# Mining Enriched Contextual Information of Scientific Collaboration: A Meso Perspective

Bing He, Ying Ding, Chaoqun Ni
{binghe; dingying; chni}@indiana.edu
School of Library and Information Science
Indiana University Bloomington

## Abstract

Studying scientific collaboration using coauthorship networks has attracted much attention in recent years. How and in what context two authors collaborate remain among the major. Previous studies, however, have focused on either exploring the global topology of coauthorship networks (macro perspective) or ranking the impact of individual authors (micro perspective). Neither of them has provided information on the context of the collaboration between two specific authors, which may potentially imply rich socioeconomic, disciplinary, and institutional information on collaboration. Different from the macro-perspective and micro-perspective, this paper proposes a novel method (meso perspective) to analyze scientific collaboration, in which a contextual subgraph is extracted as the unit of analysis. A contextual subgraph is defined as a small subgraph of a large-scale coauthorship network that captures relationship and context between two coauthors. This method is applied to the field of library and information science (LIS). Topological properties of all the subgraphs in four time spans are investigated, including size, average degree, clustering coefficient, and network centralization. Results show that contextual subgprahs capture useful contextual information on two authors' collaboration.

## Introduction

The trend of scientific collaboration has become more and more prominent within and across different disciplines in the past decades. The idea that scientific research is moving from a personal, disciplinary-based, and location-restricted practice towards a collective, problem-oriented and geographical-distributed activity is well-accepted nowadays (Sonnenwald, 2007). Scientific collaboration advances professional development and increases the integration of knowledge. Access to expertise, facility, and connections from multiple sides is shared in scientific collaboration, providing a stronger whole than any individual side. In particular, it also enhances the visibility of aspiring young scientists (Beaver & Rosen, 1978, 1979). Therefore, in recent years, numerous institutional and governmental initiatives are intended to encourage collaboration among scientists, institutions, and countries. Coauthorship is an explicit and critical product of scientific collaboration, and has been used extensively to explore the patterns and potential of scientific collaboration and the impact of individual scholars. Many aspects of scientific collaboration, including the investigation of global topology of coauthorship networks or ranking the impact of individual authors, can be tracked by analyzing the coauthorship network. However, some questions, such as in what context do two specific coauthors actually collaborate and why, are still remained unanswered.

Previous studies of coauthorship networks can be generally categorized into two directions. One focuses on the global structure and evolution of coauthorship networks (a macro perspective) (Barabasi, et al., 2002; Moody, 2004; Newman, 2001a, 2001b; Leydesdorff & Wagner, 2008). The other emphasizes various indicators of the impact/prestige of individual researchers (a micro perspective). For example, different types of centrality and weighted PageRank are performed based on coauthorship networks (Borner, et al., 2005; Liu, et al., 2005; Yan & Ding, 2009; Yan et al., 2010). Although Integration of coauthorship networks on institutional and international levels can noticeably reflect the language and geographical factor in scientific collaboration, yet it loses some information and makes the personal-level factors invisible. They haven't shed light on how to characterize and contextualize the collaborative relationship of a coauthor pair.

In large-scale coauthorship networks, searching for the relationship between two specific coauthors (i.e., two people who have co-published a paper, which is referred to as "a coauthor pair" in the rest of the paper), usually yields the edge between them weighted by the number of coauthored papers (Figure 1 (a)). However, such edges disregard a large amount of contextual information between the co-author pair. First, a direct concern is that a single edge in coauthorship networks omits important information in the case of more than two researchers collaborating on one paper (i.e., multi-authorship). Second, a single edge in coauthorship networks cannot display the broad environment of collaboration, such as disciplinary, socioeconomic, institutional, and geographical factors (Sonnenwald, 2007). Instead, relevant authors, who are directly or indirectly involved in the collaboration of the coauthor pair, can imply rich contextual information. The subgraph formed by these relevant authors and the coauthor pair is referred to as the contextual subgraph characterized by the co-author pair (a meso perspective). Table 1 summarizes the macro, micro and meso perspectives of coauthorship networks.

**Table 1 A summary of macro, micro and meso level perspectives**

| | Measure | Characteristics |
|---|---|---|
| Macro-level | size, largest component, geodesic distance, degree distribution, clustering coefficient, k-core, and so forth | detecting the global pattern of scientific collaboration |
| Micro-level | degree centrality, closeness centrality, betweeness centrality, eigenvector centrality, and PageRank | identifying most collective authors and ranking impact of individual authors |
| Meso-level | number, size, and other topological properties of contextual subgraphs | characterizing and contextualizing collaborative relationships between coauthor pairs |

In order to address those questions, this paper defines a contextual subgraph that captures the link and context information for a coauthor pair. More specifically, the research questions addressed in this paper is "*in what context do two specific coauthors actually collaborate and why*". Here we provide an example of contextual subgraphs. As shown in Figure 1, assume that people want to find out how and in what context M. Thelwall and D. Wilkinson have collaborated. By looking at the edge between the two scientists (Figure 1a), people can only know that they have coauthored a certain number of papers. By contrast, the contextual subgraph (Figure 1b) shows seven more researchers involved in their collaboration. Through examining those researchers' affiliations, we found that M. Thelwall and D. Wilkinson are both faculty members of Statistical Cybermetrics Research Group at University of Wolverhampton, as are R. Binns, L. Price, and P. Musgrove. In addition, G. Harries, X.M. Li, and T. Page-Kennedy are faculty members in the same department with M. Thelwall and D. Wilkinson. Many papers coauthored by M. Thelwall and D. Wilkinson also involved those other nodes in the subgraph as coauthors (multiple authors). Contextual information supplied by the subgraph is more informative in helping us to understand these collaborations.

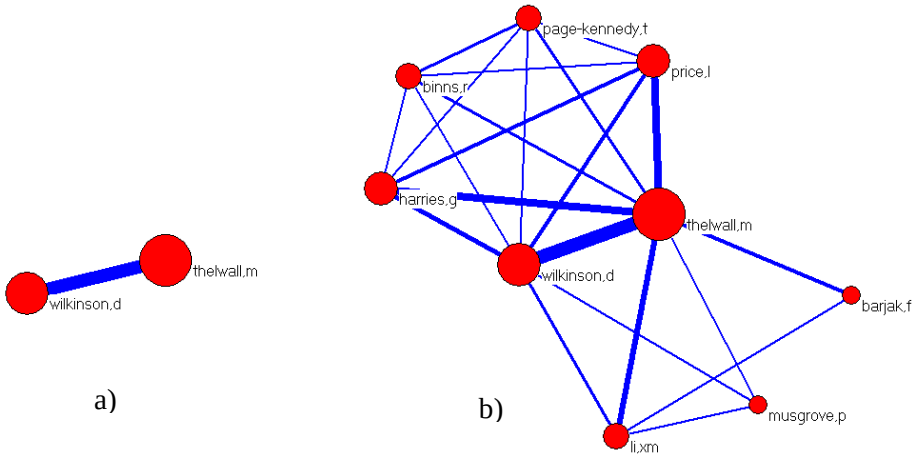


**Figure 1 (a) The edge between M. Thelwall and D. Wilkinson in coauthorship network; and 1 (b) the contextual subgraph of M. Thelwall and D. Wilkinson**

Taking contextual graph as the unit of analysis, statistical features of the topological properties of subgraphs can be investigated and correlated to other aspects of coauthorship (demographics, journal, institution, nations, mentorship, etc.) to uncover underlying mechanisms of scientific collaboration. In this paper, the method is applied to the coauthorship networks in LIS field. In addition to diachronically analyzing the topological properties of thousands of coauthor subgraphs, this paper also explores how topological properties of contextual subgraphs correlate with productivities and citations of coauthor pairs. This paper is organized as follows: section 2 states related works; section 3 elaborates on the methodology and the sample data; section 4 presents the results; and section 5 concludes the study.

