Mining Patterns of Author Orders in Scientific Publications

Bing He, Ying Ding, Erjia Yan
{binghe; dingying; eyan}@indiana.edu
School of Library and Information Science, Indiana University, Bloomington

Abstract

The author order of multi-authored papers can reveal subtle patterns of scientific collaboration and provide insights on the nature of credit assignment among coauthors. This article proposes a sequence-based perspective on scientific collaboration. Using frequently occurring sequences as the unit of analysis, this study explores (1) what types of sequence patterns are most common in the scientific collaboration at the level of authors, institutions, U.S. states, and nations in Library and Information Science (LIS); and (2) the productivity (measured by number of papers) and influence (measured by citation counts) of different types of sequence patterns. Results show that (1) the productivity and influence approximately follow the power law for frequent sequences in the four levels of analysis; (2) the productivity and influence present a significant positive correlation among frequent sequences, and the strength of the correlation increases with the level of integration; (3) for author-level, institution-level, and state-level frequent sequences, short geographical distances between the authors usually co-present with high productivities, while long distances tend to co-occur with large citation counts; (4) for author-level frequent sequences, the pattern of "the more productive and prestigious authors ranking ahead" is the one with the highest productivity and the highest influence; however, in the rest of the levels of analysis, the pattern with the highest productivity and the highest influence is the one with "the less productive and prestigious institutions/states/nations ranking ahead."

Keywords: author orders, author sequence, scientific collaboration

1 Introduction

Collaboration is becoming a common practice in scientific research. It brings the complementary backgrounds of participating experts into one project, resulting in more publications, and providing more opportunities for graduate students and junior faculty members. Clearly stating the identity and order of the authors gives information about who is accountable for the integrity of the reported study and who deserves what amount of credit for the work (Savitz, 1999; Rennie, Yank, & Emanuel, 1997; Rennie & Flanagin, 1994). Meanwhile, researchers use author lists to form impressions about the capabilities and achievements of the authors. People who sit on committees of recruit, promotion, awards, and honors greatly base their assessment of a candidate on his/her position in the author list of his/her publications. The author order of scientific publication can be closely related to the fairness of evaluation systems and the unwritten rules of credit assignment, which are crucial sectors in the sustainable development of academic communities. The author order also has practical implications for global scientific policies of

government and funding agencies. It is therefore crucial to uncover the underlying patterns of author orders stated in scientific publications.

Several patterns of author order have been already noticed by researchers in different fields. In the early 20th century, alphabetical ranking of authors was used in political sciences and economics (Endersby, 1996). Nowadays, authors are generally ranked by the significance and amount of their contributions to the reported research. In biomedical research, however, some found that the last author noted makes the most contribution, followed by the second author (Tscharntke et al., 2007). Others added footnotes to elaborate on each author's contribution. In computer science, a common practice has been to mark several authors and indicate that they have equal contributions. In clinical research, the last author is usually "the person in whose laboratory the study was done and who was peripherally involved with the details of the study, but who also participated in either the general conception, supplying the administrative support, or overseeing the general progression of the study" (Burman, 1982). Different author orders thus reflect different epistemic cultures in scientific collaboration practices of differing fields.

As early as the 1960s, researchers have studied the ordering of authors in scientific publications (Zuckerman, 1968; Floyd, Schroeder, & Finn, 1994; Rennie, Yank, & Emanuel, 1997; Joseph, Laband, & Patil, 2005). There is a rich collection of literatures discussing the significant shift from alphabetic ordering to a contribution-based ordering of authors (Peffers & Hui, 2003; Riesenberg & Lundberg, 1990; Tscharntke et al., 2007). These studies have explored the order of authors from the social, ethical, disciplinary, and intellectual property perspectives, mostly within medical-related fields. Yet due to the difficulty of analyzing author orders as well as the lack of a framework for quantitative analysis, few studies have been conducted with a large-scale quantitative investigation of author orders.

In this paper, we design a framework for analyzing author orders and take advantage of frequent sequence mining algorithms to empirically study the author orders in the field of library and information science (LIS). One of the core parts of our framework is that the unit of analysis is set to be the subsequences of adjacent co-authors that frequently occur in the published papers. Similar to the cases using individual authors as the unit of analysis, the productivity (measured by number of papers published) and influence (measured by the sum of citation counts of published papers) of the frequent sequences are analyzed. Note that "influence" and "prestige" are used interchangeably in this paper. Those frequent sequences are grouped into four categories according to the relative level of productivity, and the influence of the two individual authors who comprise the sequences. Moreover, our analysis is conducted at different levels of integration, including author-level analysis, institution-level analysis, state-level analysis, and nation-level analysis (i.e., international collaboration).

