Text and Graph Based Approach for Analyzing Patterns of Research Collaboration: An analysis of the TrueImpactDataset

Drahomira Herrmannova[†], Petr Knoth[‡], Christopher Stahl[†], Robert Patton[†], Jack Wells[†]
[†]Oak Ridge National Laboratory; [‡]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 frequently have they previously collaborated together). These two views of scientific collaboration are used to analyze publications in the TrueImpactDataset (Herrmannova et al., 2017) (Herrmannova et al., 2017), 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.

Keywords: collaboration networks, publication impact, text mining, semantic similarity, semantometrics

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan[1].

1. Introduction

Many studies have focused on scientific collaboration networks (Newman, 2004), patterns of scientific collaboration across disciplines (Friedkin, 1980), and on how these patterns affect scientific production and impact (Guimerà et al., 2005). Within this area, it has been shown that newcomers in a group of collaborators can increase the impact of the group (Guimerà et al., 2005), and that high impact scientific production occurs when scientists create connections across otherwise disconnected communities from different knowledge domains (Lambiotte and Panzarasa, 2009).

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 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* (Montolio et al., 2013) – 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 (Glenisson et al., 2005; Janssens et al., 2006), 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 (Herrmannova et al., 2017), belongs to a class of methods referred to as "semantometrics" (Knoth and Herrmannova, 2014). 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. In this work we test our approach on a dataset of publications regarded as seminal and publications regarded as literature reviews by field experts, and compare these two publication types in terms of collaboration patterns.

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

Uncovering and studying patterns of academic social networks has been applied to many problems ranging from identifying influential researchers (Fu et al., 2014) and ranking conferences (Silva et al., 2014) to measuring re-

[1] http://energy.gov/downloads/doe-public-access-plan

search contribution (Rocha and Moro, 2016) and the diffusion of innovation (Valente, 1996). One of the first studies focusing on the strength of ties in social networks (Granovetter, 1973) 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. 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 (Granovetter, 1973). Other approaches used to measure the strength of ties have been the frequency of contact (Granovetter, 1983), mutual acknowledgement of contact (Friedkin, 1980), or the likelihood of a tie re-appearing in the future (Brandão et al., 2017). (Newman, 2004) 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 (Granovetter, 1973) and later (Guimerà et al., 2005), who classified agents in a network as incumbents and newcomers, and have shown newcomers to a group help to improve its performance, (Montolio et al., 2013) 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. (Montolio et al., 2013) 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. (Silva et al., 2014) 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 (Silva et al., 2014) 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 (Ding, 2011). These two approaches have also been used together to create maps of scientific communities in a specific field (Glenisson et al., 2005; Janssens et al., 2006) and to identify similar researchers (Cabanac, 2011). 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 (Lin et al., 2012). 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 (Ding, 2011). 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 (Herrmannova and Knoth, 2015), 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 (Herrmannova et al., 2017) (Herrmannova et al., 2017) 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.

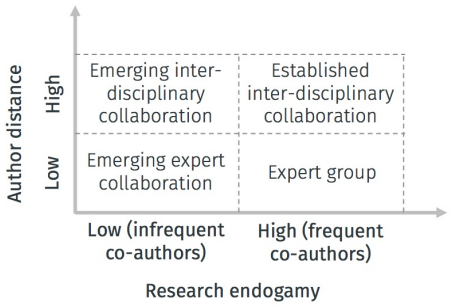

Figure 1: Types of research collaboration based on semantic distance of authors, and their collaboration frequency.

The semantic distance of a pair of authors is calculated using their previous publication record.

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

Here $A(p)$ is a set of authors of publication p. As explained in (Herrmannova and Knoth, 2015), we calculate the distance for a pair of authors $d(a_i, a_j)$ by concatenating the publications 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 (Montolio et al., 2013) 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 (Montolio et al., 2013; Silva et al., 2014) (Equation 2). The research endogamy $e(A)$ of a set of authors is calculated as follows:

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

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.
Endogamy of a publication p is then defined as a mean of endogamy values of the power set of its authors (Montolio et al., 2013; Silva et al., 2014) (Equation 3).

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

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 (Herrmannova et al., 2017), 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. 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[2] (Herrmannova et al., 2017) (Herrmannova et al., 2017), which provides us with seminal publications and literature reviews,
2. Microsoft Academic (MA) API[3] (Sinha et al., 2015) which we use to collect metadata (particularly the information about authors and their publications) of the papers in the TrueImpactDataset,
3. Mendeley API[4] 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 row *Number of authors* shows the total number of (non-disambiguated) authors of all papers in the dataset.