## Related Work

Before 2000, studies of coauthorship networks focused on the validity of using coauthorship data to analyze research collaboration and how coauthorship can be retrieved, refined, and analyzed (Lukkonen et al., 1992; Kretschmer, 1994; Persson & Beckmann, 1995; Melin & Persson, 1996). The coauthorship networks in these studies are usually of relatively small size. Beginning in 2000, several researchers

started to construct large-scale networks using coauthorship data representing research collaborations in various disciplines (Newman, 2001a, 2001b, 2001d, 2004; Barabási et al., 2002; Newman, 2004; Moody, 2004). Topological properties of networks that have been much discussed include graph size, largest components, geodesic distance, degree distribution, clustering coefficient, centrality, and k-core. While Newman (2001a, 2001b) performed analysis on a static network at a specific time point, Barabási et al. (2002) presented the evolution of topological properties of coauthorship networks in mathematics and neuroscience for an eight-year period (1991-98) and built a model to simulate the structural mechanisms that govern the evolution. Moody (2004) explored how variations of the global network topology in sociology collaboration networks have affected the field's research practice in the last 30 years.

Another direction of studies aimed to construct various indicators of the impact of individual authors/institutions/countries through manipulation of coauthorship network properties from a micro perspective (Borner et al., 2005; Liu et al., 2005; Yan & Ding, 2009). Assorted measurements of centrality and adapted models of PageRank are two popular topics of such studies. Borner et al. (2005) proposed a novel local, author-centered measure based on the entropy contribution of a single author's impact across all of its coauthorship relations. Yan and Ding (2009) compared authors' impact ranked by PageRank and various centrality measures over a time span of 20 years and verified their usability. Liu et al. (2005) proposed a weighted PageRank algorithm which takes the number of papers coauthored into consideration. Few of those studies, however, have analyzed the relationship and context between a coauthor pair from a meso perspective. As proposed in this paper, a subgraph that captures important connections between two-coauthors can fill this gap.

Meanwhile, extensive literatures have been devoted to study the internal and external factors that affect scientific collaboration. They have emphasized different aspects of scientific collaboration, including: (1) Cognitive/disciplinary factor; for example, the emerging interdisciplinary areas require collaboration, etc. (Katz & Martin, 1997; Beaver, 2001; Hara, Solomon, Kim, & Sonnenwald, 2003); (2) Geographic factor; for example, researchers who are geographically closer are more likely to collaborate (Katz, 1994; Luukkonen et al., 1992;Schubert & Braun, 1990); (3) Organizational factor; for example, leadership and management of scientific collaboration also play a noticeable role (Finholt & Olson, 1997); (4) Political factor; for example, governments are keen to encourage the level of participation in scientific collaboration (Clarke, 1967; Smith, 1958); (5) Socioeconomic factor (Maglaughlin & Sonnenwald, 2005); (6) Resource accessibility (Cohen, 2000); and (7) Social networks and personal factors; prestige and productivity of researchers also impact their participation in scientific collaboration (Egghe, 2008; Glänzel, 2000; Glänzel & Schubert, 2001). However, most of the previous studies have either analyzed various possible factors theoretically and qualitatively, or verified only an individual factor with quantitative evidences. There lacks a method that can be used to quantitatively analyze all possible factors on a unified platform. In fact, all those factors that affect the scientific collaboration are buried in the background information (e.g., nationality, affiliation, position, expertise, prestige, etc.) of coauthors in the identified contextual subgraph.

Another group of related work addresses graph mining. Subgraph extraction and matching is an emerging topic in the area of graph mining. Estrada et al. (2005) defined a novel centrality measure, referred to as subgraph centrality, which characterizes nodes in a network according to the set of subgraphs formed by random walks starting and ending at the node (i.e., closed walk). The influence of closed walks on the centrality decreases as the length of the walk increases. Their experiments showed that subgraph

centrality is more discriminative for the nodes of a network than degree, betweenness, closeness, or eigenvector centrality. Faloutsos et al. (2004) extracted a subgraph that best captures the relationship between two nodes based on a large graph, using an electricity circuit analogue. Their algorithm was adapted and applied by Ramakrishnan et al. (2005) to multi-relational graphs. Another study utilized subgraphs in measuring proximity between nodes in graphs (Koren et al., 2006). Work of Faloutsos et al. (2004) extended the definition of subgraph to identifying the most important set of intermediate nodes among more than two predefined nodes (Tong & Faloutsos, 2006). While these studies emphasized similarity between indirectly connected pair of nodes, this paper concentrates on contextualizing pairs of authors who are directly connected in a coauthorship network. On the other hand, those studies address the problem of the subgraph from an algorithm perspective, while this study tailors the problem according to specific features of coauthorship networks and shows rich possibilities of exploring scientific collaboration using contextual subgraphs proposed in this paper.

# Methodology

The contextual subgraph between a coauthor pair is defined as a subgraph of the large coauthorship graph that is formed by paths within a certain length between two directly connected authors (i.e., a coauthor pair). A contextual subgraph is thus characterized or defined exclusively by a coauthor pair. The contextual subgraph of a coauthor is created through two steps: 1) identify all the paths within a certain length of the coauthor pair; and 2) merge those paths into a graph.

## Algorithm

A modified heap-based Dijkstra path-finding algorithm is used to efficiently identify the paths within a certain length between two specific nodes in a large-scale graph (Tang et al., 2008). Length denotes the number of jumps needed to reach from one node to another in the undirected coauthorship network. Identified paths are further merged to form the contextual subgraph. More specifically, the approach contains two steps:

1. Enumeration of all paths within a certain length (predefined threshold): a heap-based Dijkstra algorithm with complexity of O($n$log$n$) ($n$ is the size of the original graph) and a depth-first search are used to locate all the paths within a certain length between two specific nodes. Intuitively, search processes begin at the starting node and ending note at the same time. The process systematically explores all the neighboring nodes in sequence, where for each of those neighboring nodes, it visits their unexplored neighbor nodes and records/updates all its stretching-out paths. One path is identified when the two processes visit the same node. Thus the path is recognized by combining the recorded paths between the staring node and the coincidental node, and between the coincidental node and the ending node (see Figure 2). In Figure 2, supposing the starting node is 1 and the ending node is 26:
    - Breadth first search (BFS)explores the nearest neighbor of node 1 and reaches node 3, 4, 6, 7, 10 (Figure 2-b);
    - Meanwhile, another BFS similarly explores the nearest neighbor of node 26 and it reaches node 19, 21, 23, 24, 25 (Figure 2 c);
    - The former BFS further explore all the nearest neighbors of node 3, 4, 6, 7, 10, and reaches 2, 5, 8, 9, 11, 14, 18 (Figure 2 d);

- Meanwhile, the latter BFS explore all the nearest neighbors of node 19, 21, 22, 23, 24, 25, and it reaches 15, 16, 18, 22 (Figure 2 e); and
- A node (i.e., node 18) is visited by both BFS processes; the algorithm ends. The shortest path between node 1 and node 26 is 1 – 10 – 18 – 21 – 26 (Figure 2 f).

