Open Research Online

The Open University's repository of research publications
and other research outputs

Research Collaboration Analysis Using Text and Graph Features

Conference or Workshop Item

oro.open.ac.uk

Research Collaboration Analysis Using Text and Graph Features

Drahomira Herrmannova[1] and Petr Knoth[2] and Christopher Stahl[1] and
Robert Patton[1] and Jack Wells[1]

Oak Ridge National Laboratory and The Open University
Oak Ridge, TN, USA; Milton Keynes, UK
{herrmannovad, stahlcg, pattonrm, wellsjc}@ornl.gov; petr.knoth@open.ac.uk

Abstract. Patterns of scientific collaboration and their effect on scientific production have been the subject of many studies. In this paper we analyze the nature of ties between co-authors and study collaboration patterns in science from the perspective of semantic similarity of authors who wrote a paper together and the strength of ties between these authors (i.e. how much have they previously collaborated together). These two views of scientific collaboration are used to analyze publications in the TrueImpactDataset [11], a new dataset containing two types of publications – publications regarded as seminal and publications regarded as literature reviews by field experts. We show there are distinct differences between seminal publications and literature reviews in terms of author similarity and the strength of ties between their authors. In particular, we find that seminal publications tend to be written by authors who have previously worked on dissimilar problems (i.e. authors from different fields or even disciplines), and by authors who are not frequent collaborators. On the other hand, literature reviews in our dataset tend to be the result of an established collaboration within a discipline. This demonstrates that our method provides meaningful information about potential future impacts of a publication which does not require citation information.

1 Introduction

It has been shown scientific authorship is shifting from single-authored publications towards team production [23] and international collaboration [22]. Consequently, many studies have focused on scientific collaboration networks [17], patterns of scientific collaboration across disciplines [4], and how these patterns affect scientific production and impact [9]. Many such studies have focused on the concept of "bridges" – collaboration ties formed by authors from different communities or fields which create bridges between these distinct communities or fields [8]. Within this area, it has been shown that newcomers in a group of collaborators can increase the impact of the group [9], and that high impact scientific production occurs when scientists create connections across otherwise disconnected communities from different knowledge domains [14].

Existing works studying scientific collaboration networks have often focused either on properties of the network or on topical information pertaining to the nodes in the network. In this work we develop and test an approach which combines both network and topical information about the nodes. In order to gain insight into the types of collaboration between authors, we investigate the possibility of utilizing semantic distance in co-authorship networks together with the concept of *research endogamy* [16] – the tendency to collaborate with the same authors or within a group of authors; and study how these types of collaboration reflect scientific importance.

In contrast to previous studies combining topical and network information [6,12], our approach is beneficial in that it does not require citation information or a complete network, and can therefore be applied to newly published works. This approach, which we have introduced in a previous publication [11], belongs to a class of methods referred to as "semantometrics" [13]. In contrast to the existing metrics such as bibliometrics, altmetrics or webometrics, which are based on measuring the number of interactions in the scholarly network, semantometrics build on the premise that full-text is needed to understand scholarly publication networks and the value of publications.

The content of this paper is organized as follows. First, in Section 2, we discuss previous work related to our research, and our motivation for utilizing research endogamy and semantic distance of authors. In Section 3, we first define research endogamy and author distance and present a classification of research publications created using these two measures. Next, in Section 3.1 we describe our methodology and in Section 3.2 we describe the dataset used in our study.

2 Related Work

In this section, we review previous literature relevant to our study. First, we discuss methods for measuring the strength of ties in academic social networks, particularly research endogamy. Next, we briefly discuss methods for detecting communities in scholarly networks.