2 Literature Review

2.1 Author Orders

With the prominent trend of scientific collaboration, the topic of author orders has attracted much attention from researchers in many areas, most of whom have focused on the correspondence between the author order and the relative amount of credit assigned to the authors. Early literatures on author orders mostly approached the problem from a social or ethical perspective. Von Glinow and Novelli (1982) asked the question of how authorship orders should be determined and argued that credit based on

prestige was viewed as unfair, but that listing authors alphabetically was sometimes seen as appropriate. Fine and Kurdek (1993) focused on the specific context of author orders in relation to collaboration between graduate students and faculty members. They presented hypothetical cases that describe typical ethical dilemmas occurring in the context, and made recommendations to faculty that highlight ethical principles. They suggested that the relative scholarly abilities and professional contributions of the collaborators should be used as the criteria to decide authorship credit and order, and that decision-making processes about authorship order should start early in the collaborative endeavor. Floyd et al. (1994) developed a theoretical framework to account for conflicts over credit for collaborative research. They provided evidence for the effect of individuals' motives and attitudes on the criteria for author orders, and proposed that judging the degree of contribution from author orders should be done with caution. Tscharntke et al. (2007) summarized different methods of assigning credit based on author lists: (1) the "sequence-determines-credit" approach (SDC); (2) the "equal contribution" norm (EC); (3) the "first-last-author-emphasis" norm (FLAE); and (4) the "percent-contribution-indicated" approach (PCI). Savitz (1999) proposed to build a reflection of a consensus about the interpretation of credit accountability from the author orders. Other relevant literatures include that of Renni, Yank, and Emanuel (1997), Laurance (2006), Riesenberg and Lundberg (1990), and Rennie and Flanagin (1994).

Another set of literature took an empirical perspective on the issue of author orders and conducted quantitative analyses. Hunt and Blair (1987) explored the correlation between authors' prestige and their ranks in the author lists. They found that tenure is negatively related to ranks, indicating that more prestigious authors are less concerned with order as they tend to be given more credit. Hunt and Blair called this phenomenon the Matthew Effect. Peffers and Hui (2003) computed the percentages of papers with alphabetically ranked author lists in journals with high impact factors versus journals with median or low impact factors in the field of Information Management Systems (IS). They found that in top IS journals, the alphabetical ranking of authorship tends to disappear. In Baerlocher et al.'s study (2007), by means of designed questionnaires, authors were asked to assess the contributions of each author in eleven categories. Their results showed that the first authors presented the highest level of participation in most categories. Zuckerman (1968) interviewed Nobel laureates concerning their positions in the author lists, and compared their rank in these lists to that of their co-workers. He showed that Nobel laureates tend not to be ranked top in the author list as their reputation grows. Different from previously discussed papers, which concentrated on the relationship between author orders and credit assignment in collaborative papers, Joseph et al. (2005) raised the question of how author order is related to the quality of the paper. They built a stochastic model of author orders under the assumption that each author works equally hard to get priority in ranking. They found that in the field of economics, the quality of alphabetically ranked papers is higher than the quality of the non-alphabetically ranked papers. In this paper, we propose a new framework of quantitatively analyzing author orders that decomposes the author list into frequently occurring subsequences, and provides a way to correlate different patterns of subsequences to the general productivity and influence of the collaboration.

2.2 Sequential Pattern Mining

Sequential pattern mining, as one subarea of frequent pattern mining, has been a focused theme in data mining research for over a decade. The goal of sequential pattern mining is to find the frequent patterns from a collection of sequences, such as finding personal shopping preferences from customer shopping sequences, finding user behavior patterns from Web clickstreams data, and finding functional areas of

DNA from gene sequencing data. Since its introduction by Agrawal and Srikant (1995), sequential pattern mining has become an important topic in data mining. Various algorithms have been proposed to provide optimal solutions to this problem, among which Apriori-based algorithms are a classic family of algorithms. Srikant and Agrawal (1996) proposed an algorithm called Generalized Sequential Patterns (GSP), which uses the downward-closure property of sequential patterns and adopts a multiple pass, candidate generate-and-test approach. Zaki (2001) extended the vertical format-based frequent itemset mining methods Eclat (Zaki, 1998) to a sequential pattern mining method, referred to as SPADE. Other relevant studies include that of Pei et al. (2001, 2004) and Yan et al. (2003). Note that all these algorithms share the same goal but differ in their efficiencies. In this study, an implementation of SPADE in R is used (Buchta & Hahsler, 2010).