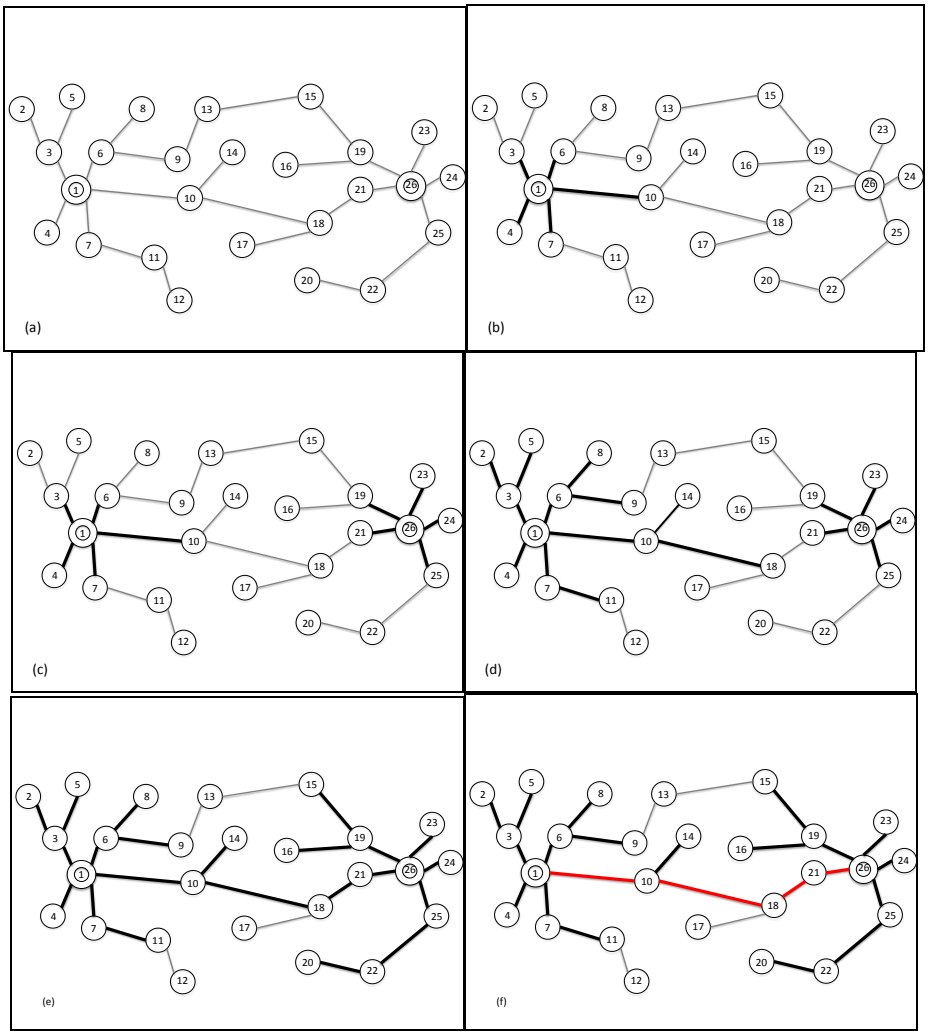


**Figure 2 Finding paths within a certain length between two nodes**

2. Construction of contextual subgraphs: based on the set of paths identified in Step 1, the contextual subgraph is formed in the way that the same nodes in different paths are recognized and merged as one, and that all edges are reserved and kept their association with nodes (see Figure 3).

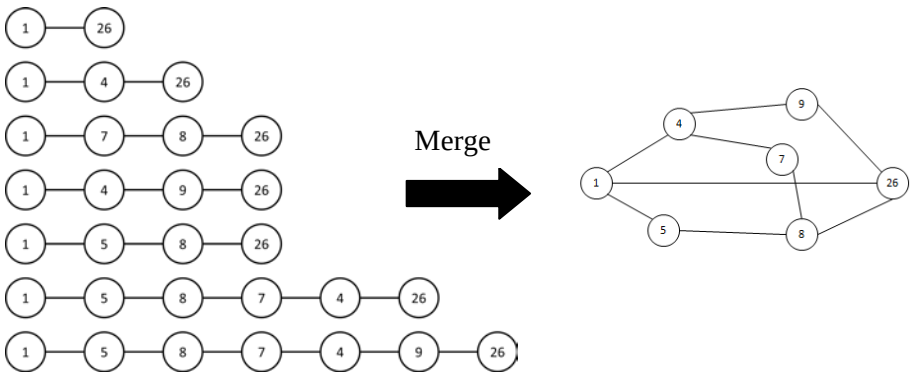


**Figure 3 Merging the set of paths into a contextual subgraph.**

## Data

This proposed methodology is applied to the field of LIS. 50,920 articles written by 42,991 researchers published during 1955-2009 in journals categorized into "INFORMATION SCIENCE & LIBRARY SCIENCE" were downloaded from ISI. In order to explore the dynamics of contextual subgraphs over time, the data were further divided to four accumulative time spans of 1955-1980, 1955-1990, 1955-2000, and 1955-2009. Table 2 shows the descriptive statistics of the data in the four time spans. Based on this dataset, a coauthorship network was built. Each author is a node, and a linkage is created if the two authors have coauthored at least two papers.[1] Table 3 shows the overview of coauthorship network of LIS in the four time spans. Subgraphs for all the possible pairs of coauthors were extracted and investigated through topological properties, including size, average degree, clustering coefficient, network centralization, as well as the correlations between any two of them.

**Table 2 Overview of the LIS dataset**

| | 1955-1980 | 1955-1990 | 1955-2000 | 1955-2009,9 |
|---|---|---|---|---|
| Number of papers | 10,318 | 17,540 | 32,314 | 50,920 |
| Number of papers with more than one author | 1,309 | 3,255 | 8,379 | 17,936 |
| Number of authors* | 2,482 | 5,641 | 14,492 | 30,503 |

*Only authors who have collaborated with at least one other author are included.

[1] We didn't consider homonyms in authors' names. There are two reasons. First, there is no standard way of disambiguating authors' names. Various ways suggested by literatures don't give very satisfactory performance, and they either need more information or are set in a different context. Second, as shown by previous bibliometrics study (Barabasi et al., 2002; Moddy, 2004), name disambiguation doesn't make a big difference in the results in this context. Barabasi et al. (2002) argued that for coauthorship networks, author disambiguation may not be critical. Moody (2004) found no significant difference in the results in coauthorship networks using the methods for name disambiguation.

**Table 3 Overview of LIS coauthorship network***

| | 1955-1980 | 1955-1990 | 1955-2000 | 1955-2009,9 |
|---|---|---|---|---|
| Number of nodes** | 174 | 534 | 1,843 | 4,405 |
| Number of edges | 114 | 375 | 1,684 | 4,793 |

*All components are all included.
**Only authors who have collaborated with at least one other author no less than two times are included.

The maximum length of paths that is allowed to be included in contextual subgraphs is set at six for two reasons. First, for more than 80% of all the pairs of coauthors in 2009, paths within length six cover all the possible paths between them. Second, an intuitive explanation of threhold six can be informed by the theory of six degrees of separation (Milgram, 1967). The coauthorship network is usually seen as a social network, because coauthoring a paper often requires intensive communication of ideas and exchange of expertise. Moreover, six degrees of separation in coauthorship networks is also supported by previous studies. For example, Newman (2001d) found that the typical distance between any two randomly selected scientists is approximately six links.

## Results and Analysis

In this section, a statistical overview of topological properties of all the contextual subgraphs in four time spans is presented along with analyses of typical cases. Furthermore, correlations between topological properties of a subgraph and the productivity, as well as between those properties and citations of the coauthor pair, are calculated and analyzed. Table 4 shows the number of contextual subgraphs in each time span. The numbers of contextual subgraphs in each time span are consistent with the number edges shown in Table 3, since each directly connected pair of authors in coauthorship network corresponds to a contextual subgraph.

**Table 4 Total number of contextual subgraphs in four time spans**

| Year | 1955-1980 | 1955-1990 | 1955-2000 | 1955-2009,9 |
|---|---|---|---|---|
| Total number of contextual subgraphs | 114 | 375 | 1,684 | 4,793 |
| Number of contextual subgraphs with size larger than two | 38 (33.33%) | 133 (35.47%) | 877 (52.08%) | 3,001 (62.61%) |

As shown in Table 4, the percentage of subgraphs with size larger than two increases over time, indicating that collaboration between two authors tends to involves more and more other researchers. This fact reflects the global trend of broadened collaboration. More importantly, it gives prominence to contextual subgraphs in representing the actual practices of modern science, because subgraphs stretch out from the single edge between a coauthor pair and capture a broader range of the actual collaboration relationship.

