Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks

Ying Ding
School of Library and Information Science, Indiana University, Bloomington, IN, 47405

Abstract
Scientific collaboration and endorsement are well-established research topics which utilize three kinds of methods: survey/questionnaire, bibliometrics, and complex network analysis. This paper combines topic modeling and path-finding algorithms to determine whether productive authors tend to collaborate with or cite researchers with the same or different interests, and whether highly cited authors tend to collaborate with or cite each other. Taking information retrieval as a test field, the results show that productive authors tend to directly coauthor with and closely cite colleagues sharing the same research interests; they do not generally collaborate directly with colleagues having different research topics, but instead directly or indirectly cite them; and highly cited authors do not generally coauthor with each other, but closely cite each other.

Keywords
Scientific collaboration, scientific endorsement, topic modeling, path-finding algorithm

1. Introduction

Bibliometrics measures the standing or influence of an author, journal or article in scholarly networks based on various citation analyses. Citations are understood to serve as carriers of authority and correspond to different endorsements. Scientific collaboration and endorsement are well-established research topics utilizing three kinds of methods (Milojevic, 2010): qualitative methods (e.g., surveys/questionnaires, interviews, or observations), bibliometric methods (e.g., publication counting, citation counting, or co-citation analysis), and complex network methods (e.g., shortest path, centralities, network parameters, or PageRank/HITS). These studies either provide quantitative analytical results about the global network features or rankings of individual network nodes, or qualitative content analysis of survey/observation results. But they are still not able to address the features of the scientific collaboration and endorsement by considering scholar's research interests without involving labor-intensive interviews. This derives the research question of this paper that how to apply quantitative methods to analyze scientific collaboration and endorsement patterns of researchers by considering their research interests. It will contribute to the current state of the art by analyzing scientific collaboration and endorsement from the research topic perspectives to see whether authors tend to collaborate with or cite those having the same or different research topics. Assuming that coauthorship indicates a level of scientific collaboration (Newman, 2004), this paper takes information retrieval (IR) as a test field and applies both a topic modeling algorithm and a path-finding algorithm in coauthorship and citation networks to address these issues.

Easley and Kleinberg (2010) summarized several kinds of common networks: collaboration graphs (i.e., coauthorship networks), who-talks-to-whom graphs (i.e., email communication graphs), information linkage graphs (i.e., citation networks), and technological networks (i.e., peer-to-peer networks). Complex network methods provide several parameters (e.g., degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, diameter, shortest path, distance, clustering coefficient, and geodesic) to characterize the features of these different networks. Most of these network parameters focus on the macro features of the networks by investigating the relation of a node to the rest of other nodes in the graph. Some ranking algorithms (e.g., PageRank, HITS) take the whole network as a graph and utilize the probability propagation of random surfers based on how many times a node is linked by other nodes and how many times a node is linked by other important nodes. Mainstream network analysis research aims to categorize nodes based on network connectivity properties and summarize their distribution patterns. Few studies consider the path between two individual nodes. In a graph, the relationship is manifested as a path between node A and node B with zero or a number of nodes in between, such as, a direct edge between node A and node B is the path with zero nodes in between. Although the shortest path has been applied to capture macro level network features (e.g., betweenness centrality, closeness centrality, distance, and geodesic), it has not been fully applied to identify relationships between two given nodes in a graph. Analyzing specific paths between two nodes can thus reveal micro level features of complex networks.

Most network analyses do not consider topic features of nodes because capturing these features has been a challenge until the Latent Dirichlet Allocation (LDA) (Zhai & Lafferty, 2001; Blei, Ng, & Jordan, 2003). LDA postulates a latent structure between words and documents, which can be extended to include authors to enable the simultaneous modeling of document contents and author interests (Rozen-Zvi, et al, 2004). McCallum, Wang and Corrada-Emmanuel (2007) were among the first to present the extended

LDA model to ascertain topic distributions based on the directed messages sent between people in email communication networks. Their model synchronizes the content of messages and the directed social network within email communications, and can thus discover topics influenced by the social structure of people who send or receive emails. The combination of topic modeling and network analysis extends traditional social network analysis from categorizing network connectivity and distribution to capturing the content richness of social interactions.

The contribution of this paper can be summarized as follows: 1) From the methodology perspective, this study applied the combination of a topic modeling algorithm (especially, the extended LDA model) and a path-finding algorithm to mine research topics of scientists based on their publications and identified their semantic associations based on coauthorship networks and author citation networks. The author is aware of no similar approach being applied in the bibliometric area; 2) From the result perspective, this paper was able to address the collaboration patterns and citation patterns at the topic level rather than at the domain/disciplinary level (Newman, 2004, Milojevic, 2010). Most citation databases (e.g. Web of Science) provide subject categories for papers and journals, but these categories are much broader than the topics mentioned in this paper. When based purely on subject categories provided by citation databases, it is impossible to illustrate the collaboration and citation patterns at further detailed granularities (Bollen, Rodriguez, & Van de Sompel, 2006), such as different research topics within one subject category. Based on methodologies provided in this study, collaboration and citation patterns of different research topic groups within one subject category are detectable; 3) This paper associated topics with authors, while previous studies associated subject categories with papers or journals (Bollen, Rodriguez, & Van de Sompel, 2006); 4) This paper identified the dynamic changes of the collaboration and citation patterns over more than a 40-year time span; 5) Based on the path-finding algorithm, this paper found that there are more than six degress of separation for the current IR coauthorship network and less than three degrees of influence for the current IR author citation network; and 6) This paper proposed the Salton number to measure the distance between any given IR researcher and Salton in the IR coauthorship network.

This paper is organized as follows: Section 2 introduces related work categorized based on applied methodologies. Section 3 presents the dataset, formed networks, and methods applied in this study. Section 4 discusses the findings and their impact. Section 5 concludes this study and suggests future work.

2. Related Work

Studying scientific collaboration and citation networks has become increasingly important in order to better facilitate scientific collaboration and enhance scholarly communication. The related work is organized here based on the different approaches that these studies utilized, namely, network analysis, bibliometrics, and qualitative methods.

Network analysis for collaboration and citation networks

Network analysis uses mathematical models and graph theory to analyze large-scale graphs, mainly on the topological features, such as largest component, centrality, distance, diameter, and cluster coefficient (Grossman, 2002; Barabasi, 2002; Newman, 2004; and Milojevic, 2010). Collaboration graphs are scale-free networks in which the degree distribution follows a power law. The use of shortest path to analyze

network features is mainly based on betweenness and closeness centralities (Freeman, 1977). Goh, Oh, Jeong, Kahng and Kim (2002) found that the betweenness centrality distribution of the collaboration graph follows a power law, which indicates that most authors are sparsely connected while a few authors are intensively connected. Later, Goh et al. (2003) found that authors with high betweenness centralities do not prefer to collaborate with the same sort of authors. Newman (2004) found that scientists who work as a team tend to have shorter average distances to other scientists in the graph. Using Dijkstra's algorithm (Ahuja, Magnanti & Orlin, 1993) to calculate the shortest distances between nodes on a weighted collaboration graph, he found that publishing more papers with many coauthors is a good way to connect with your peers. Moody (2004) laid a sociological foundation for scientific collaboration and proposed several models for the collaboration network. A citation network is defined as a kind of information network that represents the network of relatedness of subject matter (Newman, 2010). An, Janssen, and Milios (2004) analyzed the network feature of a citation graph in the computer science area. They found that the citation graph is not connected and the probability of having a directed path between any pair of nodes is less than 2%. Newman (2010) found that citation networks in Web of Science follow a power law. Citation networks can also be extended to patent citation networks (Li, Chen, Zhang & Li, 2007) and legal citation networks (Zhang & Koppaka, 2007). Other related researches applying complex network methods to coauthorship and citation networks are Kretschmer (2004), Liu, Bollen, Nelson and Sompel (2005), Vidgen, Henneberg and Naude (2007), Rodriguez and Pepe (2008), and Yan and Ding (2009). These studies analyzed either the macro-level network features of coauthorship or citation networks, or the individual author rankings within different domains. None of them addressed the collaboration and citation patterns of researchers from the same or different research topic groups. This paper aims to fill this gap by applying the combined approach of a topic modeling algorithm and a path-finding algorithm to find whether productive authors tend to coauthor with or cite researchers sharing the same research topics or those having different research topics.

Bibliometric methods for collaboration and citation networks

Citation networks have been widely studied in bibliometrics (Small, 1973; White & Griffith, 1981). These studies took co-citation networks as similarity graphs to unveil the disciplinary/intellectual structures of a domain or school of thought by using author co-citation analysis (Ding, Chowdhury & Foo, 1999), journal co-citation analysis (Ding, Chowdhury & Foo, 2000), and co-word analysis (Ding, Chowdhury & Foo, 2000a). Citation analysis focuses on counting the number of citations in different ways, which can be viewed as calculating the degree of nodes in various citation graphs without taking graph topology into consideration (e.g., impact factor (Garfield, 1972)). PageRank or HITS related studies calculated the eigenvector centrality of nodes (Kleinberg, 1998, Brin & Page, 1998). Bibliometrically, collaboration is not normally viewed or studied graphically, but rather as a social phenomenon with different factors, including economic factors (Price, 1966), intra-scientific factors (Beaver, 2001), scientific acknowledgement factors (Cronin, Shaw, & La Barre, 2003; Giles & Councill, 2004), organizational factors (Glanzel & Schubert, 2004), geographical factors (Ding, Foo & Chowdhury, 1998; Glanzel, 2001), social stratification factors (Kretschmer, 1994), sector factors (Leydesdorff & Etzkowitz, 1996), and academic credit factors (Katz & Martin, 1997; Cronin, 2001). These bibliometric investigations studied citation patterns based either on co-citation networks with the aim to detect intellectual structures of domains, or on PageRank to rank individual authors or journals. However, the collaboration patterns were not investigated utilizing research topic factors. This paper will contribute to the current state of the art by

analyzing collaboration patterns from research topic perspectives to see whether authors tend to collaborate with those having the same research topics or having different research topics.

