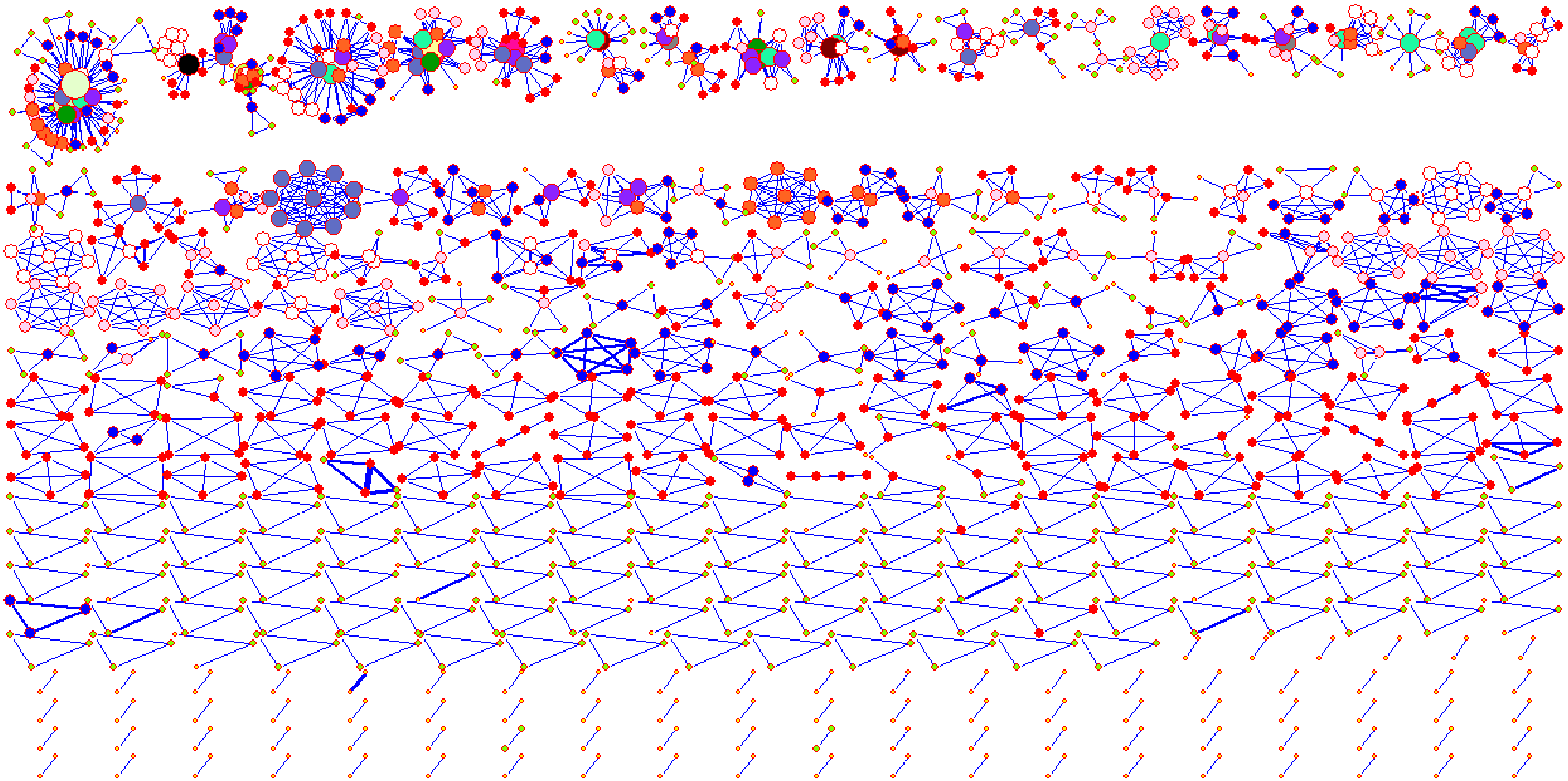
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"