### Topological properties of subgraphs over time

Topological properties, including graph size, average degree, clustering coefficient, and network centralization, are investigated for all the existing contextual subgraphs in the four time spans.

## Size of contextual subgraphs

Size is the basic topological properties of a network. Figure 4 shows the probability distribution of the subgraphs' number of nodes in different time spans. A general pattern of power law phenomena can be discerned in Figure 4 (power law regression in 1955-2009 gives an exponent 2.059 where $R^2$ equals 0.6982). This fact indicates that a small portion of researchers tend to collaborate in the context of large groups of people (e.g., 30), while a dominant majority of them work with small groups (e.g., less than five). Meanwhile, the size of the largest contextual subgraph in each time span increases from four in 1955-1980 to 35 in 1955-2009.

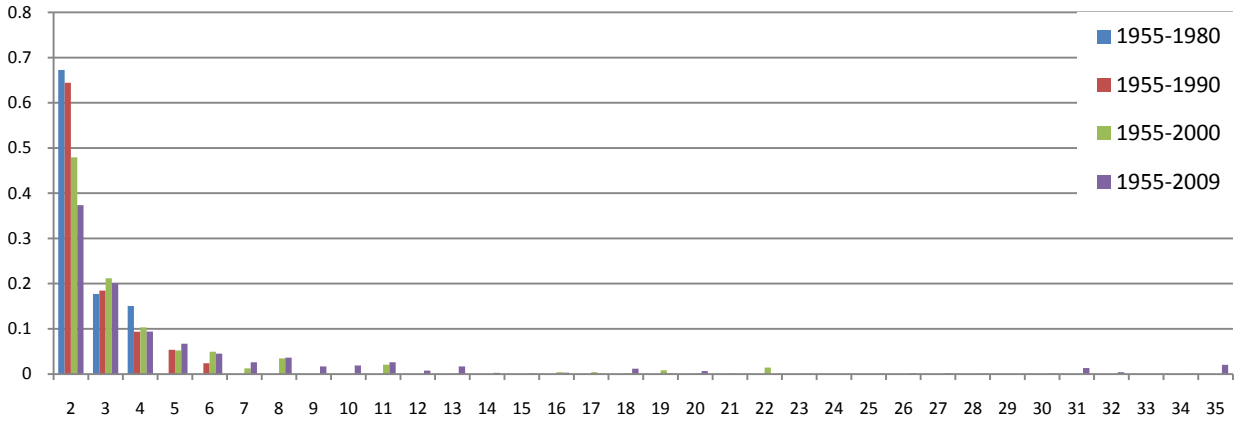


**Figure 4 Probability distribution of size of contextual subgraphs in four time spans**

Figure 5 shows one of the largest contextual subgraphs (31 nodes) in 1955-2009, which is characterized by the coauthor pair G. J. Kuperman and D. W. Bates. Both are specialized in health information technology, especially the use of computer systems to improve patient care with particular respect to clinical decision support. They are respectively affiliated with multiple institutions, including universities, research centers, hospitals and companies.[2] [3] For example, L. Leape, R. Kaushal, and D. W. Bates are all principal investigators in the Center of Excellence for Patient Safety Research and Practice (CEPSRP)[4], while D.L. Seger, J. Fiskio, A. C. Seger, A. Wright and D. W. Bates are members of the Clinical and Quality Analysis group at Partners HealthCare System, Inc. Moreover, institutional-level collaboration is also embedded in this contextual subgraph. For example, the Harvard School of Public Health (one of Bates's affiliations) is listed as a collaborating institution of CEPSRP [5]. This contextual subgraph is defined by two prestigious researchers affiliated with multiple institutions, which probably explain why it becomes one of the largest subgraphs.

[2] See also http://www.coesafety.bwh.harvard.edu/linkPages/peoplePages/core_heads/dwb.htm.

[3] See also http://people.dbmi.columbia.edu/~gjk9001/.

[4] See also http://www.coesafety.bwh.harvard.edu/linkPages/peoplePages/InvestigatorsGenlPage.htm.

[5] See also http://www.coesafety.bwh.harvard.edu/linkPages/aboutPages/collaborating_insts.htm.

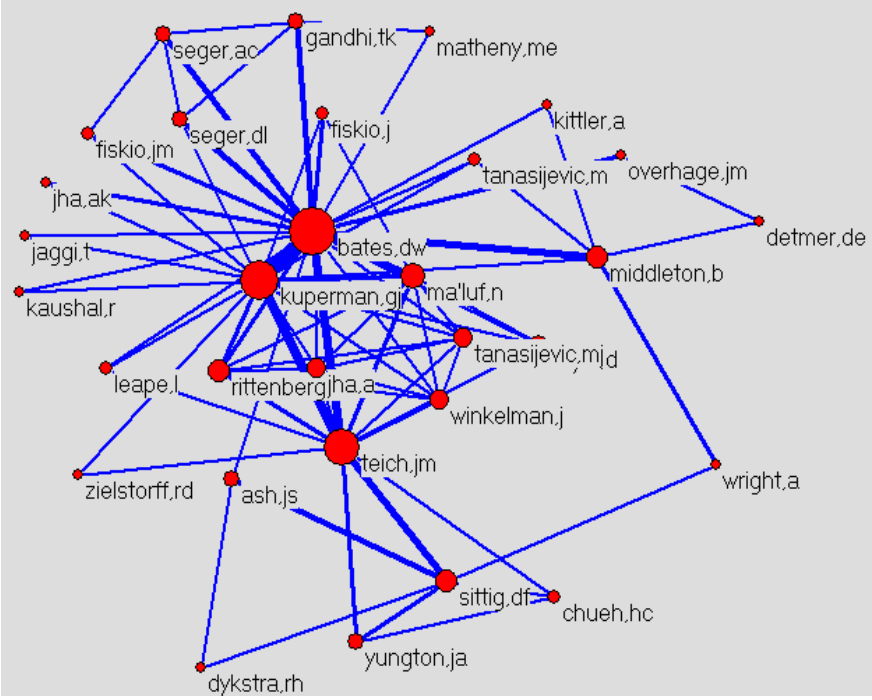


**Figure 5 One of the largest subgraphs between D. W. Bates and G. J. Kuperman in 1955-2009**

As indicated above, compared with saying that D.W. Bates and G.J. Kuperman coauthored more than ten papers, by examining the authors showed in the proposed contextual subgraph between them, we obtain enriched contextual information, much of which is essential for exploring the patterns of collaboration.

### Average degree of contextual subgraphs

Average degree evaluates the connectivity of the subgraph: a higher value suggests that more mediated nodes are shared by different paths. Figure 6 shows the probability distribution of the average degree of subgraphs in the four time spans, which also presents a power law shape in all time spans (power law regression in 1955-2009 gives an exponent as 1.363 where $R^2$ equals 0.9143). In most contextual subgraphs, an author is connected with less than three other authors on average, while in only a small number of subgraphs, an author is connected with more than 10 other authors on average. Similar to graph size, the highest value of average degree of contextual subgraphs also increases prominently over four time spans. The portion of contextual subgraphs with degree larger than one also increases over time, indicating that authors tend to collaborate with more authors across years. This phenomenon reflects the international trend of far-ranging collaborations in modern science (Sonnenwald, 2007).

