Path-Combination-Based Co-authorship Prediction in Document-Author Bipartite Networks (Postprint)

Authors: Zhang Jinzhu, Wang Xiaomei, Han Tao

Date: 2017-11-08T00:00:00+00:00

Abstract

【Purpose】 To mitigate the impact of information loss during the projection of paper-author bipartite networks onto co-authorship networks, develop co-authorship relationship prediction metrics and methods tailored to specific bipartite networks, and enhance prediction accuracy and result interpretability.

【Method】 First, construct the paper-author bipartite network and its projected co-authorship network; then extract second-order and third-order paths from the bipartite network to represent association relationships between authors; finally, employ logistic regression to learn the contribution of different paths to co-authorship relationship prediction, thereby establishing a path-combination-based co-authorship relationship prediction metric within paper-author bipartite networks.

【Results】Experiments in the field of library and information science confirm that substantial information loss occurs during the projection of paper-author bipartite networks onto co-authorship networks, which is quantitatively measured through variations in co-authorship relationship prediction accuracy; logistic regression is well-suited for learning the contribution of different paths to co-authorship relationship prediction, and the resulting path-combination metric achieves significantly higher accuracy compared to other metrics, with more interpretable prediction results.

【Limitations】 Other higher-order paths have not yet been incorporated into this model, and the generalizability of the method requires validation in other domains.

【Conclusion】 Co-authorship relationship prediction should be conducted directly on paper-author bipartite networks to reduce the impact of information loss during projection; the path-combination metric on paper-author bipartite networks is the optimal metric for co-authorship relationship prediction; this method can

be extended and applied to other types of bipartite networks, such as patent-inventor bipartite networks.

Full Text

Predicting Co-authorship with Combination of Paths in Paper-Author Bipartite Networks

Zhang Jinzhu[1], Wang Xiaomei[2], Han Tao[2]
[1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190, China

Abstract

[**Objective**] This paper aims to reduce information loss during the projection of paper-author bipartite networks into co-authorship networks, develop prediction indicators and methods adapted to specific bipartite networks, and improve prediction accuracy and result interpretability.
[**Methods**] First, we constructed a paper-author bipartite network and its projected co-authorship network. Next, we extracted second-order and third-order paths from the bipartite network to represent author associations. Finally, we employed logistic regression to learn the contribution of different paths to co-authorship prediction, forming a path-combination-based indicator for the bipartite network.
[**Results**] Experiments in library and information science confirmed substantial information loss during bipartite network projection, quantified through accuracy changes in co-authorship prediction. Logistic regression proved suitable for learning path contributions, yielding a path-combination indicator with far superior accuracy and interpretability.
[**Limitations**] Higher-order paths were not incorporated into the model, and the method' s generalizability requires validation in other domains.
[**Conclusions**] Co-authorship prediction should be conducted directly on paper-author bipartite networks to reduce projection-related information loss. The path-combination indicator in bipartite networks is optimal for co-authorship prediction and can be extended to other bipartite networks such as patent-inventor networks.

Keywords: Paper-author bipartite network; Paths combination indicator; Library and Information Science; Co-authorship network; Co-authorship prediction

1. Introduction

Against the backdrop of increasing interdisciplinary integration and the growing specialization of research directions, scientific research is increasingly shifting from individual work to collaborative team efforts, thereby enhancing research quality and efficiency. This trend has drawn significant attention to the study of research team formation tailored to contemporary needs and specific topics. Co-authorship relationships serve as a crucial manifestation of scientific collaboration and an important pathway for discovering such partnerships [1]. Consequently, the likelihood of co-authorship can, to some extent, represent the potential for scientific cooperation among authors, providing valuable insights for team member selection and composition [2].

Currently, co-authorship prediction is primarily conducted within co-authorship networks, which represent authors as nodes and co-authorship relationships as edges. Since both nodes and edges are of a single type, these networks constitute a form of unipartite network. Co-authorship prediction in such networks involves forecasting the likelihood of edge formation between currently unconnected node pairs [3]. By applying and refining various correlation metrics from complex network analysis, we can calculate the degree of relatedness between author pairs that have not yet co-authored, using this relatedness to indicate their future co-authorship potential [4]. These correlation metrics for author pairs can be categorized into common neighbor and its variants, path-based indicators, and random walk indicators [5], with numerous experiments across various domains comparing their effectiveness to identify optimal predictors [6-7].