Survey methods for collaboration and citation networks

Before analysis of large-scale networks became feasible, qualitative methods were used to study social interactions, such as collaboration and acquaintanceship. Networks were constructed by interviewing participants and distributing questionnaires. Although these studies have revealed details about the cognitive, psychological, and sociological features of networks, they suffer from several problems (i.e., time-consuming, privacy concerns, subjectiveness, sampling issues, and statistical accuracy) and limit themselves to small-size networks (Newman, 2004a). Laudel (2002) conducted an empirical investigation of research collaboration based on 101 semi-structured interviews with research group leaders and members. She identified six types of research collaborations with distinct reward patterns and found that around 50% of collaborations are not rewarded in formal communication channels and only one third are rewarded by acknowledgements. Hara, Solomon, Kim and Sonnenwald (2003) collected collaborative data by using interviews with around 100 members in four multidisciplinary research groups, observations of videoconferences and meetings, and a center-wide sociometric survey to analyze scientists' perspectives on collaboration and factors that impact collaboration. They developed a framework that identifies forms of collaboration (e.g., complementary and integrative collaboration) and associated factors (e.g., personal compatibility, work connections, incentives and infrastructure). Similar surveys have been conducted to investigate knowledge flows produced by patent citations (Jaffe, Trajtenberg & Fogarty, 2000) and to measure law reviews' influence on judicial decisions (McClintock, 1998). Although these survey-based researches can identify detailed features of collaboration patterns, none of them analyzed the topic features of collaboration patterns. Furthermore, they are costly, subjective, and based on a limited data size. This paper investigated the topic features of collaboration patterns based on large-scale collaboration and citation networks. It has also been able to identify the dynamic changes of these patterns based on different time spans.

3. Methodology

Data

Information retrieval (IR) was selected as the test area. Papers and their citations were collected from Web of Science (WOS) covering the time period of 1956-2008. Based on a set of search terms related to IR, the following queries were formed: INFORMATION RETRIEVAL, INFORMATION STORAGE and RETRIEVAL, QUERY PROCESSING, DOCUMENT RETRIEVAL, DATA RETRIEVAL, IMAGE RETRIEVAL, TEXT RETRIEVAL, CONTENT BASED RETRIEVAL, CONTENT-BASED RETRIEVAL, DATABASE QUERY, DATABASE QUERIES, QUERY LANGUAGE, QUERY LANGUAGES, and RELEVANCE FEEDBACK. In total, 15,367 papers with 350,750 citations were gathered. Citation records contain only the first author, year, source, volume, and page number. The dataset was divided into four phases: 1956-1980 (Phase 1), 1981-1990 (Phase 2), 1991-2000 (Phase 3), and 2001-2008 (Phase 4). Table 1 shows the details of the IR dataset in these four time periods.

Table 1. Overview of the IR dataset

	1956-1980	1981-1990	1991-2000	2001-2008	Total

No. of papers	1,313	1,173	4,485	8,396	15,367
No. of citations	10,862	17,874	110,454	211,560	350,750

Coauthor networks and citation networks

Coauthorship networks document scientific collaboration through published articles, where nodes are authors and a link represents the fact that two authors have written at least one paper together. Coauthorship networks are thus undirected networks. Citation networks document the citing behavior via scholarly publications, where nodes are authors and a link represents the citing of one author by another. Citation networks are thus directed networks. The last name with first initial was used as author name in this study to represent unique authors (Newman, 2004, Milojevic, 2010). Table 2 shows the details of IR coauthorship networks and citation networks.

Table 2. Overview of IR coauthorship and citation networks

(No. of nodes, No. of edges)	1956-1980 (Phase 1)	1981-1990 (Phase 2)	1991-2000 (Phase 3)	2001-2008 (Phase 4)
Coauthorship network	(930, 4256)	(961, 2252)	(6650, 24184)	(13640, 63140)
Citation network	(6054, 11192)	(5978, 17084)	(36411, 171814)	(62636,444203)

Note: Single-authored papers are not considered in the coauthorship networks. Papers that do not contain references are not included in citation networks.

Topic Modeling Algorithm

This paper applied the Author-Conference-Topic (ACT) model developed by Tang, Jin and Zhang (2008). The ACT model is an extended LDA model used to simultaneously extract topic features of papers, authors, and publication venues. Conference represents a general publication venue which also includes journals, workshops and organizations. Figure 1 displays the plate notation of the ACT model, in which gray and white circles indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and plates. Plates indicate a repeated sampling with the number of repetitions given by the variable in the lower corner (Buntine, 1994). Here d is document, w is word, a_d is the set of co-authors, x is author, z is topic, α, β and μ are hyperparameters, θ and ϕ are multinomial distributions over topics and words, respectively, and ψ is a multinomial distribution over publication venues.

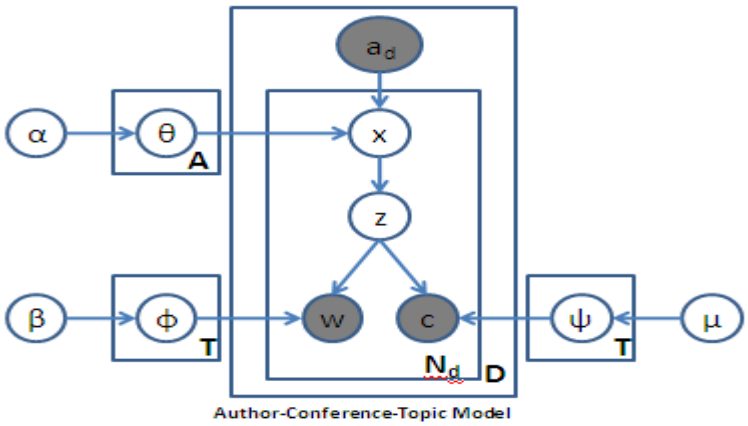


Figure 1. The plate notation of the ACT model

The ACT model calculates the probability of a topic given an author, the probability of a word given a topic, and the probability of a conference given a topic. The Gibbs sampling is used for inference, and the hyperparameters α, β, and μ are set at fixed values (α=50/T, β=0.01, and μ=0.1). The posterior distribution is estimated based on *x* and *z* only, and the results are used to infer θ, φ, and ψ. The posterior probability is calculated as:

$$P(z_{di}, x_{di} | z_{-di}, x_{-di}, w, c, \alpha, \beta, \mu) \propto \frac{m_{x_{di} z_{di}}^{-di} + \alpha_{z_{di}}}{\sum_z (m_{x_{di} z}^{-di} + \alpha_z)} \times \frac{n_{z_{di} w_{di}}^{-di} + \beta_{w_{di}}}{\sum_{w_v} (n_{z_{di} w_v}^{-di} + \beta_{w_v})} \times \frac{n_{z_{di} c_d}^{-d} + \mu_{c_d}}{\sum_c (n_{z_{di} c}^{-d} + \mu_c)}$$

After the Gibbs sampling, the probability of a word given a topic φ, the probability of a conference given a topic ψ, and the probability of a topic given an author θ can be estimated as:

$$\phi_{z w_{di}} = \frac{n_{z w_{di}} + \beta_{w_{di}}}{\sum_{w_v} (n_{z w_v} + \beta_{w_v})}$$

$$\psi_{z c_d} = \frac{n_{z c_d} + \mu_{c_d}}{\sum_c (n_{zc} + \mu_c)}$$

$$\theta_{xz} = \frac{m_{xz} + \alpha_z}{\sum_{z'} (m_{xz'} + \alpha_{z'})}$$

A paper d is a vector w_d of N_d words, in which each w_{di} is chosen from a vocabulary of size V. A vector a_d of A_d authors is chosen from a set of authors of size A, and c_d represents a conference. A collection of D papers is defined by $D = \{(w_1, a_1, c_1), \ldots (w_{D,} a_D, c_D)\}$, where x_{di} denotes an author chosen from a_d and is responsible for the *ith* word w_{di} in paper d. The number of topics is denoted as T. This paper applied the ACT model to extract the topical distribution of authors to derive their research interests. For each phase, five topics were extracted by the ACT model, and each topic contained the list of top-ranked authors and keywords which have high probabilities to be associated with this topic. Compared with tradition co-citation analysis (e.g., co-word analysis, co-author analysis, and co-journal analysis), the ACT model has the following advantages: 1) it can cluster author, word and journal at the same time so that one topic cluster contains a list of authors, a list of words, and a list of journals; 2) it can mine latent topics via *z* in Figure 1, while co-citation analysis cannot detect such hidden topics. The appendix shows the example of five extracted topics and their corresponding top 10 keywords and authors in 2001-2008. These top-ranked authors are productive authors who produce most of the words associated with this topic.