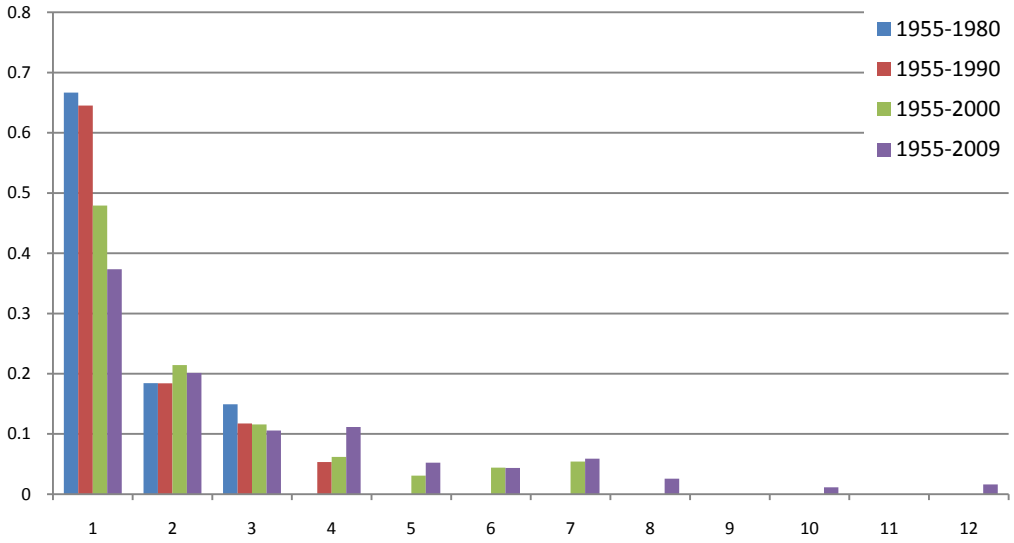


**Figure 6 Frequency distribution of average degree of subgraphs 2009, 9**

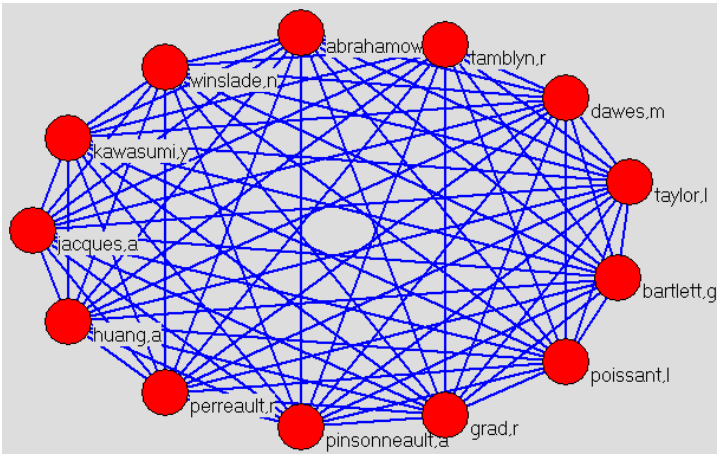


**Figure 7 One of the subgraphs with the largest average degree in 2009, 9**

Figure 7 shows the largest contextual subgraph with average degree equaling 12. This subgraph is a complete graph, in which every node is connected with every other node. The forming of this clique is due to the fact that these 13 authors worked as a research team on a project (http://moxxi.mcgill.ca/moxxihome.html) and published two papers (Tambly et al., 2006; Tamblyn et al., 2008) with all of them listed as authors. This case demonstrates that contextual subgraphs can effectively capture multi-authorships (a.k.a., hyperauthorships (Cronin, 2001)); underlying this multi-authorship are possibly institutional (same affiliations) and economical (the same funded project) factors.

## Clustering Coefficient

The clustering coefficient tells how well connected the neighborhood of one node is. If the neighboring node is fully connected, the clustering coefficient is 1, and a value close to 0 means that there are hardly any connections in the neighborhood. More formally, local clustering coefficient of a node (Wasserman & Faust, 1994) can be represented as:

$$CC_{node} = \frac{number\ of\ existing\ triangles\ formed\ by\ the\ node\ and\ its\ neighboring\ nodes}{number\ of\ all\ possible\ triangles\ formed\ by\ the\ node\ and\ its\ neighboring\ nodes}.$$

The clustering coefficient of a subgraph is obtained by averaging the local clustering coefficient of all nodes. Figure 8 presents the probability distribution of the clustering coefficient of contextual subgraphs in the four time spans. As presented in Figure 8, a substantial portion of contextual subgraphs has a clustering coefficient no less than 0.8, reflecting the strong tendency of collaboration between neighbors of the same author reported by other previous studies (Newman, 2001b; Barabasi, et al., 2002; Moody, 2004). Meanwhile, this phenomenon also reveals well-formed delicate clusters around a majority of coauthor pairs, which cannot be gleaned without proposed contextual subgraphs. In addition, the portion of contextual subgraphs with clustering coefficient equaling 1 decreases in the latter three time spans. This can be explained by the trend of global trend of researches in scientific collaborations, which might weaken the tendency of collaboration between coauthors of the same author.

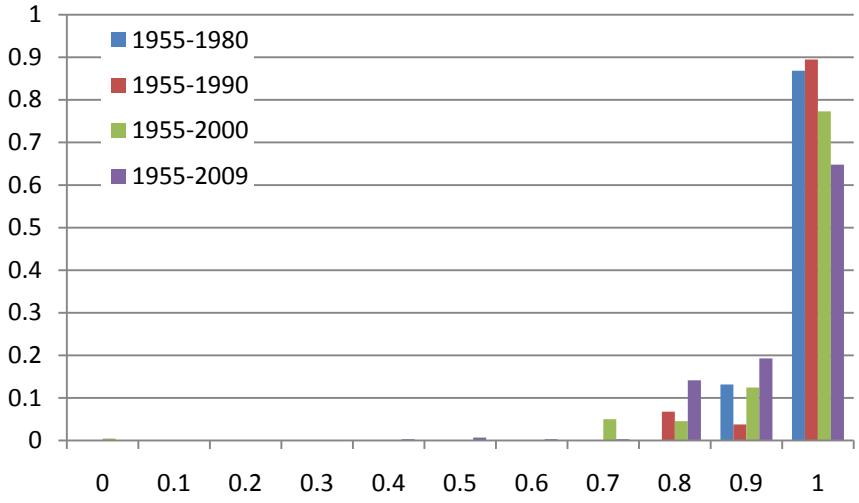


**Figure 8 probability distribution of clustering coefficient of subgraphs 2009, 9 (only subgraphs with size larger than two are included).**

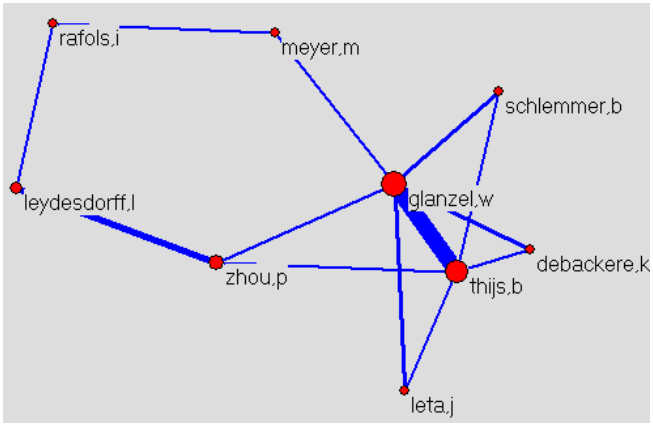


**Figure 9 The subgraph between W. Glänzel and B. Thijs in 1955-2009**

Figure 9 shows the subgraph with the lowest clustering coefficient of all subgraphs whose sizes are larger than four, which captures the context of collaborations between W. Glänzel and B. Thijs. Examination of these authors' affiliations shows that the nine affiliated authors are associated with institutions located in Hungary, Belgium, Netherlands, England, Finland, and Brazil, respectively or jointly. This broad international background probably weakens the tendency of collaboration between neighbors of the same author.