Co-authorship networks are derived from paper-author bipartite networks through projection. However, this projection process results in the loss of paper-specific information, making it difficult to track which specific publications facilitate co-authorship relationships and potentially reducing prediction accuracy [8-9]. Therefore, it is necessary to quantify this information loss and its impact on co-authorship prediction by comparing the accuracy changes of identical indicators between bipartite networks and their projected co-authorship counterparts. This challenge has motivated a new approach: conducting co-authorship prediction directly on bipartite networks. Bipartite networks consist of two distinct node types with edges existing only between nodes of different types, giving rise to specialized centrality metrics, clustering coefficients, community structures, and evolutionary models [10]. Link prediction in bipartite networks primarily involves adapting unipartite network indicators to the bipartite context, yielding bipartite versions of common neighbor and local path metrics [11] that have demonstrated promising results in domains such as product-consumer, RNA-protein, and book-borrower networks. Nevertheless, the relationships among authors in paper-author bipartite networks are more diverse and complex than those in co-authorship networks. Further research is needed to effectively extract and represent these varied associations, clarify their contributions to co-authorship prediction, and develop optimal indicators that integrate multiple relationship types.

This paper directly extracts multiple path types from paper-author bipartite networks to represent author associations, employing logistic regression—a machine learning method—to learn the contribution of each path type to co-authorship prediction. The learned weight coefficients are then used to combine these paths into a multi-path combination indicator for co-authorship prediction in bipartite networks. Building upon this foundation, we compare and analyze relevant prediction indicators across both paper-author bipartite networks and their projected co-authorship networks, quantifying information loss during projection through accuracy variations.

---

2. Co-authorship Prediction in Paper-Author Bipartite Networks

The co-authorship prediction model in paper-author bipartite networks comprises three components: (1) construction of bipartite networks and their projected co-authorship networks, (2) path-based representation and combination of author associations in bipartite networks, and (3) evaluation of co-authorship prediction indicators. First, we design a projection scheme from bipartite to co-authorship networks that ensures high consistency in co-authorship prediction, enabling fair comparisons. Next, we extract multiple path types from the bipartite network to represent author associations as driving factors for co-authorship formation, using machine learning methods to construct multi-path combination indicators. Finally, we evaluate these prediction indicators using link prediction methodologies.

2.1 Network Construction and Data Splitting

[Figure 1: see original paper]

We construct paper-author bipartite networks and their corresponding projected networks using the following method, creating training and test sets for model training and evaluation. First, we extract all co-authorship relationships from the bipartite network, representing them as "author-paper-author"triples. For instance, in the bipartite network shown in Figure 1(a), all co-authorship relationships are represented as: (A1, A3): [A1P1A3, A1P2A3], (A1, A2): [A1P2A2], (A2, A3): [A2P2A3], and (A2, A4): [A2P3A4]. Next, we generate training and test sets using 10-fold cross-validation [7,12], dividing the dataset into ten equal groups through random sampling without replacement. Each group serves once as the test set while the remaining nine groups form the training set, yielding ten distinct training-test set combinations. Finally, we construct corresponding training and test sets for the bipartite network using the "author-paper-author" relationships. For example, if (A1, A3), (A1, A2), and (A2, A3) constitute the training set and (A2, A4) the test set in the co-authorship network, then the bipartite network training set comprises [A1P1A3, A1P2A3], [A1P2A2], and [A2P2A3], with [A2P3A4] as the test set.

This data splitting approach ensures high consistency between bipartite and projected networks when comparing co-authorship prediction indicators. However, because the projected network does not retain paper information, certain inconsistencies persist between the two network types—precisely verifying the existence of information loss during projection and causing identical indicators to yield different results across network types. For example, if [(A1, A3), (A2, A3), (A2, A4)] is selected as the training set and [(A1, A2)] as the test set, the corresponding "author-paper-author" relationships in the bipartite network training set would be [A1P1A3, A1P2A3, A2P2A3, A2P3A4]. Notably, the relationships [A1P2A3, A2P2A3] in this training set would directly imply the formation of A1P2A2, requiring no prediction whatsoever.