Path-Finding Algorithm

Given a graph $G = (V, E)$, where V represents a set of nodes and E is a set of edges linking two nodes. E can be reincarnated into different relationships between two nodes, such as endorses/cites, knows, and collaborates. Edges may have different weights to illustrate *importance*, *influence* and *frequency*. The topology of the network is reflected by the asymmetric adjacency matrix $A = (A_{ij})$, where $A_{ij} = 1$ if v_i links to v_j and $A_{ij} = 0$ if not. The Breadth-First Search (BFS) algorithm is commonly used to find the shortest paths between two nodes and normally takes $O(n^2)$ time, where n is the total number of edges in the graph (Knuth, 1997). This paper applied the path-finding algorithm developed by Jie Tang (via

personal communication: for more details about this algorithm, please refer to He et al. (2010 submitted)). This algorithm can shorten the computing time to *O(nlog(n))* by simultaneously applying BFS on the two nodes until one path has formed in the middle. It also uses other optimization processes to reduce the computing complexity. Additionally, it can calculate near-shortest paths and derive subgraphs of two given nodes in a scalable way. Figure 2 illustrates the process of finding one shortest path between node 1 and node 26 (Figure 2-a):

1. BFS explores the nearest neighbor of node 1 and it reaches node 3, 4, 6, 7, 10 (Figure2-b);
2. Meanwhile, another BFS explores the nearest neighbor of node 26 similarly and reaches node 19, 21, 23, 24, 25 (Figure 2-c);
3. Explore all the nearest neighbors of node 3, 4, 6, 7, 10, and reach 2, 5, 8, 9, 11, 14, 18 (Figure2-d);
4. Meanwhile, explore all the nearest neighbors of node 19, 21, 22, 23, 24, 25, and reach 15, 16, 18, and 22 (Figure 2-e); and
5. Two BFS processes meet at node 18 and the algorithm ends. The shortest path between node 1 and node 26 is 1 – 10 – 18 – 21 – 26 (Figure 2-f).

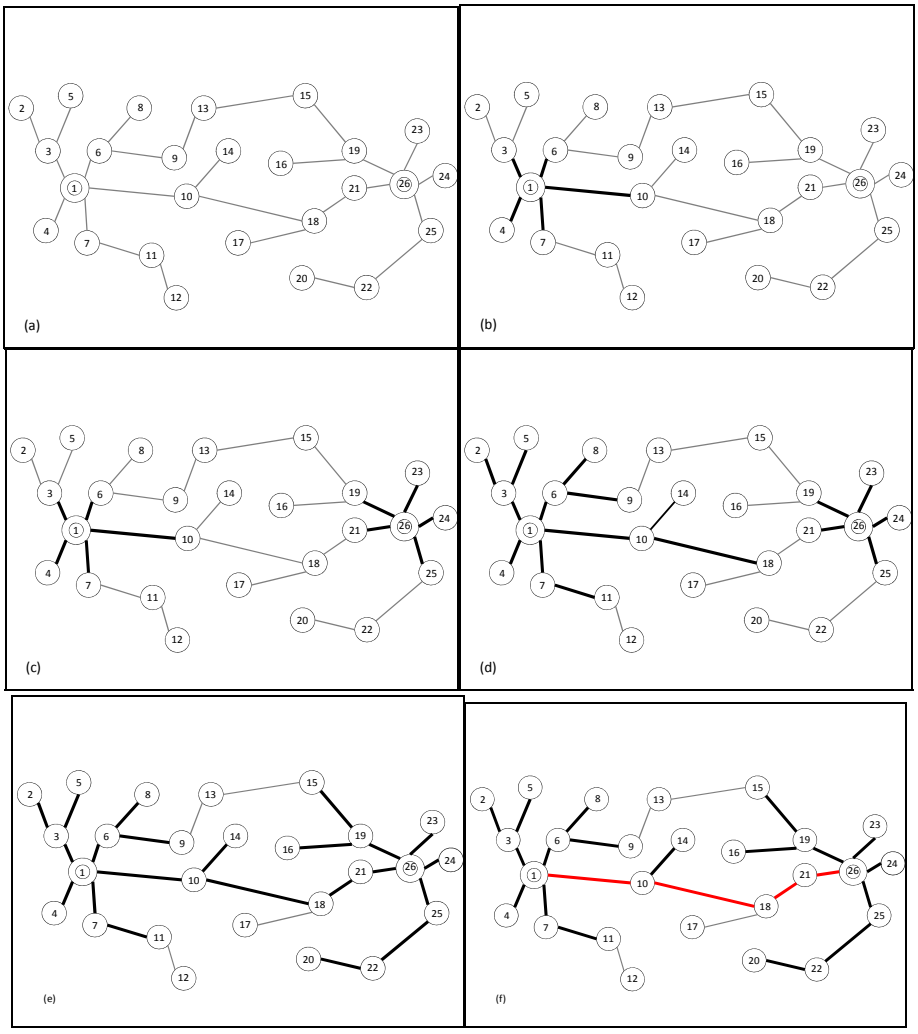


Figure 2. Example of the proposed algorithm.

4. Result and Discussion

Figure 3 shows the general statistics of the coauthorship in IR. For the multi-authored papers, each author was assigned full credit. The ratio of single-authored paper decreases from 69.53% in 1956-1980 to 15.48% in 2001-2008. The dominant coauthorship pattern changes from single author per paper in 1956-1990 to two/three authors per paper in 1991-2008. The largest number of authors per paper increases from 6 to 22.

Table 3. General statistics of coauthorship in IR

	1956-1980	1981-1990	1991-2000	2001-2008
% of Single authored papers	69.53%	62.86%	29.96%	15.48%
% of coauthored papers (2 authors)	18.05%	20.95%	33.04%	31.00%
% of coauthored papers (3 authors)	7.39%	9.80%	20.78%	26.41%
% of coauthored papers (4 authors)	3.20%	3.83%	8.85%	15.05%
% of coauthored papers (>4 authors)	1.83%	2.56%	7.38%	12.05%

In each period, five topics were extracted using the ACT model. The top 20 authors with high probability in each topic were selected as productive authors. So a total of 100 productive authors were selected and paired together within and across topics. The path-finding algorithm was applied to identify the shortest path between any given pair in the coauthorship network of each phase. The collaboration strength is used to measure the length of the shortest path (i.e. also called geodesic in graph theory), where the shorter the length, the stronger the collaboration strength will be (Newman, 2004a). Since coauthorship networks are social networks, six degrees of separation (i.e., one person is only six steps away from another person, or there are no more than six persons in between any given two persons in the world) was applied to categorize the strength of collaboration (Newman, Barabási, & Duncan, 2006): direct collaboration (e.g., author A and author B co-authored papers directly), indirect collaboration (e.g., there are six or less nodes in the shortest path of author A and B), loose collaboration (e.g., there are more than six nodes in the shortest path of author A and B), and no collaboration (e.g., a path between author A and B does not exist).

It is same for the citation strength. The top 100 highly cited authors were selected and paired together. The path-finding algorithm was applied to identify the shortest path of any given pair in the citation network of each phase. The citation strength is used to measure the length of the shortest path, where the shorter the length, the stronger citation strength will be. Since citation networks indicate the flow of influence in scholarly communications, three degrees of influence (where one person can be influenced by other person who is not more than three steps away) were applied to categorize citation strength (Christakis & Fowler, 2009): direct citation (e.g., author A directly cited author B), indirect citation (e.g., there are three or less nodes in the shortest path of author A and B), loose citation (e.g., there are more than three nodes in the shortest path of author A and B), and no citation (e.g., there is no path between author A and author B). Here the number of nodes in the shortest path of author A and B does not include the starting nodes (i.e., author A) and ending nodes (i.e., author B).

Productive authors' collaboration strength within topics

Table 4 shows the collaboration strength of the top 100 most productive authors sharing similar research interests which were extracted by the ACT model. In 1956-1980, 99.19% never collaborated and only 3.51% collaborated directly (the shortest path length is 0) within the topic of Medical IR. In 1981-1990,

there was not much change and three topics had a few direct collaborations. In contrast, things changed dramatically after 1990. In 1990-2000, the non-collaboration ratio dropped from 99.68% to 87.24% and nearly 50% of researchers collaborated with each other on the topic of Database and Query Processing. There were a few direct and indirect collaborations in each topic. After 2000, things changed dramatically again. Nearly half of the top 100 productive authors collaborated on various topics, with the topic of Data Storage and Evaluation being the highest (90%) and Online IR the lowest (10%). During this period, collaboration was more indirect (36.84%) and loose (11.05%). Overall, the collaboration strength within topics increased from 0.81% in 1956-1980 to 49.68% in 2001-2008. Figure 3 summarizes the overall collaboration strength of productive authors within topics. As each period has direct collaborations, it seems that productive authors tend to directly collaborate with colleagues sharing the same research interests. According to six degrees of separation in social networks (Newman, Barabási, & Duncan, 2006), Figure 3 shows that the percentage of loose collaboration within topics increased over time from 0% in 1956-1980 to 11.05% in 2001-2008. This indicates that the current coauthorship network in IR has more than six degrees of separation.

