Co-author Relationship Prediction in Bibliographic Network: A New Approach Using Geographic Factor and Latent Topic Information

Thi Kim Thoa Ho
CHArt Laboratory EA 4004, EPHE,
PSL Research University, Paris, France
University of Education, Hue
University, Vietnam
thi-kim-thoa.ho@etu.ephe.psl.eu

Quang Vu Bui
University of Sciences, Hue
University, Vietnam
Hue, Vietnam
buiquangvu@hueuni.edu.vn

Marc Bui
CHArt Laboratory EA 4004, EPHE,
PSL Research University, Paris, France
marc.bui@ephe.psl.eu

ABSTRACT

In this research, we propose a novel approach for co-author relationship prediction in a bibliographic network utilizing geographic factor and latent topic information. We utilize a supervised method to predict the co-author relationship formation where combining dissimilar features with the dissimilar measuring coefficient. Firstly, besides existing relations have been studied in previous researches, we exploit new relation related to the geographic factor which contributes as a topological feature. Moreover, we discover content feature based on textual information from author's papers using topic modeling. Finally, we amalgamate topological features and content feature in co-author relationship prediction. We conducted experiments on dissimilar datasets of the bibliographic network and have attained satisfactory results.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Link prediction; bibliographic network; multi-relation network; topic modeling.

ACM Reference Format:
Thi Kim Thoa Ho, Quang Vu Bui, and Marc Bui. 2019. Co-author Relationship Prediction in Bibliographic Network: A New Approach Using Geographic Factor and Latent Topic Information. In *The Tenth International Symposium on Information and Communication Technology (SoICT 2019), December 4–6, 2019, Hanoi - Ha Long Bay, Viet Nam.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3368926.3369668

SoICT 2019, December 4–6, 2019, Hanoi - Ha Long Bay, Viet Nam

ACM ISBN 978-1-4503-7245-9/19/12…$15.00
https://doi.org/10.1145/3368926.3369668

1 INTRODUCTION

1.1 Problem Definition

Link prediction in networks has been a key research branch since the emergence of online social networks. Link prediction is a significant task in link mining. Link prediction is to predict whether there will be links between two nodes based on the attribute information and the observed existing link information. Link prediction not only can be used in the field of the social network but can also be applied in other fields [14]. There are many applications of link prediction in the field of social networks, for instance, these methods can be applied for link recommendations to users in online social networks or the evaluation of evolving social network models (Lu & Zhou 2011) and so on. Moreover, link prediction can be applied to other networks such as predicting outbreak of a disease in disease networks, detecting spam emails in email networks or suggesting alternative routes for possible navigation based on the current traffic patterns, etc.

Majority existing link prediction researches [1, 4, 6–8, 17] have been studied on homogeneous networks where only one type of object and one type of link exists in the network, for instance, co-author network with object *author* and *co-author* link or object *user* and link *friendship* in friendship network. Nevertheless, majority networks in real-world are heterogeneous where there are various object's types and multiple relations in network, for instance, bibliographic network is heterogeneous network where contains multiple objects including authors, papers, venues, affiliation and so on, concurrently exist numerous relationships among authors such as co-author relation, relation with common co-author and so on. Link prediction on heterogeneous network has been studied in [11, 15, 16], however; those studies just exploited topological features. As a matter of fact, beside topological features, the content from papers contains important information which contributes to the formation of co-author relationship since two authors research in the same field has a high probability to connect compare to the dissimilar field. Therefore, in this study, we will study the problem of predicting the co-author relationship in heterogeneous bibliographic network with exploiting on both topological and content feature.