### Network Centralization

Network Centralization assesses the global centrality of the network. Network centralization is defined as the ratio of variations of degree divided by the largest possible degree variations with the same size. The possible largest degree variation of a simple graph occurs in a strict star-structure. More formally, network centralization (Wasserman & Faust, 1994) can be represented by the following formula:

$$C_{network} = \frac{degree\ variaitons}{largest\ possible\ degree\ variations\ with\ the\ same\ size}.$$

Figure 10 shows the probability distribution of network centralization values of these subgraphs. A dominant proportion of the subgraphs takes a 0 in network centralization, indicating that a substantial formation of contextual subgraphs could be attributed to multiple-authorship. Meanwhile, the portion of subgraphs take a value greater than 0 in network centralization increases in the latter three time spans and subgraphs with network centralization larger than 0.5 emerge in 1995-2009, revealing that more star-like structured contextual subgraphs tend to be popularized over time (subgraphs of network centralization larger than 0.5 are referred to as a star-like structure). Various factors may contribute to the increasingly obvious star-like structure. For example, applying to research funding usually requires one or more principal investigators (PI) who are usually prestigious scholars in the area. The standing of PIs would have much indirect impact on facilities, publications, research groups, and so forth, strengthening the preferential attachment (Newman, 2001c).

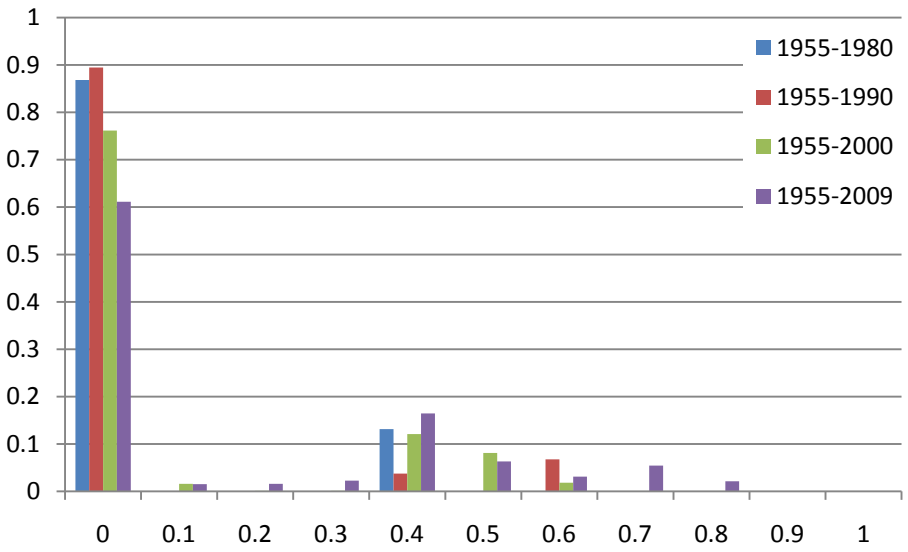


**Figure 10 probability distribution of network centralization of subgraphs in four time periods**

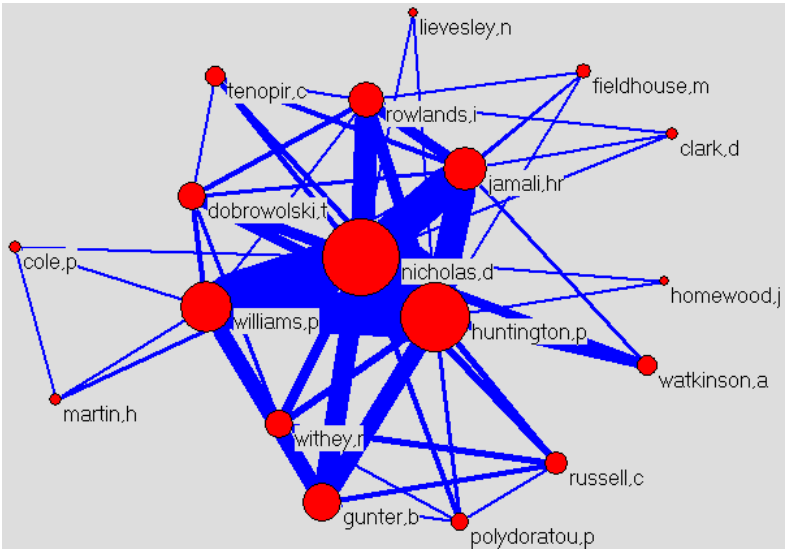


**Figure 11 Frequency distribution network centralization of subgraphs 2009, 9 between D. Nicholas and C. Tenopir (network centralization = 0.71324)**

Figure 11 shows one of subgraphs with the largest network centralization value, indicating a dominant star-like structure. This subgraph is defined by D. Nicholas and C. Tenopir. D. Nicholas is Director of the Centre for Information Behavior and the Evaluation of Research (CIBER) and Professor/Director of the School of Library of University College London; P. Huntingtona and H.R. Jamali are senior researchers and founder members of CIBER, and most of the remaining authors are also affiliated with CIBER.

## Correlations between topological properties

In addition to the topological properties of subgraphs discussed in preceding sections, correlations between them are also presented and analyzed in this section (Tables 5-8). The correlations can be used to discern relationships among different features of contextual subgraphs in terms of whether they are positively or negatively related, or in other words, whether certain features (such as large size and a high

value of average degree) are inclined to co-occur in the same contextual subgraphs. Spearman's rank correlation coefficient, rather than Pearson product-moment correlation coefficient, is adopted here for two reasons. First, as shown in previous sections, the probability distributions of those topological properties of contextual subgraphs do not meet the vibrate normality requirement, so Pearson product-moment correlation coefficient may be misleading if applied to this data set. Second, Pearson product-moment correlation coefficient assumes a linear dependence between two variables, which cannot be verified by available data. Instead, Spearman's rank correlation coefficient is a non-parametric measure of correlation between two variables, which doesn't require the dependency to be required by a linear relationship.

**Table 5 Correlation between topological properties of subgraphs in 1955-1980**

| 1955-1980 | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| size | 1 | | | |
| average degree | .964** | 1 | | |
| network centralization | .433** | .179 | 1 | |
| clustering coefficient | -.433** | -.179 | -1.000** | 1 |

**Correlation is significant at the 0.01 level (2-tailed).

**Table 6 Correlation between topological properties of subgraphs in 1955-1990**

| 1955-1990 | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| Size | 1 | | | |
| average degree | .953** | 1 | | |
| network centralization | .468** | .192* | 1 | |
| clustering coefficient | -.468** | -.192* | -1.000** | 1 |

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

**Table 7 Correlation between topological properties of subgraphs in 1955-2000**

| 1955-2000 | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| Size | 1 | | | |
| average degree | .967** | 1 | | |
| network centralization | .610** | .435** | 1 | |
| clustering coefficient | -.608** | -.412** | -.967** | 1 |

**Correlation is significant at the 0.01 level (2-tailed).

**Table 8 Correlation between topological properties of subgraphs in 1955-2009**

| 1955-2009 | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| Size | 1 | | | |
| average degree | .945** | 1 | | |
| network centralization | .692** | .475** | 1 | |
| clustering coefficient | -.693** | -.463** | -.958** | 1 |

**Correlation is significant at the 0.01 level (2-tailed).

Tables 5-8 show that phenomena of large size, high values of average degree, star-like structure, and low clustering coefficient tend to be co-presented in the same subgraphs. Meanwhile, the growing strength of the correlation between graph size and network centralization, as well as average degree and network

centralization over time, indicate the trend of more close and systematic collaboration practices. This can be explained that, with expanding collaboration space and possibilities, and increasing funding sources, researchers tend to form a relatively systematic way of collaboration, typically with a large number of authors in the core and a small number in the periphery. This situation further results in the observed correlation between topological properties of contextual subgraphs. The scatter plot of those properties in 1955-2009 is shown in Figure 12.