Top 20 authors with the highest degree scores

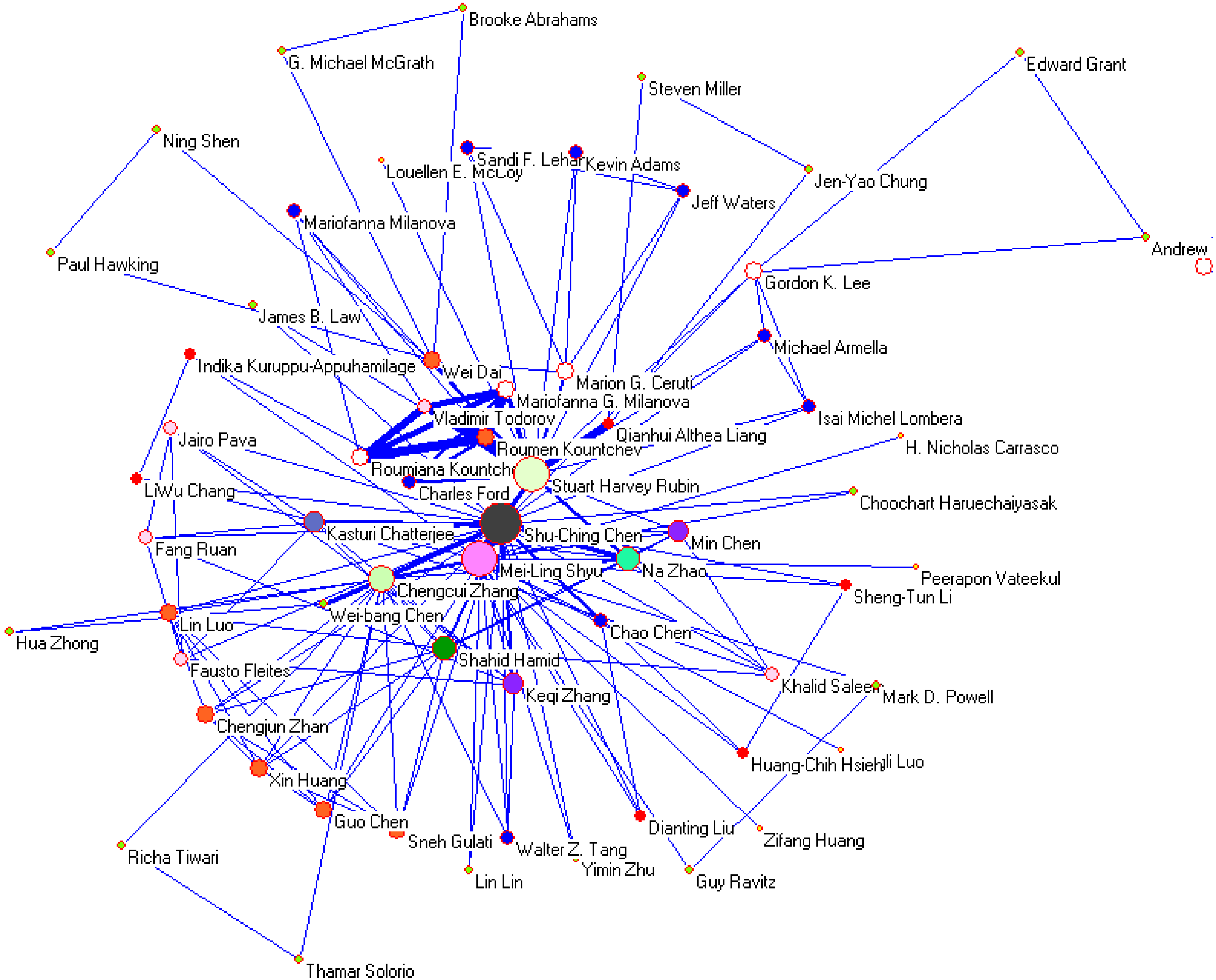
Rank	ID	Degree	Author
1	3	0.035044	Shu-Ching Chen
2	1	0.034418	Stuart Harvey Rubin
3	2	0.030663	Taghi M. Khoshgoftaar
4	6	0.028786	Reda Alhajj
5	8	0.028786	Wen-Lian Hsu
6	10	0.024406	Min-Yuh Day
7	4	0.022528	Mei-Ling Shyu
8	17	0.021277	Richard Tzong-Han Tsai
9	14	0.017522	Eduardo Santana de Almeida
10	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"