Table 4. Collaboration strength of productive authors within topics

1956-1980	**Topic 1: Thesaurus and Chemical IR**	**Topic 2: Data Storage and Evaluation**	**Topic 3: Online IR**	**Topic 4: Medical IR**	**Topic 5: IR theory and Patent**	**Average**
No Collaboration	99.47%	100%	100%	96.49%	100%	99.19%
Direct Collaboration (Shortest path=0)	0	0	0	3.51%	0	0.7%
Indirect Collaboration (Shortest path<=6)	0.53%	0	0	0	0	0.11%
Loose Collaboration (Shortest path>6)	0	0	0	0	0	0
Longest shortest path	2	NA	NA	0	NA	NA
1981-1990	**Topic 1: Automatic IR System**	**Topic 2: Online IR**	**Topic 3: Digital Library**	**Topic 4: Database and Query Processing**	**Topic 5: Evaluation**	**Average**
No Collaboration	100%	99.47%	100%	99.47%	99.47%	99.68%
Direct Collaboration (Shortest path=0)	0	0.53%	0	0.53%	0.53%	0.32%
Indirect Collaboration (Shortest path<=6)	0	0	0	0	0	0
Loose Collaboration (Shortest path>6)	0	0	0	0	0	0
Longest shortest path	NA	0	NA	0	0	NA
1991-2000	**Topic 1: Web IR**	**Topic 2: Multimedia IR**	**Topic 3: Evaluation**	**Topic 4: Medical IR**	**Topic 5: Database and Query Processing**	**Average**
No Collaboration	97.66%	91.58%	95.91%	99.47%	51.58%	87.24%
Direct Collaboration (Shortest path=0)	0.58%	1.05%	2.92%	0.53%	2.63%	1.54%
Indirect Collaboration (Shortest path<=6)	0.58%	4.21%	1.17%	0	36.32%	8.46%
Loose Collaboration (Shortest path>6)	1.17%	3.16%	0	0	9.47%	2.76%
Longest shortest path	8	14	1	0	11	6.8
2001-2008	**Topic 1: Thesaurus and Chemical IR**	**Topic 2: Data Storage and Evaluation**	**Topic 3: Online IR**	**Topic 4: Medical IR**	**Topic 5: IR theory and Patent**	**Average**
No Collaboration	28.42%	10%	90%	71.05%	52.11%	50.32%
Direct Collaboration (Shortest path=0)	1.58%	1.58%	2.11%	1.05%	2.63%	1.79%
Indirect Collaboration (Shortest path<=6)	60%	54.74%	4.74%	21.05%	43.68%	36.84%
Loose Collaboration (Shortest path>6)	10%	33.68%	3.16%	6.84%	1.58%	11.05%
Longest shortest path	9	11	7	8	7	8.4

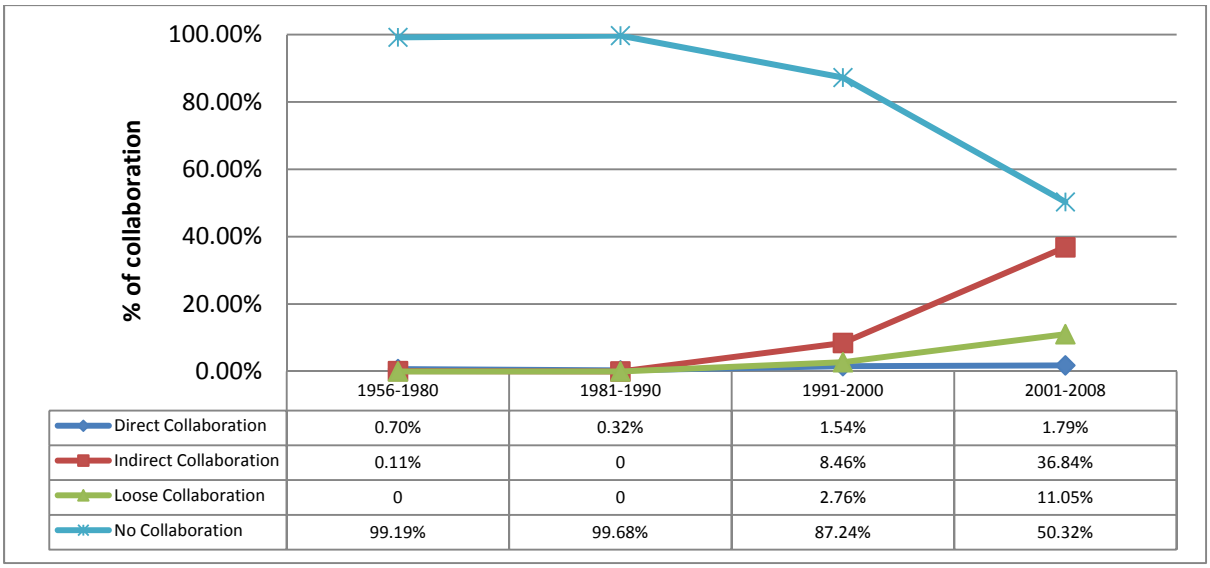


Figure 3. Summary of the collaboration strength of productive authors within topics.
(Note: Direct collaboration: shortest path=0; indirect collaboration: shortest path<=6; loose collaboration: shortest path>6; and no collaboration.)

Productive authors' collaboration strength across topics

Table 5 shows the collaboration strength of productive authors across different topics in each period, which is quite similar to the collaboration strength within topics. Before 1990, 99.92% never collaborated. In 1991-2000, people from the topic of Multimedia IR and Database and the topic of Query Processing started to collaborate (22.11%), while the rest did not. After 2000, nearly 50% of productive authors collaborated with colleagues having different topics, with the collaboration of the topic of Thesaurus and Chemical IR and the topic of Data Storage and Evaluation being the highest (80.75%), and the topic of Online IR and the topic of Medical IR the lowest (16.5%). Loose collaboration was up to 13.83% with the longest shortest path reaching 12. Overall, the collaboration strength across topics increased from 0.08% in 1956-1980 to 43.6% in 2001-2008. Figure 4 summarizes the collaboration strength of productive authors across topics. In general, direct collaboration remains consistently rare during these four periods. It seems that productive authors do not generally collaborate directly with colleagues having different research topics, and instead collaborate indirectly via other shared collaborators. Figure 4 shows that the percentage of loose collaboration across topics increased from 0% in 1956-1980 to 13.83% in 2001-2008. As seen in the collaboration within topics, this confirms that the current coauthorship network in IR has more than six degrees of separation.

Table 5. Collaboration strength of productive authors across topics

1956-1980	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Collaboration	100%	100%	100%	100%	99.47%	100%	99.72%	100%	100%	100%	99.92%
Direct Collaboration (Shortest path=0)	0	0	0	0	0	0	0	0	0	0	0
Indirect Collaboration (Shortest path<=6)	0	0	0	0	0.53%	0	0.28%	0	0	0	0.08%
Loose Collaboration (Shortest path>6)	0	0	0	0	0	0	0	0	0	0	0
Longest shortest path	NA	NA	NA	NA	1	NA	1	NA	NA	NA	NA
1981-1990	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Collaboration	100%	100%	100%	100%	100%	100%	100%	100%	100%	99.25%	99.93%

Direct Collaboration (Shortest path=0)	0	0	0	0	0	0	0	0	0	0.5%	0.05%
Indirect Collaboration (Shortest path<=6)	0	0	0	0	0	0	0	0	0	0.25%	0.03%
Loose Collaboration (Shortest path>6)	0	0	0	0	0	0	0	0	0	0	0
Longest shortest path	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
1991-2000	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Collaboration	95%	97.78%	100%	88.95%	98.95%	100%	77.89%	99.75%	100%	100%	95.83%
Direct Collaboration (Shortest path=0)	0	0	0	0	0	0	0	0	0	0	0
Indirect Collaboration (Shortest path<=6)	3.16%	2.22%	0	3.95%	1.05%	0	9.47%	0.25%	0	0	2.01%
Loose Collaboration (Shortest path>6)	1.84%	0	0	7.11%	0	0	12.63%	0	0	0	2.16%
Longest shortest path	13	3	NA	11	4	NA	15	1	NA	NA	NA
2001-2008	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Collaboration	19.25%	74.5%	53.25%	40.5%	71.5%	47.75%	33.5%	83.5%	79%	61.25%	56.40%
Direct Collaboration (Shortest path=0)	0	0.25%	0	0.25%	0	0	0.5%	0	0	0.25%	0.13%
Indirect Collaboration (Shortest path<=6)	51.5%	17.5%	35.25%	43.75%	15.75%	30.75%	41.25%	13%	17%	30.75%	29.65%
Loose Collaboration (Shortest path>6)	29.25%	7.75%	11.5%	15.5%	12.75%	21.5%	24.75%	3.5%	4%	7.75%	13.83%
Longest shortest path	12	10	10	10	11	12	10	9	8	8	10

Note: Here T1-T4 corresponds to the Topic 1-Topi 4 in Table 4.