## Productivity, citation and contextual subgraphs

In order to show the rich possibilities of quantitatively analyzing various aspects of scientific collaboration, in this section we provide two examples. Productivity and citation form the base of impact analysis. We investigate the two possible formulations of the relation between collaboration and productivity, as well as collaboration and citations, as the following two questions:

- whether higher values of average productivity or standard deviation of productivities of two coauthors is associated with larger/denser/star-like contextual subgraphs; and
- whether higher values of average number of citations or standard deviation of citations of two coauthors is associated with larger/denser/star-like contextual subgraphs.

### Productivity and contextual subgraphs

An author's productivity is defined as the number of papers in which he/she is listed as an author in the period under investigation. The great diversity that exists among scholars with respect to research productivity has been frequently documented (Wanner et al., 1981), and explanations for such differences can be categorized into background characteristics, such as gender, other demographics or socioeconomic origins, and features of the academic career, such as rank or quality of institution, career stage, disciplines, and publication type. Comparisons based on these demographic variables, however, tend to be anomalous (Borgman, et al., 2002).

**Table 9 Correlation between topological properties of subgraphs and average productivity of the coauthor pair**

| Average productivity | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| 1955-1980 | .500** | .495** | .495** | 0.172 |
| 1955-1990 | .061 | -0.072 | .452** | -.452** |
| 1955-2000 | .131** | 0.003 | .398** | -.424** |
| 1955-2009 | .213** | .075** | .430** | -.440** |

**Correlation is significant at the 0.01 level (2-tailed).

**Table 10 Correlation between topological properties of subgraphs and the standard deviation of productivity of the coauthor pair**

| Standard deviation of productivity | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| 1955-1980 | .532** | .519** | .519** | 0.209 |
| 1955-1990 | -0.013 | -0.128 | .356** | -.356** |
| 1955-2000 | .067* | -0.034 | .308** | -.321** |
| 1955-2009 | .180** | .063** | .360** | -.364** |

*Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Tables 9 and 10 show the correlation between topological properties of sets of contextual subgraphs and the average productivity and standard deviation of productivities of the coauthor pairs. As shown in Table 9, in all time spans except 1955-1990, the size of contextual subgraphs is positively correlated to the average productivity with statistic significance, indicating that high values of average productivity and large size of contextual subgraphs tend to vary in the same direction over time. It's reasonable that authors writing relatively large quantities of papers tend to be involved with more coauthors. Meanwhile, a broad connection with researchers may also tend to increase one author's productivity. This result confirms conclusions from previous studies. For example, Egghe (2008) proved that high productivity leads to high fractions of coauthored papers (but low productivity can have low or high fractions of coauthored papers) using a scientometrics dataset. The standard deviation of productivities also presents a significant positive correlation with graph size in three time spans (Table 9). Additionally, network centralization shows statistically significant correlations with average and standard deviation of productivity over all time spans. An intuitive example might be mentorship. Usually, a combination of a tenured faculty and a junior doctoral student may make both the average and standard deviation of their productivities high; the collaboration between them tends to be a star-like structure with the faculty as a star and different doctoral students as peripheral nodes. The scatter plot of those variables in 1955-2009 is shown in Figure 12.

### Citation and contextual subgraphs

As for the relationship between collaboration and citation, it is well known that multi-authored publications, publications coming from more than one institution, and publications coming from more than one country on average are cited more than single-authored publications (Glänzel, 2000; Glänzel & Schubert, 2001). It is also known that international coauthorship tends to result in publications with higher citation counts than purely domestic publications (Narin et al., 1991; Persson et al., 2004). Citation counts of authors are calculated using references of articles published in journals categorized as library and information science on ISI from 1955 to Sep 2009. In each time span, citation counts take only paper published in this time span into consideration. For example, when calculating citations counts for papers in time span 1955-1980, only citations within this time period are counted.

Tables 10 and 11 show the correlation between topological properties of contextual subgraphs and citations of the coauthor pair. The correlation between topological properties of subgraphs and citations of the coauthor pair is not as strong as that of productivity in four time spans. As shown in Table 11, in 1955-1990 and 1955-2000, there is a negative correlation between subgraph size and average citation of the coauthor pair, implying highly cited authors tend to be associated with relatively small contextual subgraphs. This is reasonable, because prestigious scholars are relatively cautious when selecting collaborators. Moreover, Table 12 shows that in all time spans, the standard deviation of citations of the coauthor pair yields a significant negative correlation with the average degree, suggesting that coauthors with relatively similar citations tend to be associated with relatively dense subgraphs. Figure 12 shows the scatter plot of those topological properties of contextual subgraphs as well as the popularity and prestige of the pair of coauthors. The box with text within it denotes the coordinates in both horizontal and vertical direction. Figure 12 is provided besides the tables of spearman's correlation index, because scatter plots can literally present the relationship between variables through plotting each individual in the sample. Therefore, it can give a more direct and vivid image of the correlation between variables under investigation.

**Table 11 Correlation between topological properties of subgraphs and average citation of the coauthor pair**

| Average citation | size | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| 1955-1980 | 0.189 | 0.293 | -0.295 | 0.295 |
| 1955-1990 | -.256** | -.278** | 0.061 | -0.061 |
| 1955-2000 | -.175** | -.246** | .086* | -.091** |
| 1955-2009 | 0.02 | -.087** | .221** | -.206** |

*Correlation is significant at the 0.05 level (2-tailed).

**Correlation is significant at the 0.01 level (2-tailed).

**Table 12 Correlation between topological properties of subgraphs and standard deviation of citations of the coauthor pair**

| Standard deviation of citation | node | average degree | network centralization | clustering coefficient |
|---|---|---|---|---|
| 1955-1980 | 0.297 | .370* | -0.158 | 0.158 |
| 1955-1990 | -.213* | -.225** | 0.041 | -0.041 |
| 1955-2000 | -.155** | -.222** | .086* | -.091** |
| 1955-2009 | 0.028 | -.065** | .193** | -.172** |

*Correlation is significant at the 0.05 level (2-tailed).

**Correlation is significant at the 0.01 level (2-tailed).


**Figure 12 Scatterplot matrix of topological properties of contextual subgraphs as well as average and standard deviation of productivities and citations of co-author pairs in 1955-2009**

Addressing the two questions introduced in the beginning of this section, the above discussion results show that higher values of average productivity and the standard deviation of productivities of a coauthor pair are positively associated with larger, denser, and especially star-like contextual subgraphs. Meanwhile, high values of average citation counts and standard deviation of citation counts of a coauthors pair are associated with smaller, sparser, and star-like contextual subgraphs.

## Discussion and Conclusion

In this paper, contextual subgraph is proposed as a novel meso perspective method to quantitatively illustrate various factors that affect scientific collaborations. It captures the authors either directly or indirectly involved with the scientific collaboration between two specific pair of authors. The identification of these authors makes it possible to analyze various possible factors that affect the scientific collaboration by exploring the background information of these authors. Thus, a unified, quantitative framework for exploring scientific collaboration has been built. More specifically, this method can (1) provide a close look at the context of collaboration between two specific coauthors in large-scale collaboration networks; (2) give a way to trace back the underlying motives of coauthorship quantitatively; and (3) the general procedure of data processing makes it applicable across disciplines and make the comparison across fields possible. Here we go back to the research questions raised in the introduction part and summarize the answers to them.

- In what context do two coauthors collaborate and why?