3 Methodology

3.1 Sequential Pattern Mining

The author list of a collection of published papers can be seen as a database of sequences. A frequent pattern in sequential pattern mining is a subsequence whose relative occurrence frequency is higher than a predefined threshold in the collection of sequences. As shown in Figure 1, each row represents the author list of a published paper. We define the number of occurrences of a subsequence divided by the total number of papers as the *support value* of the subsequence. The support values of subsequences <*author A, author C*>, <*author C, author B*>, and <*author A, author C, author B*> are therefore 0.6, 0.5, and 0.3 respectively, which are higher than a predefined threshold, such as 0.3. So these three subsequences are referred to as frequent sequences.

3.2 Framework of analysis

One paper may contain several frequent sequences. How citation counts of a paper should be assigned to each frequent sequence should thus be considered. We adopt the most frequently used method in splitting citations over individual authors, wherein each sequence is allocated equal citation counts, which is the same as the citation counts of the paper. All the citation counts of the papers containing one specific sequence add up to the final influence score for that very sequence; a similar process is done on productivity (See Figure 1).

Papers	Author list	Citation counts
1	author A; author C; author B; author D	1
2	author A; author C	13
3	author C; author B	5
4	author A; author C; author E; author F	0
5	author E; author A; author C; author B	1
6	author F; author A; author C	1
7	author F; author C; author B; author D	0
8	author E; Author D; author A; author C; author B	0
9	author F; author A	0
10	author E; author A; author C	0

Frequent Sequences/Patterns	Productivity	Influence	Support
author A, author C	6	16	0.6
author C, author B	5	7	0.5
author A, author C, author B	3	2	0.3

Figure 1 an example of sequential pattern mining

For each level of integration, we further summarize the frequent sequential patterns according to the relative amount of citations and productivities between authors comprising the sequence. For this part, we only analyze subsequences of two, because other sequences can be deconstructed into sequences of length two (i.e., the property of downward closure). We develop an intuitive framework of grouping:

- Citation counts of the first author is equal to that of the second author (c1=c2) and the number of papers published by the first author is equal to that of the second author (p1=p2)
- c1>c2 and p1>p2
- c1>c2 and p1<=p2
- c1<=c2 and p1>p2
- c1<=c2 and p1<=p2 (excluding c1=c2 and p1=p2)

After grouping the frequent sequential patterns into the above five categories, corresponding indexes are also aggregated. Moreover, all the processing and analyses are performed at different levels of integration, including author level, institution level, state level, and nation level.

3.3 Dataset

This proposed methodology is applied to the field of LIS. A total of 50,920 articles written by 42,991 researchers published during 1955 to 2009 in journals categorized into “INFORMATION SCIENCE & LIBRARY SCIENCE” were downloaded from ISI. In order to get the influence and productivity scores for frequent sequences at different levels of integration, data are processed in several steps. For the

analysis level of author, single-authored papers are excluded, as are the citations of papers with single authors. For institution-level, state-level, and nation-level analysis, documents without addresses are excluded since we cannot determine sequences of institutions, states, and nations without address information. Meanwhile, records with all authors from the United States are analyzed at the level of states, while records with at least one author from non-U.S. areas are investigated at the level of nations. Other details of processing the data, including extracting unique institution names, state names, and nation names from the address information, are elaborated in Yan and Sugimoto (2011). Table 1 shows the descriptive statistics of the processed data.

Table 1 Basic statistics of the dataset

Four levels	number of records	Support value	No. of frequent sequences of length two
Author-level	17,938	0.00011	2,524
Institution-level	1,572	0.0012	188
State-level	588	0.003	209
Nation-level	488	0.000	348

4 Results and Analysis

In this section, results are reported at author-level, institution-level, state-level, and nation-level analysis. At each level, top-ranked frequent patterns in terms of productivity and influence are presented, followed by the results under the framework of analysis.

4.1 Author-level analysis

As shown in Table 2, the top 10 sequences according to the productivity and influence are listed. In the second column, <{Nicholas,D},{Huntington,P}>, <{Nicholas,D},{Williams,P}>, <{Huntington,P},{Williams,P}>, and <{Nicholas,D},{Jamali,HR}> are among the top sequences that occur most frequently. Nicholas, D, Huntington, P, Williams, P, and Jamali, HR are actually group members of the same institution, the Centre for Information Behavior and the Evaluation of Research at University College London. This suggests that when collaborating with other members in the Center, Nicholas, D tends to precede others in author lists of published papers. <{Glanzel,W},{Schubert,A}>, <{Braun,T},{Schubert,A}>, and <{Braun,T},{Glanzel,W}> are also among the most frequently occurring sequences, which indicates that these three often collaborate and thus form a relatively stable pattern of author orders. Another noticeable frequent sequence is <{Egghe,L},{Rousseau,R}>. Egghe, L and Rousseau, R are both prestigious researchers in LIS. In the second column, <{Salton,G},{Buckley,C}>, <{Jansen,BJ},{Spink,A}>, and <{Spink,A},{Saracevic,T}> are top-ranked frequent patterns in terms of their influence. Empirical evidences show that these three authors have published some high-quality papers, such as Jansen, Spink, and Saracevic (2000). <{Karahanna,E},{Straub,DW}> and <{Gefen,D},{Straub,DW}> are also frequent sequences that have generated high-quality papers. Note that Straub, DW, a senior faculty member at Georgia State University, tends to be preceded by others in the author list.