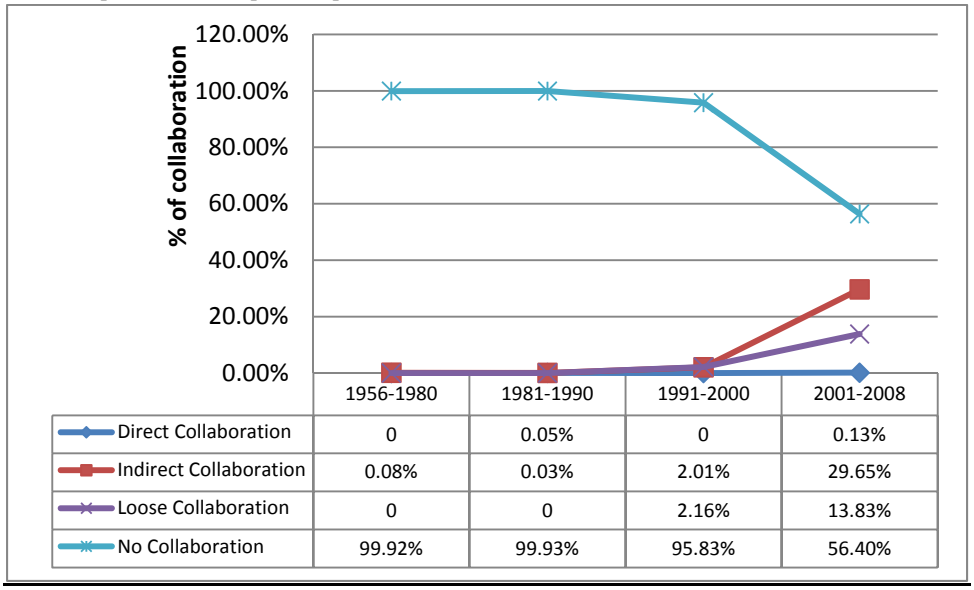


Figure 4. Summary of the collaboration strength of productive authors across topics

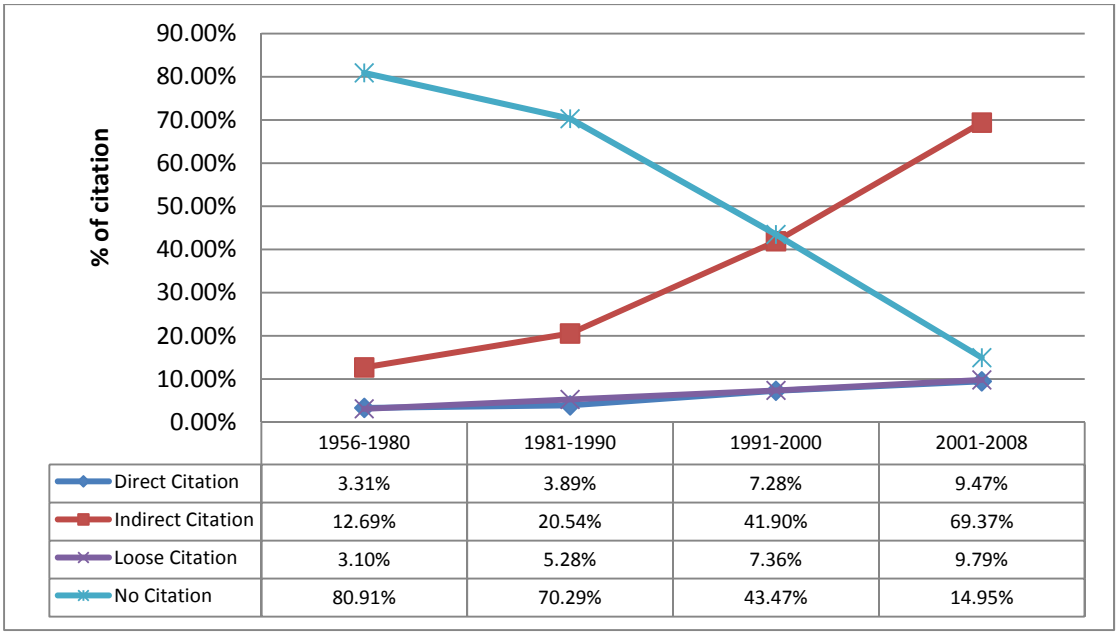
Productive authors' citation strength within topics

Table 6 shows the citation strength of productive authors within topics, which is different compared with their corresponding collaboration strengths. In 1956-1980, 57.52% of authors sharing the topic of Data Storage and Evaluation cited each other and their citation strength spans from direct citation (8.5%) to loose citation (9.83%). In 1980-1991, 84.74% of authors in the topic of Evaluation cited each other directly (13.16%) and indirectly (57.37%). In 1991-2000, 56.53% of authors cited each other within topics, and 86.84% of authors from the topic of Database and Query Processing cited each other directly and indirectly. After 2000, 85.05% of productive authors cited each other within topics, with direct and

indirect citations dominating. During these four phases, these researchers continued to cite each other, and the citation strength increased from 19.01% in 1956-1980 to 85.05% in 2001-2008, the majority being either direct or indirect citations. It seems that productive authors like to directly/indirectly cite authors with the same research interests. Figure 5 summarizes the citation strength of productive authors within topics. The citation graphs demonstrate the influence and knowledge transfer in scholarly communication. According to the three degrees of influence (Christakis & Fowler, 2009), the percentage of loose citation within topics increased slightly from 3.10% in 1956-1980 to 9.79% in 2001-2008.

Table 6. Citation strength of productive authors within topics

1956-1980	**Topic 1: Thesaurus and Chemical IR**	**Topic 2: Data Storage and Evaluation**	**Topic 3: Online IR**	**Topic 4: Medical IR**	**Topic 5: IR theory and Patent**	**Average**
No Citation	97.37%	42.48%	75.79%	98.83%	90.06%	80.91%
Direct Citation (Shortest path=0)	2.63%	8.5%	3.68%	0.58%	1.17%	3.31%
Indirect Citation (Shortest path<=3)	0	39.22%	18.95%	0	5.26%	12.69%
Loose Citation (Shortest path>3)	0	9.83%	1.58%	0.58%	3.51%	3.1%
Longest shortest path	0	6	5	6	7	4.8
1981-1990	**Topic 1: Automatic IR system**	**Topic 2: Online IR**	**Topic 3: Digital Library**	**Topic 4: Database and Query Processing**	**Topic 5: Evaluation**	**Average**
No Citation	99.47%	72.63%	88.30%	75.79%	15.26%	70.29%
Direct Citation (Shortest path=0)	0	2.63%	0	3.68%	13.16%	3.89%
Indirect Citation (Shortest path<=3)	0.53%	22.63%	5.85%	16.32%	57.37%	20.54%
Loose Citation (Shortest path>3)	0	2.11%	5.85%	4.21%	14.21%	5.28%
Longest shortest path	1	6	8	7	7	5.8
1991-2000	**Topic 1: Web IR**	**Topic 2: Multimedia IR**	**Topic 3: Evaluation**	**Topic 4: Medical IR**	**Topic 5: Database and Query Processing**	**Average**
No Citation	63.74%	35.79%	25.73%	78.95%	13.16%	43.47%
Direct Citation (Shortest path=0)	3.51%	7.37%	12.87%	0	12.63%	7.28%
Indirect Citation (Shortest path<=3)	23.98%	48.42%	60.24%	5.79%	71.05%	41.9%
Loose Citation (Shortest path>3)	8.77%	8.42%	1.17%	15.26%	3.16%	7.36%
Longest shortest path	6	6	4	6	4	5.2
2001-2008	**Topic 1: Thesaurus and Chemical IR**	**Topic 2: Data Storage and Evaluation**	**Topic 3: Online IR**	**Topic 4: Medical IR**	**Topic 5: IR theory and Patent**	**Average**
No Citation	4.74%	11.58%	20.53%	18.42%	19.47%	14.95%
Direct Citation (Shortest path=0)	5.26%	6.84%	7.37%	15.26%	12.63%	9.47%
Indirect Citation (Shortest path<=3)	89.47%	72.11%	51.58%	65.79%	67.89%	69.37%
Loose Citation (Shortest path>3)	0.53%	9.47%	20.53%	0.53%	17.89%	9.79%
Longest shortest path	4	4	6	4	3	4.2



	1956-1980	1981-1990	1991-2000	2001-2008
Direct Citation	3.31%	3.89%	7.28%	9.47%
Indirect Citation	12.69%	20.54%	41.90%	69.37%
Loose Citation	3.10%	5.28%	7.36%	9.79%
No Citation	80.91%	70.29%	43.47%	14.95%

Figure 5. Summary of the citation strength of productive authors within topics

Note: Direct Citation: shortest path=0, Indirect Citation: shortest path<=3, Loose Citation: shortest path>3, and No Citation.

Productive authors' citation strength across topics

Table 7 shows the citation strength of productive authors across topics. During the first two periods, 24% of productive authors cited colleagues with different interests and citation strength spread evenly from direct, indirect, to loose citation. The average length of longest shortest paths was 8. In 1991-2000, 59.11% of authors cited each other directly (1.2%), indirectly (49.39%), and loosely (8.51%). In 2001-2008, the majority of them (86.65%) cited each other, with indirect citation being the highest (81.15%). Figure 6 summarizes the citation strength of productive authors across topics. According to the three degrees of influence, the percentage of loose citation across topics decreased from 18.35% in 1956-1980 to 4.05% in 2001-2008. This indicates that the current citation network in IR has fewer than three degrees of influence.