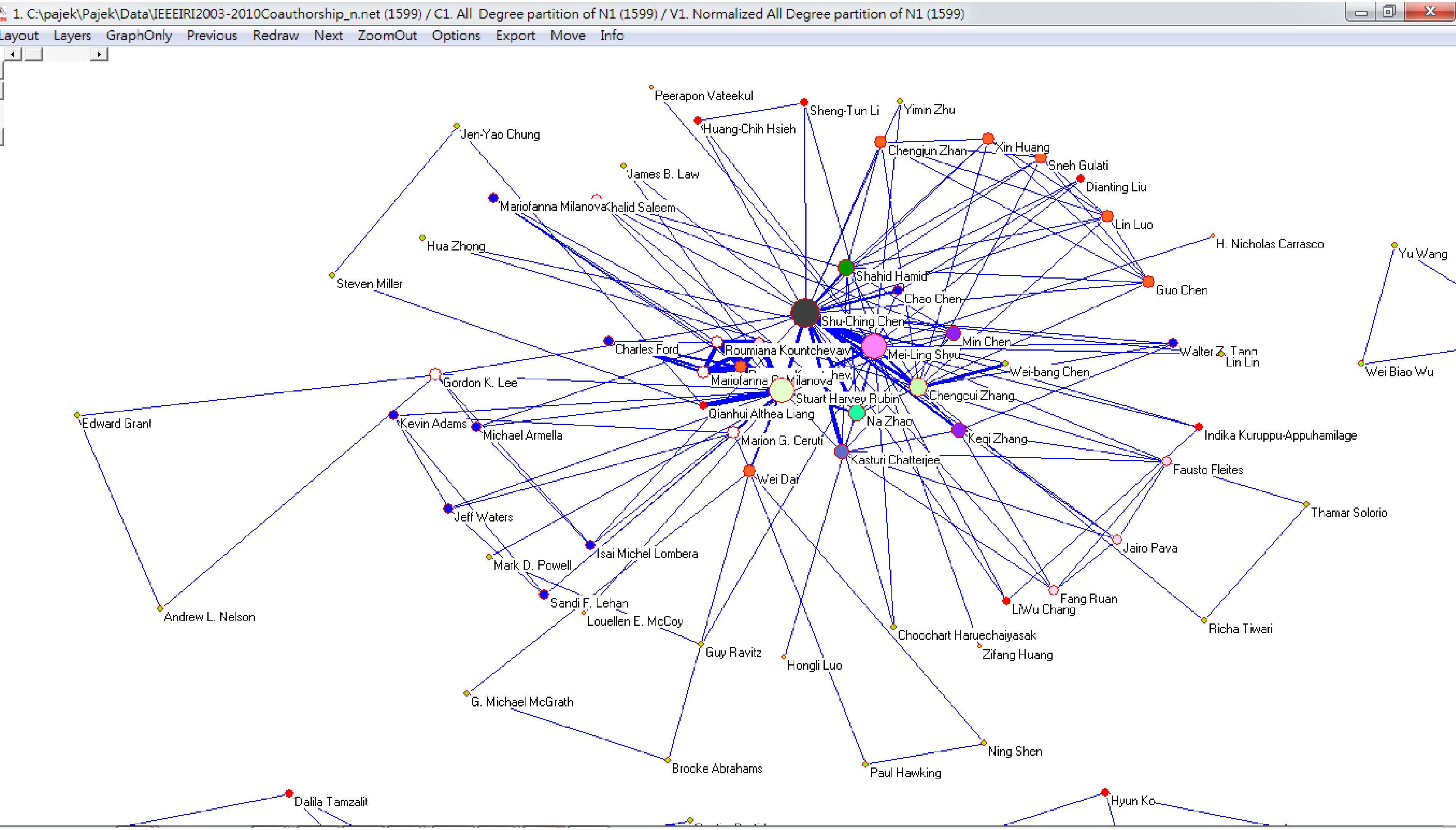
Visualization of IRI (IEEE IRI 2003-2010) co-authorship network (global view)