First, given a specific pair of coauthors in LIS as a query request, the proposed algorithm can efficiently and effectively provide and visualize the contextual subgraph characterized by them. It is useful to discover contextual information about the collaboration between these two authors (as suggested in Figure 1). Second, by conducting a global statistical analysis of the set of all the existing contextual subgraphs in LIS, this study depicts the general picture of scientific collaborations in LIS during the time period under investigation. Topological properties of contextual subgraphs in LIS during different time spans are investigated, generally showing power law shape in probability distribution. Additionally, correlations between these topological properties indicate that large size, high values of average degree, star-like structure, and low clustering coefficients tend to be co-presented in the same subgraphs.Moreover, how topological properties of contextual subgraphs correlate with productivities and citations of coauthor pairs, were explored to shed light on why contextual subgraphs present such structural features. As shown in the analysis, higher values of average and the standard deviation of productivities of a coauthor pair is positively associated with larger, denser, and especially more star-like structured contextual subgraph. Meanwhile, high values of average citation counts and standard deviation of citation counts of a coauthors pair are associated with smaller, sparser, and star-like contextual subgraph.

Contextual subgraphs can also be incorporated into analyses of coauthorship networks of a macro perspective. Generally, clustering coefficients measure the features of two linked nodes that are each linked to a third node. Consequently, these three nodes form a triangle and the clustering is frequently measured by counting the number of triangles in the network (Girvan & Newman, 2004). It has been observed that not only triangles but also other subgraphs are significant in real networks. Contextual subgraphs can be seen as a generalized form of the measure of clustering coefficients, which denotes the number of triangles (a specific type of subgraph) divided by the number of possible triangles in the graph of the same size. Meanwhile, analysis of coauthorship networks from a micro perspective can also take advantage of contextual subgraphs. For example, combining all subgraphs between one author and each of its neighbor nodes can be used to describe his/her participation and roles in all the contextualized coauthorships. From this perspective, this combined subgraph can be seen as a generalized measure of degree centrality, which is the size of the subgraph formed by all the associated paths with length one.

Our future research will focus on building new micro indicators for individual authors based on contextual subgraphs.

## Acknowledgements

We gratefully acknowledge the project ArnetMiner developed by Dr. Jie Tang at Department of Computer Science and Technology of Tsinghua University. We also want to thank Erjia Yan, Yuyin Sun, and the two anonymous reviewers for their insightful comments.

## References

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

Beaver, D. deB., & Rosen, R. (1978). Studies in scientific collaboration. Part. I, The professional origins of scientific co-authorship. *Scientometrics 1:*65-84.

Beaver, D. deB., & Rosen, R. (1979). Studies in scientific collaboration. Part. II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. *Scientometrics 1*:133-49.

Borgman, C.L., & Furner, J. (2002). Scholarly Communication and Bibliometrics. In B.Cronin (Ed.), *Annual Review of Information Science and Technology, Vol 36.* Medford, NJ: Information Today, 3-72.

Borner, K., Dall'Asta, L., Ke, W. M., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of coauthorship teams. *Complexity, 10*(4), 57-67.

Barabási, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A 311*, 590-614.

Clarke, B.L. (1976). Communication patterns of biomedical researchers. *Federation Proceeding, 26*: 1288-92

Cohen, J., (2000). Balancing the collaboration equation. *Science, 288,* 2155–2159.

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.

Egghe, L., Goovaerts, M., & Kretschmer, H. (2008). Collaboration and productivity: An investigation in Scientometrics and in a university repository. *Journal of Scientometrics and Information Management, 2*(1), 83-89.Estrada, E. & Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E, 71*(5), 056103.

Faloutsos, C., McCurley, K. S., & Tomkins, A. (2004). *Fast discovery of connection subgraphs.* Paper presented at the proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Finholt, T. A., & Olson, G. M. (1997). From Laboratories to Collaboratories: A New Organizational Form for Scientific Collaboration. *Psychological Science, 8*(1), 28-36.

Glänzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics*, 48, 121-150.

Glänzel, W., Schubert, A. (2001). Double effort-double impact? A critical view at international coauthorship in chemistry, *Scientometrics*, 50, 199-214.

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

Katz, J. (1994). Geographical proximity and scientific collaboration. *Scientometrics, 31*(1), 31-43.

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

Koren, Y., North, S. C., & Volinsky, C. (2006). Measuring and extracting proximity in networks. Paper presented at the proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kretschmer, H. (1994). Coauthorship networks of invisible colleges and institutionalized communities. *Scientometrics, 30*(1), 363-369.

Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*(4), 317-325.

Liu, X. M., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Coauthorship networks in the digital library research community. *Information Processing & Management, 41*(6), 1462-1480.

Luukkonen T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration, *Science, Technology And Human Values, 17*, 101-126.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding Patterns of International Scientific Collaboration. Science, *Technology & Human Values, 17*(1), 101-126.

Maglaughlin, K. L., & Sonnenwald, D. H., (2005). Factors that impact interdisciplinary natural science research collaboration in academia. *Proceedings of the Conference of the International Society for Scientometrics and Informetrics*, 499–508.

Melin, G. & Persson, O. (1996). Studying research collaboration using coauthorships, *Scientometrics, 36*, 363–377.

Milgram, S. (1967). The small world problem, *Psychology Today 2*, 60–67.

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

Narin, F., Stevens, E. S., & Whitlow, E. S. (1991), Scientific cooperation in Europe and the citation of multi-nationally authored papers, *Scientometrics*, 21, 313–324.

Newman, M. E. J. (2001b). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*, 016132.

Newman, M. E. J. (2001c). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102.

Newman, M. E. J. (2001d). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA 98*, 404-409.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences USA 101,* 5200–5205.

Newman, M.E.J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E, 64*(1), 016131.

Persson, O., Beckman, M. (1995). Locating the network of interacting authors in scientific specialties. *Scientometrics, 33,* 351–366.

Persson, O., Glänzel, W., Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60, 421–432.

Ramakrishnan, C., Milnor, W. H., Perry, M., & Sheth, A. P. (2005). Discovering informative connection subgraphs in multi-relational graphs. *ACM SIGKDD Explorations Newsletter,7*(2), 56-63.

Schubert, A., & Braun, T. (1990). International collaboration in the sciences, 1981-1985. *Scientometrics 19*: 3-10

Smith, M. (1958). The trend toward multiple authorship in psychology. *American Psychologist, 13*: 596-599

Sonnenwald, DH.(2007). Scientific collaboration: A synthesis of challenges and strategies, Cronin B.,(ed). *Annual Review of Information Science and Technology*, vol. 41.

Tamblyn, R., Huang, A., Kawasumi, Y., Bartlett, G., Grad, R., Jacques, A., et al. (2006). The development and evaluation of an integrated electronic prescribing and drug management system for primary care. *Journal of the American Medical Informatics Association, 13*(2), 148-159.

Tamblyn, R., Huang, A., Taylor, L., Kawasumi, Y., Bartlett, G., Grad, R., et al. (2008). A randomized trial of the effectiveness of on-demand versus computer-triggered drug decision support in primary care. *Journal of the American Medical Informatics Association, 15*(4), 430-438.

Tang, J, Zhang, J., Yao, L., Li, J., Zhang, L. & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Tong, H., & Faloutsos, C. (2006). Center-piece subgraphs: Problem definition and fast solutions. Paper presented at the proceedings of the Twelfth ACM SIGKDD International conference on Knowledge Ciscovery and Data Mining.

Wanner, R. A., Lewis, L. S., & Gregorio, D. I. (1981). Research productivity in academia: A comparative study of the sciences, social sciences and humanities. *Sociology of Education, 54*(4), 238-253.

Wasserman, S. & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge UK: Cambridge University Press.

Yan, E. J., & 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.

Yan, E., Ding, Y. & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics,* 83(1), 115-131.