2.2 Construction of Multi-Path Combination Indicators Based on Logistic Regression

Information loss occurs when bipartite networks are projected into co-authorship networks [8-9]. After constructing the bipartite network, we must extract multiple reachability paths among authors to represent their associations, employing logistic regression—a machine learning method—to learn the influence and contribution of different paths to co-authorship prediction. This process yields a path-combination-based co-authorship prediction indicator for paper-author bipartite networks.

(1) Path Representation and Extraction in Bipartite Networks In paper-author bipartite networks, author associations are formed through papers. For instance, direct co-authorship can be represented as "author-paper-author" (APA). The number of common neighbors between two authors can be represented by "author-paper-author-paper-author" (APAPA) paths. Considering only authors, the reachability path length between two authors is 2, corresponding to common neighbors in co-authorship networks. We therefore term this association a second-order path. Co-authorship relationships formed between collaborators of two authors can be represented by "author-paper-author-paper-author-paper-author" (APAPAPA) paths, with a reachability path length of 3 between the two authors, which we designate as third-order paths. In Figure 1, A1P1A3 indicates that authors A1 and A3 have a co-authorship relationship; A1P2A2P3A4 shows that author A2 is a common neighbor of A1 and A4, and since only one such path exists between A1 and A4, their second-order path count is 1. A3P1A1P2A2P3A4 represents that A1 (a collaborator of A3) and A2 (a collaborator of A4) have a co-authorship relationship; with only one such path between A3 and A4, their third-order path count is 1.

This paper employs second-order and third-order paths from paper-author bipartite networks to represent author associations, though these can be extended to fourth-order and higher-order paths.

(2) Multi-Path Combination Method Based on Logistic Regression
Multiple path types in bipartite networks may influence co-authorship prediction, with each path likely contributing differently. Therefore, machine learning methods are required to learn weight coefficients for each path type from training data, representing their respective contributions to co-authorship prediction, thereby forming a multi-path combination indicator.

Logistic Regression is a classification model in machine learning widely applied in data mining, automated disease diagnosis, and economic forecasting [13]. It addresses binary classification problems by calculating the probability of data belonging to a specific class based on one or more independent variables (i.e., influencing factors for binary classification). Using weight coefficients learned from training data, logistic regression can predict classifications for new data and compute their probabilities of belonging to specific classes.

When applied to constructing multi-path combination indicators, the independent variables are the counts of second-order and third-order paths, while the dependent variable indicates whether a co-authorship relationship exists (1 if it occurs, 0 otherwise). For each author pair (i, j) in the training set, Xk is a two-dimensional vector storing the counts of second-order and third-order paths, and yk indicates co-authorship occurrence. For example, if authors i and j have 2 second-order paths and 6 third-order paths with an existing co-authorship relationship, then Xk = [2, 12] and yk = 1. The logistic regression method uses the training set from 10-fold cross-validation as positive examples and randomly samples an equal number of negative examples to form the complete training set. We implement logistic regression using Python' s scikit-learn machine learning toolkit (specifically, the 'sklearn.linear_{model}.LogisticRegression' class). By inputting multiple Xk and yk pairs from the training set, we obtain weights for second-order and third-order paths, forming the multi-path combination indicator to calculate co-authorship probabilities for author pairs without existing collaborations.

2.3 Evaluation Based on Link Prediction

Link prediction is frequently employed to quantitatively evaluate the performance of correlation indicators in complex networks [7]. Since both paper-author bipartite networks and their projected counterparts are complex networks, and co-authorship prediction indicators constitute a type of correlation metric, link prediction theory and methods can be applied to assess various prediction indicators. Let G = (V, E) denote a paper-author bipartite network, where V represents the set of authors and E represents the set of co-authorship relationships expressed as "author-paper-author" triples. The complete set of possible co-authorship relationships U in this network contains $N\times(N-1)/2$ pairs. Given a co-authorship prediction indicator, we calculate the co-authorship likelihood for author pairs (x, y) that have not yet collaborated (U -E), sorting them in descending order of predicted probability, with top-ranked pairs having the highest future co-authorship potential.