Table 2 Top 10 FSPs in citations and productivity

Author-level	According to productivity	According to influence (citation counts)
1	{Nicholas,D},{Huntington,P}	{Benbasat,I},{Dexter,AS}

2	{Egghe,L},{Rousseau,R}	{Salton,G},{Buckley,C}
3	{Nicholas,D},{Williams,P}	{Karahanna,E},{Straub,DW}
4	{Glanzel,W},{Schubert,A}	{Nelson,RR},{Todd,PA}
5	{Huntington,P},{Williams,P}	{Jansen,BJ},{Spink,A}
6	{Braun,T},{Schubert,A}	{Adams,DA},{Nelson,RR}
7	{Nicholas,D},{Jamali,HR}	{Spink,A},{Saracevic,T}
8	{Jiang,JJ},{Klein,G}	{Doll,WJ},{Torkzadeh,G}
9	{Golderman,G},{Connolly,B}	{Brancheau,JC},{Wetherbe,JC}
10	{Braun,T},{Glanzel,W}	{Gefen,D},{Straub,DW}

Figures 2 and 3 provide a comprehensive view of the productivity and influence of all frequent sequences. Similar to cases where individual author is the unit of analysis, the productivity and citation counts of frequent sequences show a power law distribution. A small number of frequent sequences therefore produce a large number of papers or citations, while a majority of frequent sequences generate a relatively small number of papers and attract relatively few citations.

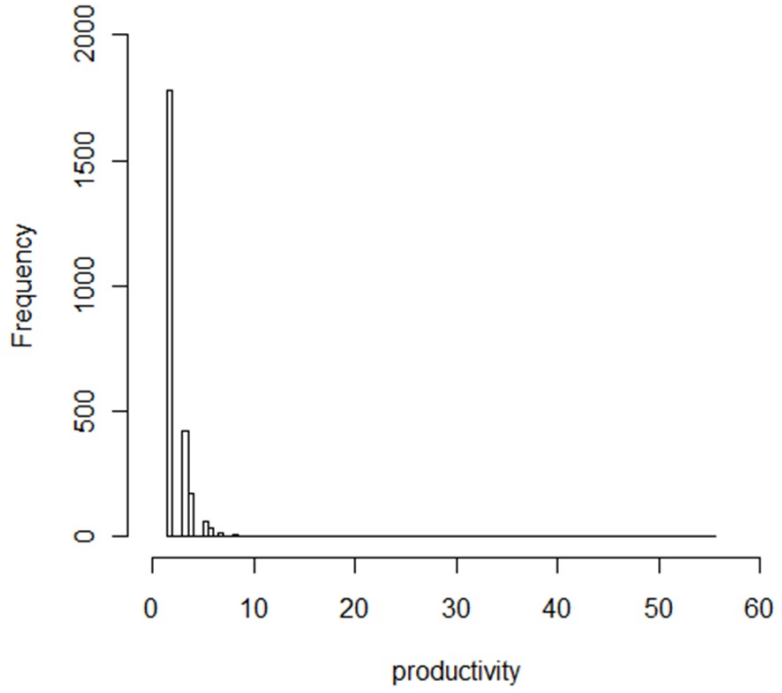

Figure 2 Frequency distribution for the productivity of frequent sequences: author-level