2.1 Strength of Ties in Academic Social Networks

Academic social networks represent relationships among researchers. Uncovering and studying patterns of academic social networks has been applied to many problems ranging from identifying influential researchers [5] and ranking conferences [19] to measuring research contribution [18] and the diffusion of innovation [21]. One of the first studies focusing on the strength of ties in social networks [8] introduced the concept "weak ties", i.e. ties across rather than within different communities or groups, and discussed the importance of these ties for diffusion processes. This has later been applied to studying academic social networks [4]. The tactic used to measure the strength of the tie between two individuals has in this case been to measure the proportion of common ties shared by the two individuals [8]. Other approaches used to measure the strength of ties have been

the frequency of contact [7], mutual acknowledgement of contact [4], or the likelihood of a tie re-appearing in the future [1]. [17] has proposed a measure of closeness of two authors which combines information about how many papers two authors wrote together and the number of other collaborators with whom they wrote them.

Following the ideas of [8] and later [9], who classified agents in a network as incumbents and newcomers, and have shown newcomers to a group help to improve its performance, [16] have used the degree of new collaborations to rank conferences. The degree of new collaborations has been quantified using a new indicator called "research endogamy", which captures the inclination of a group to usually collaborate together. [16] have shown the reputability of computer science conferences is correlated with the endogamy of their authors – low endogamy (i.e. less frequent collaboration) tends to be associated with highly reputed conferences, while lower quality conferences tend to publish articles by authors who have collaborated together on many occasions. [19] have applied the concept of endogamy to ranking publications and patents, and have shown low endogamy publications tend to receive more citations.

Overall, the aforementioned studies demonstrate the importance of connections across communities, diverse collaborations, and newcomers to a group. These patterns tend to be associated with high impact academic production. Hence, in this work, we use the concept of research endogamy of publications as defined by [19] to measure the strength of collaboration of a group of authors.

2.2 Semantic Similarity for Community Detection

Two approaches commonly used to detect communities in academic social networks are: (1) using the graph structure of the network or (2) using textual information of the nodes, e.g. by calculating semantic similarity between the nodes [3]. These two approaches have also been used together to create maps of scientific communities in a specific field [6,12] and to identify similar researchers [2]. However, the network-based approach poses a significant challenge. Community detection in incomplete networks is a challenging task which requires the use of non-traditional methods [15]. However, the complete network may not always be available, or may be difficult to obtain. For example, in order to identify whether two authors are members of the same community or of different communities, complete information about each of their communities (other authors and links between them) are needed.

Furthermore, network-based community detection has been shown to result in communities which span diverse topics, while text-based community detection helps in detecting nodes focusing on a specific topic [3]. As we are interested in studying individual publications for which we may not have complete neighborhood information, we chose the text-based approach, and use semantic distance (the inverse of similarity) to measure the similarity of authors. This is also beneficial, as the textual similarity provides information complementary to the endogamy measure, which is calculated using topological information. By combining these two approaches, we are able to study collaboration networks

not only from the perspective of tie strength, but also from the perspective of whether each tie represents potential knowledge transfer within or across disciplines.

3   Approach and Dataset

In [10], we have proposed a classification of research publications in which publications are divided into four groups (Figure 1) according to the semantic distance and the strength of ties between the publications' authors. In this paper, we provide an evaluation of this approach. To do this, we use the recently released TrueImpactDataset [11] which contains publications of two types, seminal publications and literature reviews, and compare the collaboration patters of these two types of publications in terms of author distance and collaboration frequency.
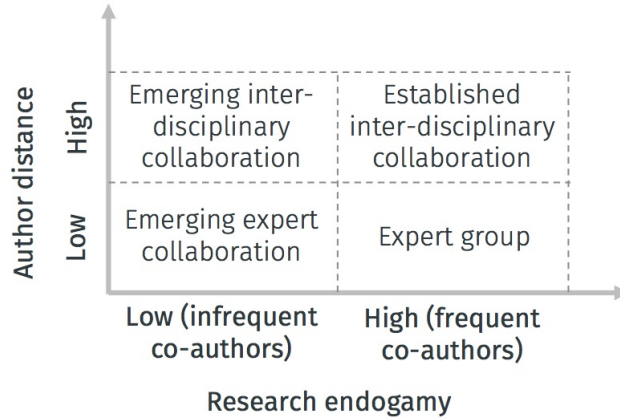

Fig. 1. Types of research collaboration based on semantic distance of authors, and their collaboration frequency.