To evaluate co-authorship prediction indicators, we partition the co-authorship relationship set E into ten training sets ET and test sets EP through 10-fold cross-validation. We compute co-authorship likelihoods for author pairs in the training set using the prediction indicator and assess result accuracy on the test set. Link prediction primarily employs two accuracy metrics: AUC (Area Under ROC Curve) and Precision [7], with reported values representing averages across ten runs. These metrics emphasize different aspects of accuracy. AUC provides a holistic measure of indicator precision but exhibits low discriminative power, as multiple indicators may show similar AUC values, potentially yielding high AUC scores even for poorly performing predictors. Precision, conversely, measures whether top-L ranked co-authorship predictions are accurate, with L being adjustable. This paper uses R-Precision to evaluate co-authorship prediction accuracy, considering both correctness and ranking quality, where L equals the number of co-authorship relationships in the test set.

---

3. Experiments and Results

To validate our approach, we constructed paper-author bipartite networks and corresponding co-authorship networks in library and information science, applying co-authorship prediction indicators to both networks. By comparing prediction accuracy and AUC values, we quantified information loss during projection, calculated the contributions of different paths in bipartite networks, and verified the effectiveness of our path-combination indicator and methodology.

3.1 Data Description

We downloaded data from the Web of Science (WoS) for publications indexed in the Science Citation Index Expanded (SCIE) under the subject category Information Science & Library Science, covering the period from 2005 to 2009. We excluded data from *Scientist* journal because it contains numerous short papers and belongs to multiple subject categories, which would otherwise skew results by making frequent authors predominantly those publishing in this journal, thereby reducing experimental credibility. The search query for our dataset was:

(WC = Information Science & Library Science) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article) Indexes=SCI-EXPANDED Timespan=2005-2009 Refined by: [excluding] SOURCE TITLES: (Scientist)

Data preprocessing primarily involved removing anonymous author information (authors named "[anonymous]"). We then selected authors with publication frequencies greater than or equal to 3 and their corresponding papers to construct the paper-author bipartite network and its projected co-authorship network. Table 1 provides detailed data statistics. The number of isolated authors refers to high-frequency authors who did not engage in any collaborations. The train-

ing set contains 90% of all co-authorship relationships, with the remaining 10% allocated to the test set.

3.2 Information Loss in Projecting Paper-Author Bipartite Networks

Significant information loss occurs when projecting paper-author bipartite networks into co-authorship networks. The CN (Common Neighbor) and second-order path indicators represent common neighbor counts in co-authorship networks and their bipartite counterparts, respectively. In co-authorship networks, the CN indicator achieves 27.1% accuracy and 85.4% AUC, whereas the second-order path indicator in bipartite networks attains 48.3% accuracy and 85.5% AUC. Notably, the bipartite network accuracy is 21.2 percentage points (78.2%) higher than that of the co-authorship network, quantitatively demonstrating the information loss during projection. Meanwhile, AUC, as a macro-level evaluation metric, shows minimal change and poor discriminative power. These results indicate that reduced co-authorship prediction accuracy is closely related to information loss during bipartite network projection, and that the magnitude of information loss can be quantified through accuracy variations, offering a novel approach for quantifying projection-related information loss.

The inability of projected co-authorship networks to retain paper information not only reduces prediction accuracy but also complicates result interpretation. For instance, among the top 10 successfully predicted co-authorship pairs in the first experimental run, both CN and second-order paths correctly predicted that Bates DW and Jenter CA would co-author, with CN count of 6 and second-order path (APAPA) count of 14 (all representing common neighbors). Conversely, among the top 119 predicted pairs (matching the test set size), second-order paths successfully predicted co-authorship between Markpin T and Sombatsompop N, which CN failed to predict, showing CN count of 2 and second-order path (APAPA) count of 13. These examples demonstrate that some common neighbor relationships are lost during projection, ultimately reducing prediction accuracy in projected networks. In contrast, paper-author bipartite networks facilitate tracking of which specific papers enable co-authorship, revealing the underlying reasons and motivations for collaboration and providing better interpretability for predictions.