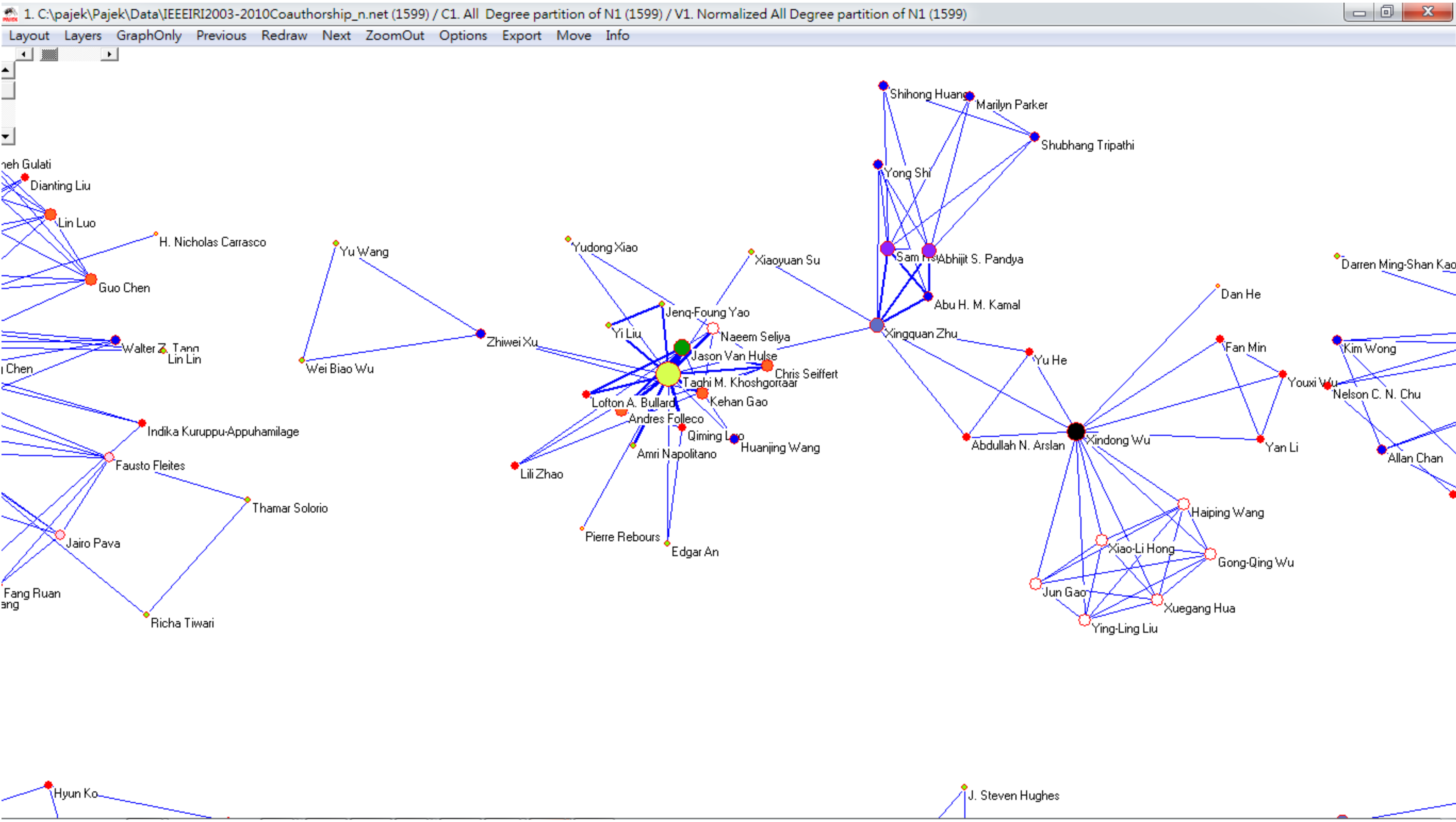
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"