The semantic distance of a pair of authors is calculated using their previous publication record. Figure 2 illustrates which publications are used in the calculation. For example, distance between authors a_1 and a_2 in Figure 2 would be calculated using distance between the following two sets of publications: $\{p_1, p_2, p\}$ and $\{p, p_3, p_4\}$. Specifically, we measure the semantic distance $d(p)$ between authors of publication p as a mean semantic distance between all pairs of authors:

$$d(p) = \frac{1}{|A(p)| \cdot (|A(p)| - 1)} \sum_{a_i \in A(p), a_j \in A(p), a_i \neq a_j} d(a_i, a_j) \qquad (1)$$

Here $A(p)$ is a set of authors of publication p. As explained in [10], we calculate the distance for a pair of authors $d(a_i, a_j)$ by concatenating the publications
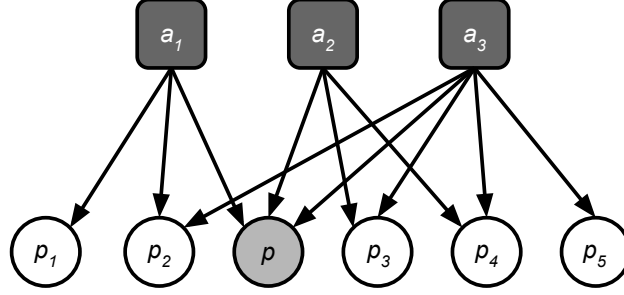
Fig. 2. A sample network showing the set of publications (round nodes) and authors (squared nodes) used in the calculation of author distance and research endogamy of publication p.

of each author into a single document. While this is a very simplistic approach, it is also beneficial in terms of complexity of the calculation.

In order to measure the strength of ties between authors, we combine the semantic distance with research endogamy value of the publication. Research endogamy [16] is the tendency to collaborate with the same authors or within a group of authors. The research endogamy of a publication is calculated based on research endogamy of a set of authors A, which is defined similarly as the Jaccard similarity coefficient [16,19] (Equation 2). The authors and publications used in the calculation are depicted in Figure 2. The research endogamy $e(A)$ of a set of authors is calculated as follows:

$$e(A) = \frac{|\bigcap_{a \in A} P(a)|}{|\bigcup_{a \in A} P(a)|} \tag{2}$$

Here $P(a)$ represents a set of papers written by author a. Higher endogamy value is related to more frequent collaboration between authors in A – a value of 1 means all authors in A have written all of their publications together. On the other hand, a group of authors who have never collaborated together will have an endogamy value of 0. For example, the research endogamy of authors a_1 and a_2 in Figure 2 is 1/5, while the endogamy of authors a_2 and a_3 is 3/5, i.e. authors a_2 and a_3 tend to collaborate more frequently than authors a_1 and a_2.

Endogamy of a publication p is then defined as a mean of endogamy values of the power set of its authors [16,19] (Equation 3).

$$e(p) = \frac{\sum_{x \in L(p)} e(x)}{|L(p)|} \tag{3}$$

Here $L(p)$ is the set of all subsets with at least two authors of p, $L(p) = \bigcup_{k=2}^{k=|A(p)|} L_k(p)$, where $L_k(p) = C(A(p), k)$ is the set of all subsets of $A(p)$ of length k.

3.1 Methodology

To study the relation between author distance and research endogamy we use our TrueImpactDataset, a multidisciplinary dataset of research publications containing seminal publications and literature reviews. We are interested in how these two types of papers are situated with regard to author distance and research endogamy. However, we also look at whether the two measures relate to the number of citations each publication received. A correlation would suggest the two metrics could potentially assist in predicting the future citation counts. Finally, we compare research endogamy and author distance, and citation counts in terms of how well each method distinguishes between seminal publications and literature reviews.