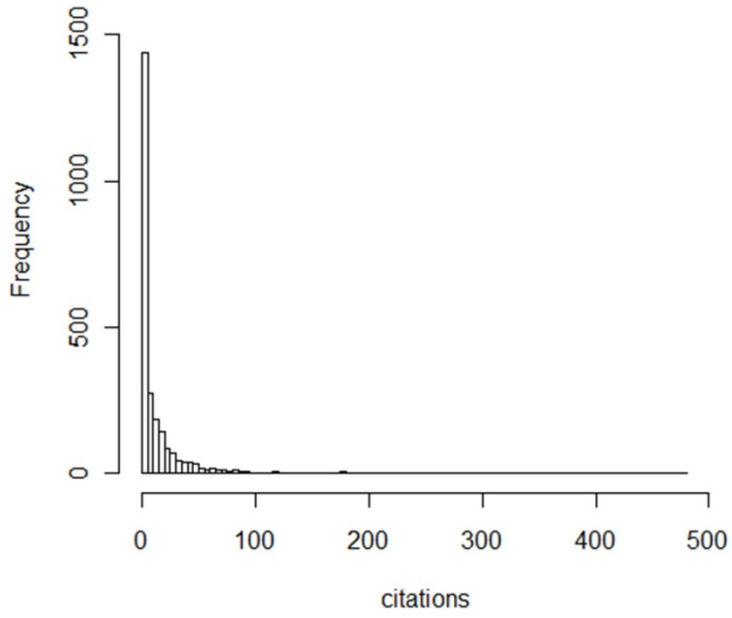

Figure 3 Frequency distribution for the influence/citations of frequent sequences: author-level

Table 3 shows the results after the aggregation of author-level frequent sequences are grouped according to the relative amount of productivity and influence of individual authors forming the sequence and corresponding indexes. The sequence pattern (c1>c2 & p1>p2) accounts for the largest amount for both the productivity and citation counts, followed closely by the sequence pattern (c1<=c2 & p1<=p2), indicating that in the field of LIS, more productive and prestigious authors tend to precede less productive and prestigious authors in author lists of published papers. In contrast with Zuckerman's (1968) study, in which he found more prestigious authors (i.e., Nobel laureates) tend to be ranked at the bottom of the author lists, the opposite situation is found in LIS. This result can be explained in that Nobel laureates are usually scientists in hard sciences that possess very different properties than LIS, which essentially resides in social sciences and the humanities. The sequence patterns (c1>c2 & p1<=p2) and (c1<=c2 & p1>p2) account for only around 20 percent for both productivity and influence. This is partly due to the significant correlation between single authors' productivity and influence.

Table 3 summarization of frequent sequences according the three criteria: author-level

Sequence patterns	Raw Prod	% of productivity	Raw citation counts	% of citations
c1>c2 & p1>p2	2,676	39.59%	18,236	45.83%
c1>c2 & p1<=p2	1,184	17.52%	6,454	16.22%
c1<=c2 & p1>p2	397	5.87%	2,016	5.07%
c1<=c2 & p1<=p2	2,502	37.02%	13,083	32.88%

4.2 Institution-level Analysis

This section provides the results for institution-level analysis. Table 4 shows the top 10 frequent sequences at the institution level. Note that the individual institutions forming a sequence might be the same as the two authors from the same institution. In the first column, a prominent trend is that many top-ranked frequent sequences either consist of the same institution or geographically close institutions. In the second column, some very prestigious universities are noted, such as Harvard University. It is interesting

that the frequent sequence < {Univ Hasselt},{Univ Antwerp}> ranked first according to productivity, while the frequent sequence <{Univ Antwerp},{Hasselt Univ}> ranked first according to citation counts. These are the two top universities among the six major universities in Belgian. Due to geographical closeness and shared first language among faculties in these two universities, it is reasonable that the productivity of the first sequence is top ranked. Meanwhile, the sharp contrast between the productivity and the influence measures between the two sequences < {Univ Hasselt},{Univ Antwerp}> and <{Univ Antwerp},{Hasselt Univ}> shows that the orders and correspondingly the implied collaboration patterns do matter in terms of their productivity and influence.

Table 4 Top 10 frequent sequences in citations and productivity

Author-Level	According To Productivity	According To Citation Counts
1	{Univ Hasselt},{Univ Antwerp}	{Univ Antwerp},{Hasselt Univ}
2	{Univ Illinois},{Univ Illinois}	{Vrije Univ Amsterdam},{Carnegie Mellon Univ}
3	{Khbo Assoc Ku Leuven},{Katholieke Univ Leuven}	{Northwestern Univ},{Northwestern Univ}
4	{Fudan Univ},{Beijing Univ Aeronaut & Astronaut}	{Partners Healthcare Syst},{Brigham & Womens Hosp}
5	{Vanderbilt Univ},{Vanderbilt Univ}	{Trinity Coll Dublin},{Vrije Univ Amsterdam}
6	{Univ Maryland},{Univ Maryland}	{Brigham & Womens Hosp},{Harvard Univ}
7	{Vrije Univ Amsterdam},{Carnegie Mellon Univ}	{Univ Pittsburgh},{Univ Pittsburgh}
8	{Univ Alberta},{Univ Alberta}	{Mit},{Mit}
9	{Penn State Univ},{Penn State Univ}	{Univ Antwerp},{Khbo}
10	{Univ Tehran},{Wolverhampton Univ}	{Econ & Social Res Inst},{Vrije Univ Amsterdam}