In this study, we propose a new approach that amalgamates topological features and content feature in co-author relationship prediction on the bibliographic network. Previous studies [11, 15, 16]

considered relations such as co-author relation, relation with common co-author, the relation is to publish the paper in the same venue and so on. However, there is no study take into account relation related to geographic factor, for instance, two authors work the same laboratory, same research institution, same city, country or geographic distance from their workplaces and so on. In fact, the geographic factor can create an important opportunity in co-author relationship establishment since there is a high probability for two researchers to corporate together if they work in the same laboratory or they can have the opportunity to meet and discuss if they work in the same country and so on. Therefore, besides existing topological features, we exploit a new topological feature related to geographic factor. On the other hand, we discover content feature based on textual information from the author's papers. The textual information can be extracted from the title, abstract, keywords or whole text of the paper. Textual information had been used for link prediction on the homogeneous network, but with methods is to count number common keywords [6] or utilize Term Frequency – Inverse Document Frequency (TFIDF) feature vector representation and the cosine measure to compute similarity from the title of papers [17]. In this study, we apply topic modeling to estimate the topic's probability distribution of authors. After that, we will utilize distance measure related to the probability distribution to measure interest's similarity of authors. Finally, we utilize a supervised learning framework to learn the best weights associated with topological features and content feature. Experimental results demonstrate that feature related to geographic factor contribute to enhance the performance of co-author relationship prediction. Moreover, the content feature is estimated using topic modeling can improve accuracy in co-author relation prediction. Particularly, the combination of topological features and content feature with our proposed method can obtain the highest accuracy. Our research has the following contributions:

- We exploit geographic factor as a topological feature which contributes to the performance enhancement of co-author relationship prediction.
- We discover a content feature in co-author relation prediction based on textual information from author's papers using topic modeling technology; from that estimate similarity of author's research interest.
- We propose a new method that is to combine topological features and content feature in co-author relation prediction on the bibliographic heterogeneous network where there is a contribution of new topological feature related to geographic factor and the application of topic modeling in content feature extraction. We conducted experiments and obtained satisfying results.

The structure of our paper is organized as follows: section 1 introduce problem definition and related works; section 2 reviews preliminaries; our approach are proposed in section 3; section 4 illustrates experiments and results; we conclude our work in section 5.

1.2 Related Works

Link prediction is one of the core tasks of social network analysis. On one hand, the link prediction on social networks has been researched broadly on homogeneous networks where there is just only one type of object that is considered such as author on co-author network or user on friendship network and only one type of link including co-author or friendship relationship respectively. The first works mostly studied using unsupervised methods [1, 7] in which dissimilar similarity measures had been proposed to estimate the similarity of each pair of node x and y. All non-observed links are ranked according to their similarity scores and the links are supposed to be higher connection probability if there is a higher likelihood. Subsequently, supervised methods were proposed for link prediction [4, 6, 8, 17] where we can combine different features with different measuring coefficient. In [4, 8], a spectrum of common topological features was utilized for link prediction including common neighbors, Jaccard coefficient, Adamic/Adar, and so on. Moreover, in [6, 17], authors exploited both topological features and content feature. Topological features consist common neighbors, shortest distance, clustering index, the shortest distance in author-KW graph, etc. Besides, the content feature was exploited based on the number of common keywords [6] or title of papers, after that TFIDF feature vector representation and the cosine measure to compute similarity [17]. Several survey on link prediction can be found in [5, 9, 18].

On the other hand, the link prediction problem also was studied on heterogeneous network [11, 15, 16] where there are multiple object types and relationships in network, for instance, bibliographic network is heterogeneous network with various types of objects such as authors, papers, venues, affiliation of authors and so on; there are numerous relationships between authors including co-author relation, relationship is to publish papers with the same venues, and so on. Link prediction in relational data which involves different types of objects and complex relationships between objects in [11, 16]. Supervised methods were utilized in [11, 15] while probabilistic model in [16]. However, those studies just explored topological features from relationships such as publish the same venues, citation relation, and so on. Therefore, in this study, we extend the link prediction problem on the heterogeneous network-bibliographic network by combining topological features and content feature. Supervised methods will be utilized for link prediction. In addition to the existing relationships in previous studies, we exploit a new relation as a topological feature related to geographic factor. Moreover, we expand to discover content feature based on textual information from papers where text information can be extracted from keywords, titles, abstracts or whole papers. Concurrently, we apply topic modeling to estimate textual information similarity instead of two existing methods that are number common keywords and TFIDF.