We use the following methodology. For the publications in the dataset we collect and/or calculate the following measures: (1) author distance, (2) research endogamy, (3) collaboration category (assigned to publications using author distance and research endogamy, Figure 1), (4) total number of citations per publication, (5) number of citations normalized by number of authors, and (6) number of citations normalized by publication age. To compare seminal publications and literature reviews in our dataset with regards to author distance and research endogamy we use t and χ^2 tests to determine whether the values of the measures are statistically significant for seminal publications and literature reviews. To analyze whether author distance and research endogamy help in distinguishing between seminal publications and literature reviews in our dataset we also analyze the distributions of both features and the placement of seminal publications and literature reviews within the four collaboration categories (Figure 1).

3.2 Data

To collect all data needed for studying the measures introduced in Section 3, we have used three data sources:

1. TrueImpactDataset[1] [11], which provides us with seminal publications and literature reviews (i.e. the p node in Figure 2),
2. Microsoft Academic (MA) API[2] [20] which we use to collect metadata (particularly the information about authors and their publications, gray and yellow nodes in Figure 2) of the papers in the TrueImpactDataset,
3. Mendeley API[3] which we use to collect publication abstracts.

Table 1 shows the size of the dataset. After collecting all needed data the size of the original dataset was reduced to 144 publications (i.e. publications for which we were able to obtain author information) – 75 literature reviews and 69 seminal publications. The rows *Total authors* and *Unique authors* show the total number of authors of all papers in the dataset and the number of unique

---

[1] `trueimpactdataset.semantometrics.org/`
[2] `aka.ms/academicgraph/`
[3] `dev.mendeley.com/`

author names, respectively. To count the unique names, we have compared the surname and all first name initials, in case of a match we consider the names to be the same (e.g. J. Adam Smith and John A. Smith will be counted as one unique name). The number of unique author names doesn't show the number of disambiguated authors, but gives us an indication of how many of the author names repeat in the dataset.

Table 1. Dataset size. The table shows for how many of the TrueImpactDataset publications we managed to get the needed metadata and how many additional publications we collected (i.e. including all other publications of the authors in the TrueImpact-Dataset – row *Total number of publications*).

Publications in TrueImpactDataset	314
TrueImpactDataset publications in MA	298
Pubs with author information in MA	144
Total authors	758
Unique authors	727
Total number of publications	27,653

4 Experiments

We begin by comparing the properties of survey publications and literature reviews. We investigate how these two types of papers are situated with regard to the extracted features. To do this, we use the following methodology: we take all of the 144 core papers and for each of them collect the features defined in section 3.1. To understand whether seminal publications and literature reviews differ in terms of these features we calculate an independent one-tailed t-test for each feature except for the collaboration category feature which is categorical and for which we calculate χ^2 test. The t-test is a measure commonly used to assess whether two sets of data are statistically different from each other. In other words, it helps to determine the features that can distinguish survey papers from seminal papers. To test the significance, we set the significance threshold at 0.05. Furthermore, for each feature we create a histogram and by comparing these histograms for the two publication types we gain insight into norms and placement of seminal and survey publications in terms of metrics.

The complete results of the t-test are presented in Table 2 and the histograms for the five numerical features are shown in Figure 3. For four of the features we reject the null hypothesis of equal means. The t-test tells us the values of these four features are significantly different for the two sets of papers.

Next, we analyze the collaboration category feature which is assigned to publications using the values of author distance and research endogamy (Figure 1). We calculate χ^2 test, which is a statistical test for categorical variables for testing whether the means of two groups are the same, to test whether the
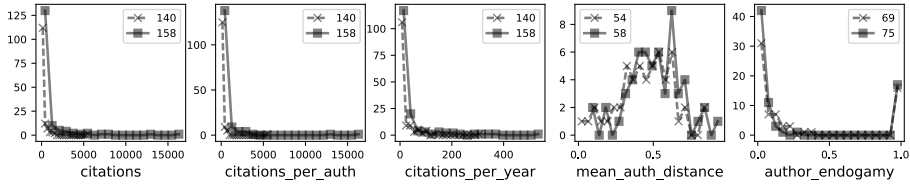
Fig. 3. Histograms of the five numerical features.