Similar to author-level analysis, at the institution level, the productivity and influence of frequent sequences also present power law characteristics. Table 5 gives the results of analysis at the institution level under our proposed framework of analysis. In terms of productivity, frequent sequences consisting of the same institution (i.e., sequence pattern c1=c2 & p1=p2) clearly account for the largest part, which is reasonable, since researchers in the same institution have a very convenient environment for scientific collaborations. This phenomenon is also identified by other types of studies. For example, Katz (1994) found that scientific collaboration decreases exponentially with the geographical distance between authors. By contrast, for influence citations, the sequence pattern (c1<=c2 & p1<=p2) has the highest percentage, indicating that at the institution level, more prestigious institutions tend to be preceded by other institutions in LIS in the author lists of collaborative papers. This result may be useful in ranking the LIS institutions, in that when taking the number of publications into the ranking measure, the position of listed affiliations is actually negatively related to the institution's prestige.

Table 5 summarization of frequent sequences according the three criteria: institution-level

Sequence Patterns	% of productivity	% of citation counts
c1>c2 & p1>p2	20.63%	22.39%
c1>c2 & p1<=p2	1.26%	0.00%

c1<=c2 & p1>p2	4.84%	5.97%
c1<=c2 & p1<=p2 (c1=c2 & p1=p2 excluded)	32.63%	43.78%
c1=c2 & p1=p2 (i.e., same institution)	40.63%	27.86%

4.3 State-level Analysis

In this section, results of state-level analysis are presented. Note that only records for all authors within the U.S. are included in this level of analysis. As shown in Table 6, in the second column, according to productivity, geographical effect becomes more prominent. All of the top-ranked frequent sequences at the state-level analysis consist of the same state. It is interesting to note the differences between the second column and the third column. When ranked by citation counts, some frequent sequences consisting of different states show up in the top 10, indicating collaboration over longer distances might generate influential papers.

Table 6 Top 10 frequent sequences in citations and productivity

Author-level	According to productivity	According to citation counts
1	{MA},{MA}	{MA},{MA}
2	{CA},{CA}	{CA},{MI}
3	{IL},{IL}	{MA},{CA}
4	{NY},{NY}	{PA},{WI}
5	{PA},{PA}	{WI},{PA}
6	{MD},{MD}	{IL},{IL}
7	{OH},{OH}	{DC},{DC}
8	{TX},{TX}	{CA},{CA}
9	{DC},{DC}	{MA},{NY}
10	{IN},{IN}	{MD},{MD}

Without surprise, the distribution of productivity and influence also present the power law shape. As shown in Table 7, geographical closeness dominates the sequence patterns in both productivity and citations. Frequent sequences consisting of the same state make up the largest percentage in both columns. Meanwhile, the second most-popular sequence pattern is (c1<=c2 & p1<=p2), similar to the case at the institution level, suggesting that more scientifically prestigious states tend to be preceded by others in author lists of collaborative papers.

Table 7 summarization of frequent sequences according the three criteria: state-level

Sequence Patterns	Prod percentage (overlap exists)	Citation percentage
c1>c2 & p1>p2	25.31%	29.94%
c1>c2 & p1<=p2	2.96%	2.64%
c1<=c2 & p1>p2	3.09%	2.07%
c1<=c2 & p1<=p2 (c1=c2 & p1=p2 excluded)	29.26%	31.26%
c1=c2 & p1=p2 (i.e., same state)	39.38%	34.09%

4.4 Nation-level analysis

This section presents the results for the nation-level analysis. Note that only the records with at least one author from outside the U.S. are included in this level of analysis. Table 8 shows the top 20 frequent patterns with regard to productivity and citation counts. It is not surprising that the sequence < {USA}, {USA}> ranks first in both productivity and citation counts. The geographic closeness, language boundaries, and cultural differences may explain the frequent sequences consisting of the same country, as well as English-speaking countries (e.g., < {Canada}, {USA}>) tending to rank high in productivity. In a relative sense, frequent sequences consisting of different countries tend to rank higher in citation counts than productivity. Also, it's interesting to note that the difference in ranking between a frequent sequence and its reversed correspondence. For example, < {Netherlands}, {USA}> ranks 13th with respect to productivity and fourth in citation counts.