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

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 TrueImpactDataset – row *Total number of publications*).

4. Experiments

In this section, we investigate how seminal publications and literature reviews 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 2. For four of the features we reject the

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

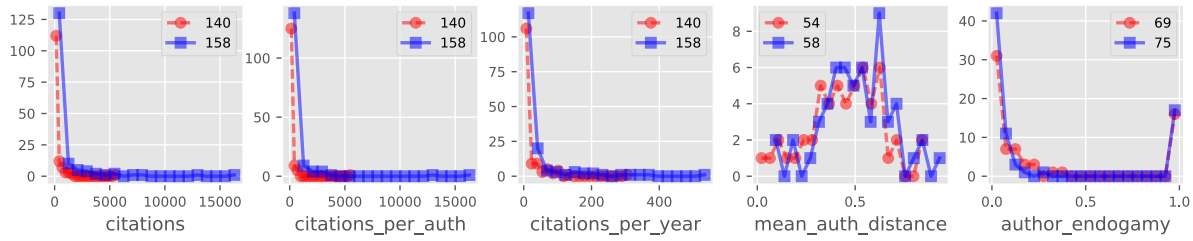

Figure 2: Histograms of the five numerical features.

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.

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

Table 2: Results of t- and χ^2 tests.

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 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.

Figure 2 shows 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, which is visible in Figure 3. Figure 3 shows the relation between author distance and research endogamy, represented as the number of publications belonging to each collaboration category introduced in Figure 1. To create this figure, we have first assigned each publication two values – its author distance and research endogamy. We have then used median endogamy (0.0297) and median author distance (0.4996) to separate the publications in the dataset into the four categories presented in Figure 1.

The figure 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-

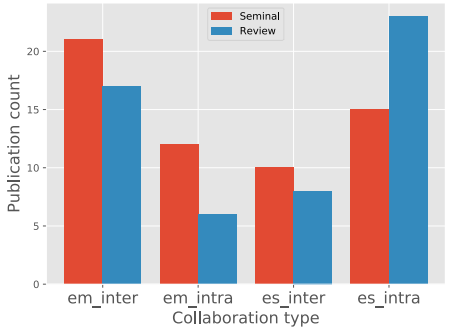

Figure 3: Number of publications belonging to each collaboration category across both publication types.

community citation and collaboration patterns are characteristic for high impact scientific production (Guimerà et al., 2005; Lambiotte and Panzarasa, 2009; Montolio et al., 2013). 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 (Montolio et al., 2013)). 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

Brandão, M. A., Vaz de Melo, P., and Moro, M. M. (2017). Tie strength persistence and transformation. *AMW (to appear)*.

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87(3):597–620.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498–514.

Friedkin, N. (1980). A test of structural features of granovetter's strength of weak ties theory. *Social networks*, 2(4):411–422.

Fu, T. Z., Song, Q., and Chiu, D. M. (2014). The academic social network. *Scientometrics*, 101(1):203–239.

Glenisson, P., Glänzel, W., Janssens, F., and De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6):1548–1572.

Granovetter, M. S. (1973). The strength of weak ties. In *American Journal of Sociology*, volume 78, pages 1360–1380. Elsevier.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, pages 201–233.

Guimerà, R., Uzzi, B., Spiro, J., and Nunes Amaral, L. A. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(April):697–702.

Herrmannova, D. and Knoth, P. (2015). Semantometrics in coauthorship networks: Fulltext-based approach for analysing patterns of research collaboration. *D-Lib Magazine*, 21(11/12).

Herrmannova, D., Patton, R. M., Knoth, P., and Stahl, C. G. (2017). 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*.

Janssens, F., Leta, J., Glänzel, W., and De Moor, B. (2006). Towards mapping library and information science. *Information processing & management*, 42(6):1614–1642.

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

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

Lin, W., Kong, X., Yu, P. S., Wu, Q., Jia, Y., and Li, C. (2012). Community detection in incomplete information networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 341–350. ACM.

Montolio, S. L., Dominguez-Sal, D., and Larriba-Pey, J. L. (2013). Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, 42(2):11–16.

Newman, M. E. (2004). Who is the best connected scientist? a study of scientific coauthorship networks. In *Complex networks*, pages 337–370. Springer.

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

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira Jr., W., and Laender, A. H. F. (2014). Community-based Endogamy as an Influence Indicator. In *Digital Libraries 2014 Proceedings*, page 10.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-j. P., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM.

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

7. Language Resource References

Herrmannova et al. (2017). *TrueImpactDataset*. Distributed via `http://trueimpactdataset.semantometrics.org/`, ISLRN 197-407-228-291-9.