Table 2. Results of t- and χ^2 tests.

Metric	p-value
Mean author distance	0.0327
Endogamy	0.3217
Citations	0.0012
Citations per year	0.0073
Citations per author	0.0110
Collaboration category	0.0218

seminal publications and literature reviews differ in terms of the collaboration category. The resulting p-value is 0.0218 (Table 2), which is lower than our significance threshold of 0.05. This tells us that the means of the two sets of papers differ.

The relation between author distance and research endogamy is shown in Figure 4. The labels in the figure correspond to the four collaboration categories presented in Figure 1. Each point in the figure represents one publication, with seminal publications and literature reviews distinguished by color. The horizontal and vertical lines in the figure represent median values for each axis – the vertical line represents median endogamy value (0.0297) and the horizontal line represents median author distance value (0.4996). The median values were used to separate the publications into the four categories (Figure 1). Figures 4 and 3 show the endogamy values for the dataset are strongly skewed towards 0. Furthermore, the results of the t-test suggest research endogamy by itself does not distinguish between the two publication types. However, when combined with the author distance measure, a clear pattern emerges. This is visible in Figure 5, which shows number of publications of each types belonging to each collaboration category.

Figure 5 shows there are some differences between seminal publications and literature reviews. In particular, the main difference between the two classes is that emerging collaborations (i.e. when the authors have not collaborated frequently together previously) are in our dataset more common for seminal publications. On the other hand, literature reviews seem to be a result of established collaborations within a discipline. These observations are consistent with previous studies which have shown that cross-community citation and collabo-
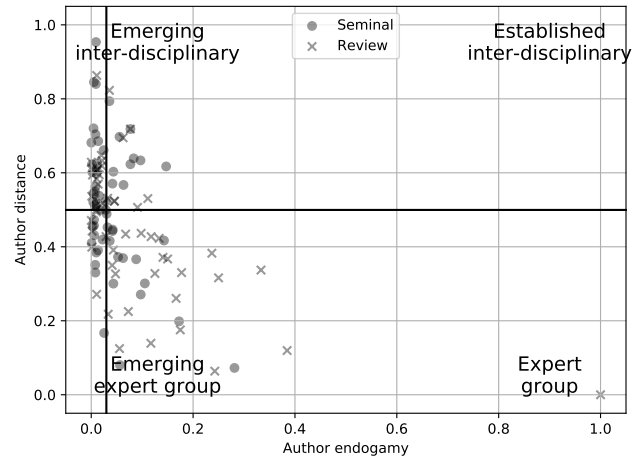
Fig. 4. Distribution of publications according to author distance and author endogamy. The horizontal and vertical lines are used to separate the publications into the four quadrants presented in Figure 1.
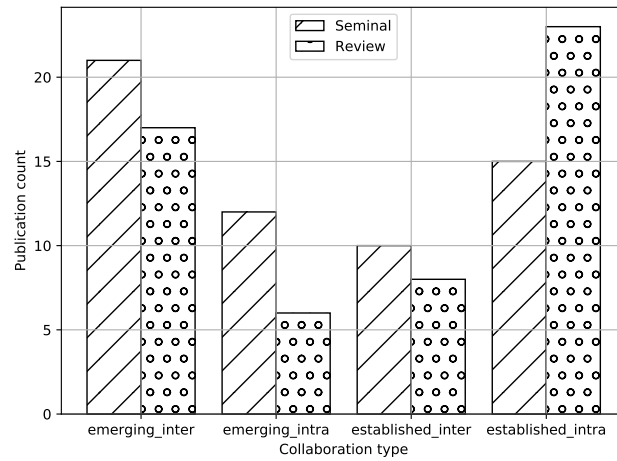

Fig. 5. Number of publications belonging to each collaboration category across both publication types.

ration patters are characteristic for high impact scientific production [9,14,16]. We believe this is an encouraging result which suggest semantic distance of authors combined with their endogamy value might be helpful in providing early indication of future impacts of a publication.