In summary, substantial information loss during the projection from paper-author bipartite networks to co-authorship networks significantly reduces prediction accuracy. Therefore, co-authorship prediction should be conducted directly on paper-author bipartite networks to mitigate projection-related information loss and enhance result interpretability.

3.3 Comparative Analysis of Path Combination and Other Indicators

We selected three path-based indicators from bipartite networks for comparative analysis: (1) the second-order path indicator (count of second-order paths) corresponding to Common Neighbor; (2) the path combination indicator corre-

sponding to Local Path, representing combinations of second-order and third-order paths—specifically, Local Path Indicator 1 fixes third-order path weight at 0.1, while Local Path Indicator 2 fixes it at 0.01; and (3) the third-order path indicator representing third-order path counts. For a comprehensive and fair comparison, we also selected representative indicators from co-authorship networks: Common Neighbor and Resource Allocation as representatives of common neighbor and its variants, Local Path and Full Path as representatives of path combination indicators, and SimRank as a representative of random walk indicators—all of which have demonstrated excellent performance in their respective categories [7]. Comparing indicators across bipartite networks reveals the varying contributions of different paths and identifies optimal indicators for co-authorship prediction. Comparing bipartite and co-authorship network indicators highlights the significant impact of path weighting and identifies the most influential factors in co-authorship prediction. Tables 2 and 3 present the accuracy and AUC values for these indicators in paper-author bipartite networks and their co-authorship counterparts.

Table 2. Accuracy and AUC Values of Co-authorship Prediction Indicators in Co-authorship Networks

Indicator	Accuracy
Common Neighbor	12.9
Resource Allocation	25.5
Local Path Indicator 1	25.5
Local Path Indicator 2	20.8
Full Path	27.1
SimRank	30.2

Table 3. Accuracy and AUC Values of Co-authorship Prediction Indicators in Paper-Author Bipartite Networks

Indicator	Accuracy
Second-order Path	28.6
Third-order Path	59.1
Path Combination (Second-order + Third-order)	48.3

The optimal indicator for co-authorship prediction is the path combination indicator that integrates both second-order and third-order path information, demonstrating that paths of different lengths all influence prediction outcomes. Across both paper-author bipartite and co-authorship networks, the path combination indicator achieves the highest accuracy and AUC values. Specifically, its accuracy is 10.8 percentage points (22.4%) higher than the second-order path indicator, 30.5 percentage points (63.1%) higher than the third-order path indicator, 28.9 percentage points (95.7%) higher than the best-performing Resource

Allocation indicator in co-authorship networks, and 46.2 percentage points (3.58 times) higher than the worst-performing SimRank indicator. AUC values show minimal variation, serving as a macro-level evaluation that inadequately distinguishes indicator quality for co-authorship prediction.

The path combination indicator' s status as the optimal predictor stems from its use of machine learning to learn path-specific contributions tailored to the dataset, confirming the crucial role of weighting. Comparisons with multiple co-authorship network indicators further demonstrate that the influence of different path lengths is not static and requires dataset-specific learning and adjustment. In both network types, both the path combination and local path indicators consider second-order and third-order paths, but differ in that the path combination indicator learns path contributions from the specific dataset, whereas local path indicators use fixed empirical values (typically 0.1 or 0.01) for third-order path weights. Accuracy results from Tables 2 and 3 show that the path combination indicator surpasses Local Path Indicator 1 by 38.3 percentage points (1.84 times), Local Path Indicator 2 by 33.6 percentage points (1.32 times), and the Full Path indicator by 33.6 percentage points (1.32 times). These findings confirm that different paths contribute differently to co-authorship prediction and require dataset-specific adjustment to achieve optimal results. The universal weight settings employed by local path and full path indicators are unsuitable for library and information science co-authorship prediction and necessitate relearning and adjustment.