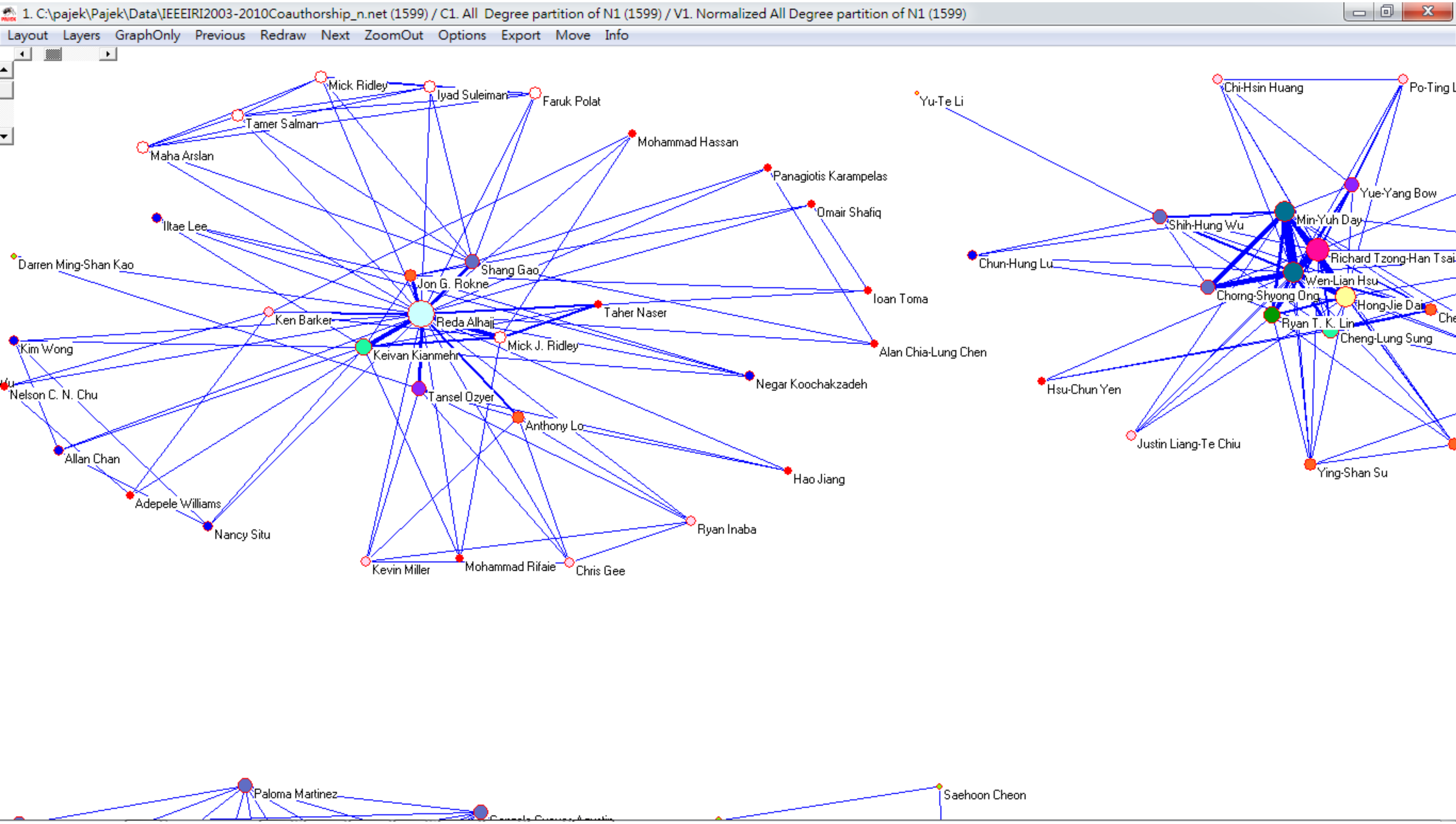
Source: Min-Yuh Day, Sheng-Pao Shih, Weide Chang (2011),
"Social Network Analysis of Research Collaboration in Information Reuse and Integration"



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"

Summary

- Social Network Analysis (SNA)
  - Degree Centrality
  - Betweenness Centrality
  - Closeness Centrality
- Applications of SNA

References

- Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” 2nd Edition, Springer. http://www.cs.uic.edu/~liub/WebMiningBook.html
- 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.