5 Conclusions

This paper studied the relationship between semantic distance of authors which collaborated on a publication and the strength of ties between these authors, which was assess using research endogamy measure (a measure of collaboration frequency introduced by [16]). More specifically, we compared publications of two types – seminal publications and literature reviews – in terms of their author distance and research endogamy values. Our results show that there are distinct differences between these two publication types in terms of collaboration patters. In particular, we found that seminal publications tend to be written by authors who have previously worked on dissimilar problems (i.e. authors from different fields or even disciplines), and by authors who are not frequent collaborators (i.e. emerging inter-disciplinary collaborations). On the other hand, literature reviews in our dataset tend to be the result of an established collaboration within a discipline (an "expert group"). This demonstrates content analysis might provide valuable information for research evaluation and meaningful information about potential future impacts of a publication which does not require citation information.

6 Bibliographical References

References

1. Michele A Brandao, POS Vaz de Melo, and Mirella M Moro. Tie strength persistence and transformation. *AMW (to appear)*, 2017.
2. Guillaume Cabanac. Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87(3):597–620, 2011.
3. Ying Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498–514, 2011.
4. Noah Friedkin. A test of structural features of granovetter's strength of weak ties theory. *Social networks*, 2(4):411–422, 1980.
5. Tom ZJ Fu, Qianqian Song, and Dah Ming Chiu. The academic social network. *Scientometrics*, 101(1):203–239, 2014.
6. Patrick Glenisson, Wolfgang Glänzel, Frizo Janssens, and Bart De Moor. Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6):1548–1572, 2005.
7. Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, pages 201–233, 1983.
8. Mark S Granovetter. The strength of weak ties. In *American Journal of Sociology*, volume 78, pages 1360–1380. Elsevier, 1973.
9. Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(April):697–702, 2005.
10. Drahomira Herrmannova and Petr Knoth. Semantometrics in coauthorship networks: Fulltext-based approach for analysing patterns of research collaboration. *D-Lib Magazine*, 21(11/12), 2015.

11. Drahomira Herrmannova, Robert M. Patton, Petr Knoth, and Christopher G. Stahl. Citations and readership are poor indicators of research excellence: Introducing trueimpactdataset, a new dataset for validating research evaluation metrics. In *Proceedings of the 1st Workshop on Scholarly Web Mining*, 2017.

12. Frizo Janssens, Jacqueline Leta, Wolfgang Glänzel, and Bart De Moor. Towards mapping library and information science. *Information processing & management*, 42(6):1614–1642, 2006.

13. Petr Knoth and Drahomira Herrmannova. Towards semantometrics: A new semantic similarity based measure for assessing a research publication's contribution. *D-Lib Magazine*, 20(11):8, 2014.

14. Renaud Lambiotte and Pietro Panzarasa. Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190, 2009.

15. Wangqun Lin, Xiangnan Kong, Philip S Yu, Quanyuan Wu, Yan Jia, and Chuan Li. Community detection in incomplete information networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 341–350. ACM, 2012.

16. Sergio Lopez Montolio, David Dominguez-Sal, and Josep Lluis Larriba-Pey. Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, 42(2):11–16, 2013.

17. Mark EJ Newman. Who is the best connected scientist? a study of scientific coauthorship networks. In *Complex networks*, pages 337–370. Springer, 2004.

18. Lais Rocha and Mirella M Moro. Research contribution as a measure of influence. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2259–2260. ACM, 2016.

19. Thiago H. P. Silva, Mirella M. Moro, Ana Paula C. Silva, Wagner Meira Jr., and Alberto H. F. Laender. Community-based Endogamy as an Influence Indicator. In *Digital Libraries 2014 Proceedings*, page 10, 2014.

20. Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

21. Thomas W Valente. Social network thresholds in the diffusion of innovations. *Social networks*, 18(1):69–89, 1996.

22. A Witze. Research gets increasingly international. *Nature*, 785:6–8, 2016.

23. Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.