Table 7. Citation strength of productive authors across topics

1956-1980	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Citation	33.61%	60.5%	68.68%	73.16%	70%	77.89%	78.95%	81.58%	82.46%	98.06%	72.49%
Direct Citation (Shortest path=0)	1.67%	0.5%	0.79%	0	0.83%	0.26%	2.92%	0.53%	2.34%	0	0.98%
Indirect Citation (Shortest path<=3)	33.89%	16.25%	5.79%	15.79%	23.61%	4.74%	12.57%	6.05%	11.7%	1.39%	13.18%
Loose Citation (Shortest path>3)	30.83%	22.75%	24.74%	11.05%	5.56%	17.11%	55.56%	11.84%	3.51%	0.55%	18.35%
Longest shortest path	8	7	9	11	6	9	11	7	9	5	8.2
1981-1990	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Citation	92%	96.05%	94.75%	87.25%	85.53%	80.75%	53.25%	83.42%	59.74%	32%	76.47%
Direct Citation (Shortest path=0)	0	0	0	0.5%	0.53%	0.25%	4.25%	0.26%	0.79%	1.75%	0.83%
Indirect Citation (Shortest path<=3)	4.25%	1.58%	1.75%	8%	8.68%	14.25%	35.75%	7.11%	27.11%	30.5%	13.90%
Loose Citation (Shortest path>3)	3.75%	2.37%	3.5%	4.25%	5.26%	4.75%	6.75%	9.21%	12.37%	35.75%	8.80%
Longest shortest path	8	10	6	9	8	5	7	7	8	11	7.9
1991-2000	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Citation	36.84%	29.09%	74.74%	24.21%	24.21%	73%	19%	71.58%	14.74%	41.5%	40.89%
Direct Citation	2.37%	4.43%	0.53%	1.05%	1.32%	0	0.75%	0.53%	1.05%	0	1.2%

(Shortest path=0)											
Indirect Citation (Shortest path<=3)	50.53%	65.93%	16.32%	68.95%	74.21%	16.75%	75.5%	20%	69.74%	36%	49.39%
Loose Citation (Shortest path>3)	10.26%	0.55%	8.42%	5.79%	0.26%	10.25%	4.75%	7.89%	14.47%	22.5%	8.51%
Longest shortest path	7	4	5	4	4	5	4	5	5	6	6
2001-2008	**T1-T2**	**T1-T3**	**T1-T4**	**T1-T5**	**T2-T3**	**T2-T4**	**T2-T5**	**T3-T4**	**T3-T5**	**T4-T5**	**Average**
No Citation	10%	20%	10%	10%	20%	10%	10%	14.5%	14.5%	14.5%	13.35%
Direct Citation (Shortest path=0)	0	0.25%	2%	2.5%	0	0.25%	2.5%	1%	0.75%	5.25%	1.45%
Indirect Citation (Shortest path<=3)	82.75%	77.25%	87.5%	87.5%	75%	86.5%	87.5%	71.5%	75.75%	80.25%	81.15%
Loose Citation (Shortest path>3)	7.25%	2.5%	0.5%	0	5%	3.25%	0	13%	9%	0	4.05%
Longest shortest path	5	4	4	3	4	4	3	6	6	3	4.2

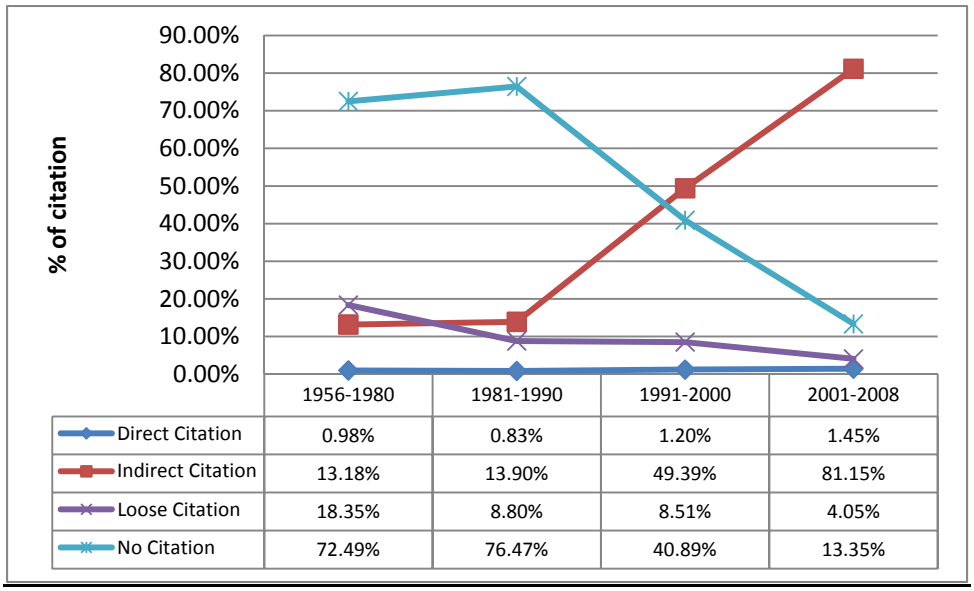


Figure 6. Summary of the citation strength of productive authors across topics

5. Highly cited authors

In each phase, the top 100 highly cited authors were selected and paired together, and the path-finding algorithm was used to calculate their collaboration and citation strengths. Since the ACT model cannot extract the topical distribution for cited authors, it is not possible to calculate the collaboration and citation strength of highly cited authors within or across topics.

Highly cited authors' collaboration strength

Table 8 shows the collaboration strength of highly cited authors. The majority did not coauthor before 2000. After 2000, however, 48.48% started to write papers together in indirect ways. Direct collaboration was only 0.38%, and 29.11% of collaborations have three or less colleagues in the shortest paths. The longest shortest path was 15, which means that author A and author B were connected via 15 different authors. Figure 7 summarizes the collaboration strength of highly cited authors. The collaboration percentage increased from 0.37% in 1956-1980 to 48.48% in 2001-2008. As Moody (2004, p217) pointed out, funding requirements, the rise of large-scale data collection and analysis efforts, the subtle division

between specialty and training, and scientific laboring distribution explain the increase in coauthorship over time. The collaboration strength of highly cited authors further confirmed that the current coauthorship network in IR has more than six degrees of separation.

Table 8. Collaboration strength of highly cited authors

	1956-1980	1981-1990	1991-2000	2001-2008
No Collaboration	99.63%	98.92%	86.38%	51.52%
Direct Collaboration (Shortest path=0)	0.22%	0.83%	0.55%	0.38%
Indirect Collaboration (Shortest path<=6)	1.44%	0.25%	5.02%	29.11%
Loose Collaboration (Shortest path>6)	0	0	8.05%	18.99%
Longest shortest path	1	2	19	15