Table 8 Top 20 frequent patterns in citations and productivity

Author-level	According to productivity	According to citation counts
1	<{USA},{USA}>	<{USA},{USA}>
2	<{Belgium},{Belgium}>	<{United_Kingdom},{United_Kingdom}>
3	<{Peoples_R_China},{Peoples_R_China}>	<{France},{USA}>
4	<{Peoples_R_China},{USA}>	<{Netherlands},{USA}>
5	<{Canada},{USA}>	<{Belgium},{Belgium}>
6	<{USA},{Peoples_R_China}>	<{United_Kingdom},{Netherlands}>
7	<{United_Kingdom},{USA}>	<{USA},{United_Kingdom}>
8	<{USA},{United_Kingdom}>	<{Ireland},{Netherlands}>
9	<{USA},{Canada}>	<{Netherlands},{United_Kingdom}>
10	<{South_Korea},{USA}>	<{United_Kingdom},{France}>
11	<{USA},{Taiwan}>	<{Switzerland},{United_Kingdom}>
12	<{Taiwan},{Taiwan}>	<{Peoples_R_China},{Belgium}>
13	<{Netherlands},{USA}>	<{Belgium},{India}>
14	<{United_Kingdom},{United_Kingdom}>	<{Taiwan},{Taiwan}>
15	<{Belgium},{Peoples_R_China}>	<{USA},{Canada}>
16	<{Peoples_R_China},{Belgium}>	<{Iran},{United_Kingdom}>
17	<{Taiwan},{USA}>	<{USA},{Spain}>
18	<{Netherlands},{United_Kingdom}>	<{Peoples_R_China},{Peoples_R_China}>
19	<{Iran},{United_Kingdom}>	<{U_Arab_Emirates},{USA}>
20	<{USA},{South_Korea}>	<{United_Kingdom},{Switzerland}>

At the nation level, the productivity and influence of frequent sequences still follow the power law distribution. Table 9 presents the results under our proposed framework for the nation-level analysis. The sequence pattern (c1<=c2 & p1<=p2) accounts for the largest part, followed by sequence pattern c1>c2 & p1>p2.

Table 9 summarization of frequent sequences according the three criteria: nation-level

Sequence Patterns	% of productivity	% of citation counts
c1>c2 & p1>p2	33.33%	34.68%
c1>c2 & p1<=p2	2.62%	1.76%

c1<=c2 & p1>p2	2.62%	3.70%
c1<=c2 & p1<=p2 (c1=c2 & p1=p2 excluded)	38.14%	37.32%
c1=c2 & p1=p2 (i.e., same country)	23.28%	22.54%

5 Discussion and Conclusion

In this paper, we propose a new framework for analyzing author orders in scientific collaboration. Sequential pattern mining is introduced and applied to author orders in LIS. Investigations are conducted at different levels of integration, including author-level, institution-level, state-level, and nation-level analysis. Our results can be summarized as follows:

- the productivity and influence approximately follows the power law for frequent sequences in all levels of analysis;
- for frequent sequences of author level, the pattern of "the more productive and prestigious author ranking ahead" (i.e., frequent sequences <author A, author B> with author A as the more productive and more prestigious one) is the one with the highest productivity and influence; however, in the rest of the levels, the pattern with the highest productivity and influence is the one with "the less productive and prestigious institutions/states/nations ranking ahead" (i.e., frequent sequences <A, B> with A as the less productive and less prestigious one; sequences consisted of the same entity are not considered here); and
- for frequent sequences of the institution level, state level, and nation level, close geographic locations, implying language boundaries, different native cultures, cost of communication, etc., usually co-present with high productivities, while distant locations tend to co-occur with high citation counts.

While the results of this paper are based on LIS research, we argue that the framework of analysis is readily applicable to other scientific disciplines, as we only make very general assumptions that: (1) scientific collaboration patterns can be captured through author lists of multi-authored publications; (2) the number of publications reflects the productivity; and (3) the citation counts reflect the influence. One limitation of our paper is that when assigning credit (measured by citation counts) to the frequent author sequences generated from one paper, we do not consider the position of the author sequences in the original author lists. In other words, the frequent author sequences in the front are assigned the same influence measure with the author sequences in the back. This might introduce biases in individual cases. Another limitation is that authors with the same last name and the same initial of the middle and/or the first name are not distinguishable in the author-level analysis (no such problem exists for other levels of analysis).