Weight analysis across different paths in bipartite networks confirms that second-order paths are more important than third-order paths, with weight values varying across datasets. As shown in Table 4 , the weights for second-order and third-order paths differ across the ten bipartite networks constructed through cross-validation, indicating that their contributions to co-authorship prediction require dataset-specific learning rather than universal optimal values. Additionally, Table 4 reveals that second-order path weight coefficients are substantially higher than those for third-order paths, confirming that second-order paths contribute far more significantly to prediction. The path combination indicator' s modest 10.8 percentage point (22.4%) improvement over the second-order path indicator further validates that common neighbors remain the most influential factor in co-authorship formation.

Prediction indicators from co-authorship networks similarly confirm that common neighbors are the most influential factor in co-authorship prediction. In Table 2, both local path and full path indicators show lower accuracy than the common neighbor indicator, suggesting that third-order or higher-order paths contribute limited value to prediction and may even negatively impact performance when using fixed empirical weights. Notably, the Resource Allocation indicator—a direct improvement over common neighbor that distinguishes author influence using neighbor degrees—achieves 3.1% higher accuracy than common neighbor, indirectly confirming the significant impact of second-order paths. Consequently, numerous studies continue to refine common neighbor indicators

from various perspectives, yielding promising results [14-15].

3.4 Co-authorship Prediction Examples

We selected the best prediction indicator from co-authorship networks (Resource Allocation) and the two best indicators from bipartite networks (second-order path and path combination) to illustrate co-authorship prediction examples, listing the top 10 predicted author pairs in Table 5 . Bold italic text indicates successfully predicted co-authorship relationships, while non-emphasized text denotes failed predictions. Since the experiments employ 10-fold cross-validation, we present results from the first run only, with indicator accuracy displayed alongside (e.g., "Resource Allocation (31.9%)" indicates its accuracy in the first run).

Table 5. Top 10 Co-authorship Predictions by Three Indicators

Rank	Resource Allocation (31.9%)	Second-order Path (49.2%)	Path Combination (60.8%)
1	**Detmer DE**	**Wang Y**	**Steen EB**
2	**Zhang L**	**Van Leeuwen TN**	**Costas R**
3	**Huntington P**	**Nicholas D**	**Jamali HR**
4	**Teo HH**	**Wei KK**	**Huntington P**
5	**Nicholas D**	**Rowlands I**	**Bates DW**
6	**Jamali HR**	**Rowlands I**	**Rowlands I**
7	**Williams P**	**Jenter CA**	**Zubair M**
8	**Jayakanth F**	**Poon EG**	**Jenter CA**
9	**Li JX**	**Zhang Z**	**Chen YC**
10	**Hwang SJ**	**Sia CL**	**Benbasat I**

Note: Bold indicates successful predictions. The table shows partial examples; full table would include all predicted pairs.

Table 5 demonstrates that all three indicators perform reasonably well, with Resource Allocation successfully predicting 8 out of 10 co-authorship relationships, while both second-order path and path combination indicators correctly predict all 10 top-ranked pairs, confirming the path combination indicator as the optimal predictor. Furthermore, among the top 20 and 30 predictions, Resource Allocation successfully predicts 10 and 14 pairs, respectively; second-order paths predict 19 and 27 pairs; and the path combination indicator predicts 19 and 29 pairs. These results reaffirm the negative impact of information loss during network projection on prediction accuracy and demonstrate that path-based indicators in bipartite networks enable superior co-authorship prediction.

4. Conclusion and Future Work

Information loss occurs when projecting paper-author bipartite networks into co-authorship networks, necessitating the development of prediction indicators and methods tailored specifically for bipartite networks to better analyze and reveal the mechanisms underlying co-authorship formation. This paper proposes a path-combination-based co-authorship prediction indicator and methodology for paper-author bipartite networks to improve both prediction accuracy and interpretability. Experiments in library and information science confirm that the second-order path indicator in bipartite networks significantly outperforms the common neighbor indicator in co-authorship networks, quantifying information loss during projection through accuracy differences. This demonstrates that co-authorship prediction should be conducted directly on paper-author bipartite networks to enhance accuracy and interpretability. Furthermore, the path combination indicator, which integrates second-order and third-order path information, substantially outperforms other indicators, showing that different paths contribute to prediction but require dataset-specific learning rather than universal empirical values. Additionally, second-order paths contribute significantly more than third-order paths, confirming common neighbors as the most influential factor in co-authorship prediction.