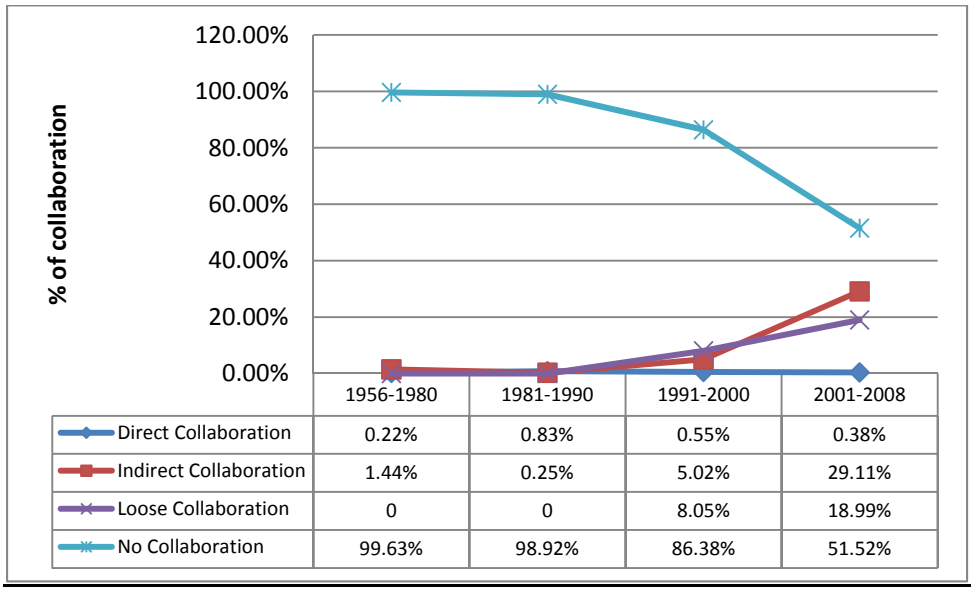


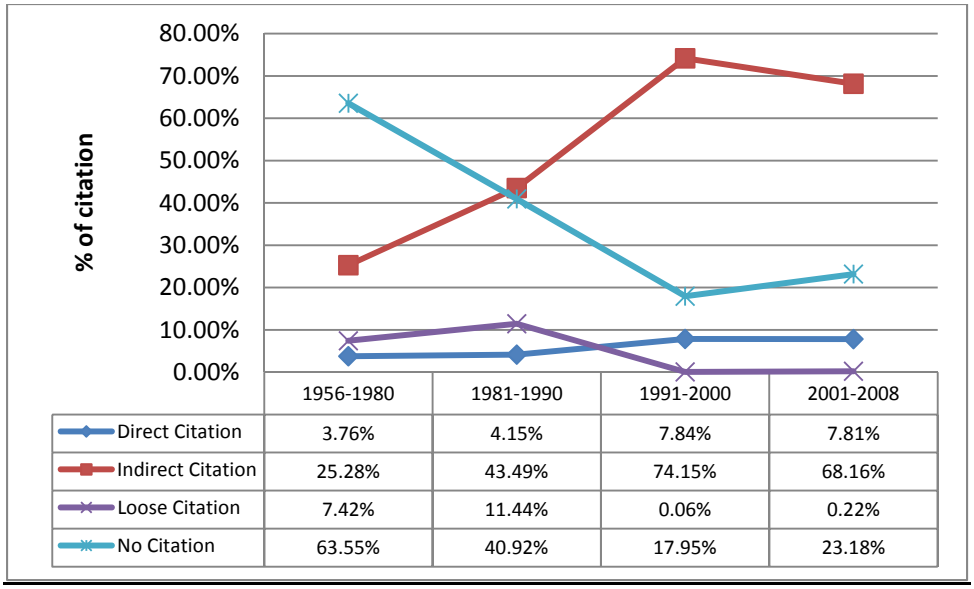
Figure 7. Summary of the collaboration strength of highly cited authors

Highly cited authors' citation strength

Table 9 shows the citation strength of highly cited authors in different phases. In 1956-1980, less than 40% cited each other. After that, 72.64% cited each other directly (6.6%), indirectly (61.93%), and loosely (3.91%). Figure 8 summarizes the citation strength of highly cited authors. It seems that highly cited authors cited each other more directly. They did not tend to coauthor with each other and their collaboration strength was very loose. The citation strength of highly cited authors further confirms that the current citation network in IR has less than three degrees of influence.

Table 9. Citation strength of highly cited authors

	1956-1980	1981-1990	1991-2000	2001-2008
No Citation	63.55%	40.92%	17.95%	23.81%
Direct Citation (Shortest path=0)	3.76%	4.15%	7.84%	7.81%
Indirect Citation (Shortest path<=3)	25.28%	43.49%	74.15%	68.16%
Loose Citation (Shortest path>3)	7.42%	11.44%	0.06%	0.22%
Longest shortest path	12	10	4	4



	1956-1980	1981-1990	1991-2000	2001-2008
Direct Citation	3.76%	4.15%	7.84%	7.81%
Indirect Citation	25.28%	43.49%	74.15%	68.16%
Loose Citation	7.42%	11.44%	0.06%	0.22%
No Citation	63.55%	40.92%	17.95%	23.18%

Figure 8. Summary of the citation strength of highly cited authors

6. Salton Number

Following the idea of the Erdős number in mathematics (http://www.oakland.edu/enp/) and the Bacon number for celebrities (http://oracleofbacon.org/), this paper proposed the Salton number for the IR community, which is the geodesic distance between one IR researcher and Salton in the IR coauthorship network. Salton was the leading researcher in IR and played an important role in establishing and developing the IR field. In his life time, he published more than 150 papers and co-authored with a wide range of colleagues, students and other collaborators. In mathematics, the average Erdős number is 4.7 and the maximum is 15 (Grossman, 2002). In this section, a coauthorship network was formed based on the whole IR dataset covering the period from 1956 to 2008. The top 100 highly cited authors in 1956-2008 were selected to calculate their Salton number. Among them, 14.13% had never coauthored with Salton (e.g., D. R. Swanson, T. Kohonen, S. P. Harter); 5.43% of them had directly coauthored with Salton (e.g., C. T. Yu, E. A. Fox, C. Buckley); 76.09% had indirect collaboration with Salton (e.g., N. J. Belkin [1], W. B. Croft [1], K.S. Jones [1]); and 4.35% had loose collaboration (e.g., F. W. Lancaster [6], M. Stonebraker [7], T. Imielinski [8]). Numbers in brackets show the Salton number which is the number of nodes in the shortest path without counting the starting and ending nodes. The average Salton number is 4.64, which is close to the Erdős number in mathematics, and the largest number is 8. Figure 9 shows some shortest paths between Salton and C. L. Borgman.

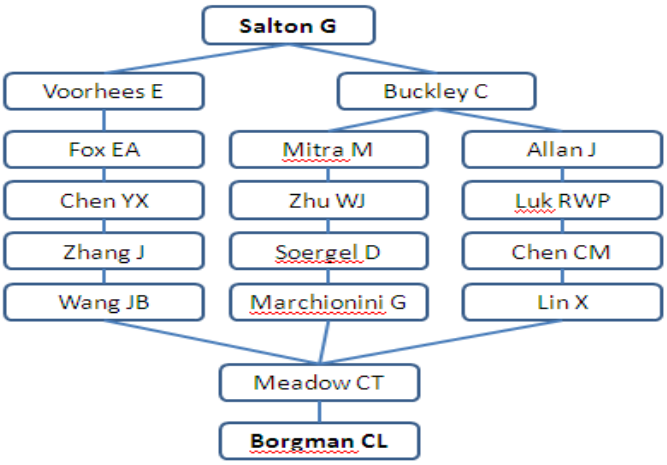


Figure 9. The shortest paths between Salton and C. L. Borgman

7. Conclusion

Bibliometrically, analyzing the relationship between two specific nodes is ignored, as ranking individual nodes dominates state-of-the-art research. But identifying the relationship between two specific nodes can reveal scholarly communication patterns (i.e., collaboration, knowledge diffusion, or authority transfer) with finer granularity. Topic modeling algorithms (e.g., LDA) can calculate the topical similarity between two nodes, but they cannot capture their relationship at a path level for a graph. A majority of current complex network analyses focus on identifying network connectivity and their distribution patterns (e.g., a power law, preferential attachment, scale-free network, or small-world network). They usually focus on the macro-level features of complex networks and do not drill down to the micro-level features of individual nodes or their paths/connected subgraphs (Faloutsos, McCurley, & Tomkins, 2004). This paper combined the method of topic modeling and network analysis to address the micro-level features of scientific collaboration and endorsement.

Information retrieval was taken as the test area, and the top 100 productive authors and highly cited authors were selected and their collaboration and citation strengths were investigated. The Salton number was introduced and the average Salton number is 4.64. The combination of LDA and path-finding algorithms enables us to address the following questions:

- Will productive authors collaborate with/cite people sharing same research interests? For the collaboration part, we found that before 1990, productive authors seldom collaborated with colleagues who have the same research interests. After 1990, their collaboration strengths increased from 12.76% to 49.68%. Productive authors tend to directly coauthor with colleagues sharing the same research interests, which is viewed as homophily in social networks (McPherson, Smith-Lovin and Cook, 2001). For the citation part, we found that productive authors consistently cite/endorse colleagues with the same research interests and their citation strengths increase from 19.09% to 85.05%, the majority of which are either direct or indirect citation. It seems that productive authors tend to directly/indirectly cite authors with the same research interests;
- Will productive authors with different research interests collaborate with/cite each other? For the collaboration part, we found that few productive authors collaborated with colleagues having different research interests before 2000. After 2000, nearly 50% of them collaborated but in an

indirect or loose manner. It seems that productive authors did not collaborate directly with colleagues having different research topics, but rather indirectly via other shared collaborators. For the citation part, we found that productive authors often cited colleagues with different research interests and their citation strengths increased from 27.51% to 86.65%. The majority of citations belong to the category of indirect citation. As found in the citation strength of productive authors within topics, these authors tend to closely cite colleagues from different research fields;

- Will highly cited authors collaborate with/cite each other? We found that the majority of highly cited authors did not collaborate with each other until 2000. After that, 48.48% of them started to coauthor with each other indirectly, while they repeatedly cited each other through these years in a direct and indirect manner; and
- Do scholarly networks in IR follow six degrees of separation or three degrees of influence? We found that the current coauthorship network in IR has more than six degrees of separation and the current citation network in IR has less than three degrees of influence. In other words, the IR researchers' collaboration strength is going beyond six persons in their shortest paths and they cite colleagues with no more than three persons away.

Citing behavior implies endorsement, confers authority, and traces provenance. It provides a unique way of allowing the network to play a role in determining researcher standing, impact and influence (Kleinberg, 1998). Citation is taken as a carrier of authority and weighted citations correspond to the strength of different endorsements. The path between two articles built through the citing activities portrays the provenance of authority transfer flow, which can be interpreted as scholarly trust. However, one can trust people with a limited or greater level of responsibility. The same goes for knowledge, influence, and impact transfer on scholarly networks. Setting up a notion of authority with different contexts is essential. So the influence flow from nodes to nodes within a scholarly network should be differentiated across different authorities, rather than merely considering every node equally or indiscriminately. In the future, we would like to utilize the topic modeling and path-finding algorithm to trace scholarly provenance and to establish scholarly trust.