With the dominant practice of multi-authored papers in modern academic communities, author orders and credit assignment will continue to gain attention. Due to the relatively complicated data structure of author orders, a rich collection of data analysis tools developed in the field of data mining could be very useful to address research challenges. With the prevalence of collaboration, the unit of analysis in the field of scholarly communication is shifting from individual to relationship (He, Ding, & Ni, 2011) to sequence, which might be an emerging paradigm. Our future work includes investigating the dynamics of author sequences across time. The temporal dynamics of the productivity and the influence of sequences

containing the same authors will be tracked and analyzed. The emerging patterns of author sequences will be investigated in difference time spans. We will show the patterns of author ordering that generate the highest productivity and the highest influence respectively within each time span

6 References

Agrawal R., Srikant, R. (1995). Mining sequential patterns. *Proceedings of the 1995 International Conference on Data Engineering*, 3-14.

Baerlocher, M. O., Newton, M., Gautam, T., Tomlinson, G., & Detsky, A. S. (2007). The meaning of author order in medical research. *Journal of Investigative Medicine, 55*(4), 174-180. 110.2310/6650.2007.06044.

ABuchta, C., & Hahsler, M. (2010). arulesSequences: Mining frequent sequences. R package version 0.1-11. http://CRAN.R-project.org/package=arulesSequences

Burman, K. D. (1982). "Hanging from the masthead": Reflections on authorship. *Annals of Internal Medicine, 97*(4), 602-605.

Endersby, W. J. (1996). Collaborative research in the social sciences: Multiple authorship and publication credit. *Social Science Quarterly, 77*, 375-392.

Fine, M. A., & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist, 48*(11), 1141-1147. doi:10.1037/0003-066X.48.11.1141

Floyd, S. W., Schroeder, D. M., & Finn, D. M. (1994). "Only if I'm first author": Conflict over credit in management scholarship. *The Academy of Management Journal, 37*(3), 734-747.

Hahsler, M., & Hornik, K. (2007). Building on the arules infrastructure for analyzing transaction data with R. In R. Decker & H. J. Lenz (Eds.), *Advances in Data Analysis.* Berlin Heidelberg: Springer-Verlag, 449-456.

Hunt, J. G., & Blair, J. D. (1987). Content, process, and the Matthew effect among management academics. *Journal of Management, 13*(2), 191-210.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227. doi: 10.1016/S0306-4573(99)00056-4

Joseph, K., Laband, D. N., & Patil, V. (2005). Author order and research quality. *Southern Economic Journal, 71*(3), 545-555.

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

Laurance, W. F. (2006). Second thoughts on who goes where in author lists, *Nature Publishing Group, 442*, 26-26.

Mary Ann Von, G., & Novelli, L., Jr. (1982). Ethical standards within organizational behavior. *The Academy of Management Journal*, *25*(2), 417-436.

Peffers, K., & Hui, W. (2003). Collaboration and author order: Changing patterns in IS research. *Communications of the Association for Information Systems, 11*(10). Retrieved from http://aisel.aisnet.org/cais/vol11/iss1/10

Pei, J, Han, J, & Lakshmanan, L.V.S. (2001). Mining frequent itemsets with convertible constraints. *Proceeding of the 2001 International Conference on Data Engineering,* 433-332.

Pei, J, Han, J, Mortazavi-Asl, B, Wang, J, Pinto, H, Chen, Q, Dayal, U, & Hsu M. C (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Trans Knowledge Data Engineering, 16*, 1424-1440.

Rennie, D., & Flanagin, A. (1994). Authorship! authorship! guests, ghosts, grafters, and the two-sided coin. *The Journal of the American Medical Association, 271*(6), 469-471.

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. *The Journal of the American Medical Association, 278*(7), 579-585.

Riesenberg, D., & Lundberg, G. D. (1990). The order of authorship: Who's on first? *The Journal of the American Medical Association, 264*(14), 1857.

Savitz, D. A. (1999). Invited commentary: What can we infer from author order in epidemiology? *American Journal of Epidemiology, 149*(5), 401-403.

Srikant, R, Agrawal, R (1996). Mining sequential patterns: Generalizations and performance improvements. *Proceeding of the 5th International Conference on Extending Database Technology,* 3-17.

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multiauthored publications. *PLoS Biology*, 13-14. Retrieved from http://ezproxy.lib.indiana.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=25450919&site=ehost-live

Yan, E., & Sugimoto, C. R. (2011). Institutional interactions: Exploring the social, cognitive, and geographic relationships between institutions as demonstrated through citation networks. *Journal of the American Society for Information Science and Technology*, *62*(8), 1498–1514

Zaki, M. J. (1998). Efficient enumeration of frequent sequences. *Proceeding of the 7th International Conference on Information and Knowledge Management,* 68-75.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning, 40*, 31-60.

Zuckerman, H. A. (1968). Patterns of name ordering among authors of scientific papers: A study of social symbolism and its ambiguity. *American Journal of Sociology, 74*(3), 276-291.