While experiments in library and information science validate the effectiveness of path combination indicators, several issues warrant further investigation. First, the contributions of fourth-order and higher-order paths to co-authorship prediction require clarification, such as extracting all path lengths from bipartite networks and using logistic regression to learn weight coefficients for each path to form a full-path combination indicator for accuracy comparison. Second, the logistic regression-based path combination method requires testing in other domains to validate its generalizability, with alternative machine learning methods also meriting comparison. Finally, this approach can be extended to other bipartite network types, including inventor collaboration prediction in patent-inventor networks, user recommendation in microblog-user networks, and product recommendation in user-product networks.

References

[1] Barabasi A L, Jeong H, Neda Z, et al. Evolution of the Social Network of Scientific Collaborations [J]. Physica A: Statistical Mechanics and Its Applications, 2002, 311(3): 590-614.
[2] Guns R, Rousseau R. Recommending Research Collaborations Using Link Prediction and Random Forest Classifiers [J]. Scientometrics, 2014, 101(2): 1461-1473.
[3] Zhang Q, Xu X, Zhu Y, et al. Measuring Multiple Evolution Mechanisms of Complex Networks [J]. Scientific Reports, 2015, 5: Article No. 10350.
[4] Zhang B, Ma F. A Review on Link Prediction of Scientific Knowledge Network [J]. Journal of Library Science in China, 2015, 41(3): 99-113.
[5] Zhang J, Han T, Wang X. Uncovering the Mechanism of Knowledge Network

Evolution by Link Prediction [J]. Geomatics and Information Science of Wuhan University, 2015, 39(S1): 100-106.
[6] Zhao J, Miao L, Yang J, et al. Prediction of Links and Weights in Networks by Reliable Routes [J]. Scientific Reports, 2015, 5: Article No. 12261.
[7] Lv L, Zhou T. Link Prediction in Complex Networks: A Survey [J]. Physica A: Statistical Mechanics and Its Applications, 2010, 390(6): 1150-1170.
[8] Guns R. Bipartite Networks for Link Prediction: Can They Improve Prediction Performance?[C]. In: Proceedings of International Society for Scientometrics and Informetrics. 2011: 249-260.
[9] Gao M, Chen L, Xu Y. Projection Based Algorithm for Link Prediction in Bipartite Network[J]. Computer Science, 2016, 43(2): 118.
[10] Wu Y, Zhang P, Di Z, et al. Study on Bipartite Networks[J]. Complex Systems and Complexity Science, 2010, 7(1): 1-12.
[11] Daminelli S, Thomas J M, Duran C, et al. Common Neighbours and the Local-Community-Paradigm for Topological Link Prediction in Bipartite Networks[J]. New Journal of Physics, 2015, 17: 113037.
[12] Zhou T, Lv L, Zhang Y C. Predicting Missing Links via Local Information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.
[13] Hosmer Jr D W, Lemeshow S. Applied Logistic Regression [M]. New York: John Wiley & Sons, 2004.
[14] Güneş İ, Gündüz-Öğüdücü Ş, Çataltepe Z. Link Prediction Using Time Series of Neighborhood-Based Node Similarity Scores [J]. Data Mining and Knowledge Discovery, 2016, 30(1): 147-180.
[15] Sett N, Singh S R, Nandi S. Influence of Edge Weight on Node Proximity Based Link Prediction Methods: An Empirical Analysis[J]. Neurocomputing, 2016, 172: 71-83.

Author Contributions

Zhang Jinzhu, Wang Xiaomei, Han Tao: Conceptualized the research, designed the study, and revised the final manuscript.
Zhang Jinzhu: Collected, cleaned, and analyzed data; conducted experiments; drafted the manuscript.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data are stored by the authors and available upon request at zhangjinzhu@njust.edu.cn.

Received Dates

Received: 2016-06-15; Revised: 2016-08-01

Note: Figure translations are in progress. See original paper for figures.