8. Acknowledgement

I would like to thank Jie Tang for sharing the ACT and path-finding codes, Yuyin Sun for his programming help, and Bing He, Stasa Milojevic, Cassidy Sugimoto, Blaise Cronin, Ronald Rousseau, and two anonymous reveiwers for their insightful comments.

9. References

Ahuja, R.K., Magnanti, T.L., & Orlin, J.B. (1993). *Network flows: theory, algorithms, and applications*. Prentice Hall, Upper Saddle River, New Jersey.

An, Y., Janssen, J. C. M., & Milios, E.E. (2004). Characterizing the citation graph as a self-organizing networked information space. *Knowledge and Information Systems*, 6, 664-678.

Barabási, A.L. (2002). *Linked: How Everything Is Connected to Everything Else and What it Means for Business, Science, and Everyday Life*, 2002. ISBN 0-452-28439-2

Beaver, D. (2001). Reflections on scientific collaborations (and its study): Past, present and prospective, *Scientometrics,* 52, 365–377.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1033.

Bollen, J., Rodriguez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69(3), 669-687.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the seventh international conference on World Wide Web*, 107-117.

Buntine, W.L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2, 159-225.

Christakis, N.A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. New York: Little Brown.

Cronin, B. (2001). Hyperauthorship: a postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, 52*,* 558–569.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Co-authorship and sub-authorship collaboration in the twentieth century as manifested in the scholarly literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology,* 54, 855–871.

Ding, Y., Foo, S., & Chowdhury, G. (1998). A Bibliometric Analysis of Collaboration in the Field of Information Retrieval. *International Information & Library Review*, 30(4): 367-376.

Ding, Y., Chowdhury, G., & Foo, S. (1999). Mapping Intellectual Structure of Information Retrieval: An Author Cocitation Analysis, 1987-1997 *Journal of Information Science*, 25(1): 67-78.

Ding, Y., Chowdhury, G., & Foo, S. (2000). Journal as Markers of Intellectual Space: Journal Co-citation Analysis of Information Retrieval Area, 1987-1997. *Scientometrics*, 47(1): 55-73.

Ding, Y.,Chowdhury, G., & Foo, S. (2000a). Incorporating the Results of Co-word Analyses to Increase Search Variety for Information Retrieval. *Journal of Information Science*, 26(6): 429-452.

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press.

Faloutsos, C., McCurley, K. S., & Tomkins, A. (2004). Fast discovery of connection subgraphs. *The tenth ACM SIGKDD Conference*, Seattle, WA, USA, August 22-25, 2004.

Freeman, L. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40, 35-41.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178, 471-479.

Giles, C.L., & Councill, I.G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), 17599-17604.

Glänzel, W. (2001). National Characteristics in International Scientific Co-authorship, *Scientometrics*, 51, 69–115.

Glanzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In H.F. Moed, W. Glanzel & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems*, pp. 257-276.

Goh, K.-I., Oh, E.S., Jeong, H., Kahng, B., & Kim, D. (2002). Classification of scale free network, *Proceedings of the National Academy of Sciences of the USA (PNAS)*, 99**,** 12583-12588.

Goh, K.I., Oh, E., Kahng, B., & Kim, D. (2003). Betweenness centrality correlation in social networks. *Physical Review E*, 67, 017101.

Grossman, J. W. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numerantium*, 158, 202-212.

Hara, N., Solomon, P., Kim, S.L., & Sonnenwald, D. H. (2003). An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science*, 54(10), 952-965.

He, B., Sankaranarayanan, M. Ding, Y., Tang, J., Wang, H., Wild, D., Chen, B., Sun, Y., Sugimoto, C., Wu, Y., & Qiu, J. (2010 submitted). Semantic association and topic mining in linked life data.

Jaffe, A.B., Trajtenberg, M., & Fogarty, M.S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *The American Economic Review*, 90(2), 215-218.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration?, *Research Policy*, **26**, 1–18.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *Proceedings of the ACM SIAM Symposium on Discrete Algorithm*.

Kretschmer, H. (1994). Coauthorship networks of invisible colleges and institutional communities. *Scientometrics*, 30, 363–369.

Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks and visibility on the Web. *Scientometrics, 60*(3), 409-420.

Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, 11, 3–15.

Leydesdorff, L., & Etzkowitz, H. (1996). Emergence of a Triple Helix of University-Industry-Government Relations, *Science and Public Policy*, 23, 279–286.

Li, X., Chen, H., Zhang, Z., & Li, J. (2007). Automatic patent classification using citation network information: An experimental study in nanotechnology. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, p419-427, Vancouver, BC, Canada.

Liu, X., Bollen, J. Nelson, M. L., & Sompel, H. V. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management,* 41, 1462-1480.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30, 249-272.

McClintock, M. D. (1998). The declining use of legal scholarship by courts: An empirical study. *Oklahoma Law Review*, 51, 659-696.

McPherson, M., Smith-Lovin, L., & Cook, M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415-444.

Milojevic, S. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, 61(7), 1410-1423.

Moody, J. (2004). The structure of social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2), 213-238.

Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the USA (PNAS)*, 101(suppl. 1), 5200-5205.

Newman, M. (2004a). Who is the best connected scientist? A study of scientific coauthorship networks. *Complex Networks*, 650, 337-370.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, 2010

Newman, M., Barabási, A. L., & Duncan J. W. (2006). *The Structure and Dynamics of Networks.* Princeton, NJ: Princeton University Press.

Price, D. deSolla (1966). *Little Science, Big Science*, Columbia Univ. Press, New York.

Rodriguez, M. A., & Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics, 2*(3), 195-201.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, p487-494, Banff, Canada.

Small, H. (1973). Co-citation in scientific literature: New measure of relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269.

Tang, J., Jin, R., & Zhang J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. *Proceedings of 2008 IEEE International Conference on Data Mining (ICDM2008)*, p1055-1060.

Vidgen, R., Henneberg, S., & Naude, P. (2007). What sort of community is the European Conference on Information Systems? A social network analysis 1993-2005. *European Journal of Information Systems, 16*(1), 5-19.

White, H.D., & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-172.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, 60(10), 2107-2118.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p334-342, September 9-13, 2001, New Orleans, LA, USA.

Zhang, P., & Koppaka, L. (2007). Semantics-based legal citation network. *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, p123-130, Stanford, CA.

Appendix

Topic and Author ranks in 2001-2008

Topic 1: Multimedia IR		**Topic 2: Database and Query Processing**		**Topic 3: Medical IR**		**Topic 4: Web IR and Digital library**		**Topic 5: IR Theory and Model**	
WORD	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**	**WORD**	**PROB**
image	0.063250	query	0.033203	database	0.010424	web	0.023822	document	0.014450
content-based	0.017681	data	0.025732	medical	0.007140	search	0.015858	text	0.010966
learning	0.008809	xml	0.019248	health	0.004982	digital	0.008366	query	0.009878
images	0.008667	processing	0.018614	clinical	0.004513	searching	0.006395	image	0.009587
relevance	0.008383	queries	0.016147	management	0.004325	knowledge	0.006001	relevance	0.008499
color	0.008312	databases	0.012764	search	0.004138	system	0.005764	fuzzy	0.008281
feedback	0.008312	database	0.009733	design	0.004138	query	0.005764	web	0.007991
video	0.007673	efficient	0.009451	study	0.003668	user	0.005528	model	0.006539
semantic	0.007389	web	0.009381	support	0.003575	model	0.005212	system	0.006321
similarity	0.007318	querying	0.008958	knowledge	0.003575	internet	0.004424	cross-language	0.006176
AUTHOR	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**	**AUTHOR**	**PROB**
HUANG TS	0.000572	LI JZ	0.000572	KOSTOFF RN	0.000321	THELWALL M	0.000628	CRESTANI F	0.000573
ZHANG HJ	0.000528	BRY F	0.000415	BALIS UJ	0.000283	YANG CC	0.000545	JONES GJF	0.000554
LU GJ	0.000409	KIM HJ	0.000371	EYSENBACH G	0.000257	SPINK A	0.000457	HERRERA-WIEDMA E	0.000548
LI J	0.000365	PAPADIAS D	0.000364	HAYNES RB	0.000251	JACSO P	0.000444	SAVOY J	0.000510
CHANG CC	0.000358	SUBIETA K	0.000358	NILSSON G	0.000238	FOURIE I	0.000425	LALMAS M	0.000510
IZQUIERDO E	0.000352	VAN DEN BUSSCHE J	0.000339	SHATKAY H	0.000218	CHEN HC	0.000393	JARVELIN K	0.000510
LASSKSONEN J	0.000327	TANIAR D	0.000327	WILCZYNSKI NL	0.000218	FORD N	0.000381	KANDO N	0.000466
BURKHARDT H	0.000308	GEERTS F	0.000327	SHYU CR	0.000218	XIE H	0.000368	CHEN SM	0.000403
LIU CJ	0.000308	SONG M	0.000320	WESTBROOK JI	0.000218	CHOWDHURY GG	0.000355	FUHR N	0.000397
ZIOU D	0.000302	CHUNG YD	0.000320	BABNETT GO	0.000212	HJORLAND B	0.000349	OUNIS I	0.000378