2 PRIMINALARIES

2.1 Link Prediction in co-authorship networks

The link prediction task is to predict whether two authors will build co-author relationship in a future time when currently they haven't co-authored to each other. There are several frameworks for link prediction have been studied including similarity-based

algorithms, supervised learning framework or probabilistic model that we summarize at subsection 1.2. In this study, we will concentrate on supervised learning framework for link prediction. This framework takes into account link prediction as a simple binary classification problem: for any two potentially linked objects a_i and a_j, predict whether $l_{a_i a_j}$ is 1 or 0.

Generally, given a past time interval $T_1 = [t1, t2]$, we will utilize features extracted from the aggregated network in time period T_1 to predict the relationship formation in a future time interval $T_2 = [t2, t3]$. In training stage, we firstly sample a set of author pairs that haven't co-authored in T_1, extract associated features in T_1, and record whether a relationship is to occur between them in period T_2. A training model is built to learn the best coefficients associated with each feature by maximizing the likelihood of relationship formation. In the test stage, we apply the learned coefficients to features of the author's pair in the test set and compare the predicted relationship with ground truth.

To evaluate the link prediction accuracy under supervised learning framework, there are several common metrics including Accuracy, ROC-AUC, Precision-Recall and so on. In this study, we chose two metrics Accuracy and ROC-AUC to evaluate the performance of co-author relationship prediction in which the former is the classification accuracy rate for binary prediction under the cut-off score as 0.5 and the area under ROC curve (AUC) for the later.

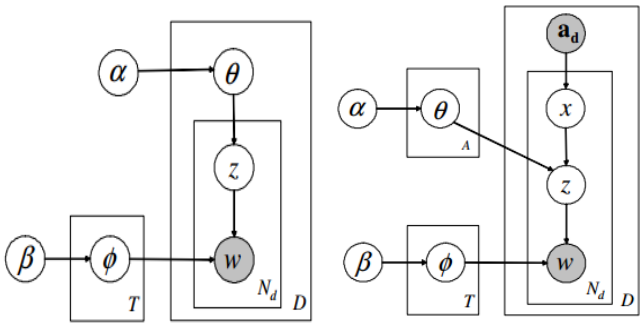
2.2 Topic Modeling

2.2.1 Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation(LDA) [10] is a generative statistical model of a corpus. In LDA, each document may be considered as an amalgamation of various topics and each topic is illustrated by a word's probability distribution. The generative model of LDA is described with the probabilistic graphical model in Figure 1a, proceeds as follows:

1. Choose distribution over topics $\theta_i \sim$ Dirichlet(α) for each document.
2. Choose distribution over words $\phi_j \sim$ Dirichlet(β) for each topic.
3. For each of the word position i, j:
   3.1. Choose a topic $z_{ij} \sim$ Multinomial(θ_i)
   3.2. Choose a word $w_{i,j} \sim$ Multinomial($\phi_{z_{i,j}}$)

2.2.2 Author-Topic Modeling (ATM). Author-Topic model (ATM) [12] is an expanded model from LDA with incorporate author's information. Each author is associated with a combination of topics where topics are multinomial distributions over words. The words in a collaborative paper are assumed to be the result of a mixture of the authors' topics. The generative model of ATM is described with the probabilistic graphical model in Figure 1b, proceeds as follows:

1. For each author a=1, ..., A choose $\theta_a \sim$ Dirichlet(α)
   For each topic t=1,..T choose $\phi_t \sim$ Dirichlet(β)
2. For each document d=1,.., D
   2.1. Given the vector of authors a_d
   2.2. For each word i=1,...,N_d
      2.2.1. Choose an author $x_{di} \sim$ Uniform(a_d)
      2.2.2. Choose a topic $z_{di} \sim$ Discrete($\theta_{x_{di}}$)
      2.2.3. Choose a word $w_{di} \sim$ Discrete($\phi_{z_{di}}$)



(a) LDA (b) ATM

Figure 1: Topic modeling

Table 1: Topological relationships in DBLP network

Relation	Semantic Meaning of Relation
APA	a_i and a_j are co-authors (target relation)
APAPA	a_i and a_j are co-authors of common authors
APVPA	a_i and a_j publish in the same venues
AAFA	a_i and a_j have relation related to geographic factor

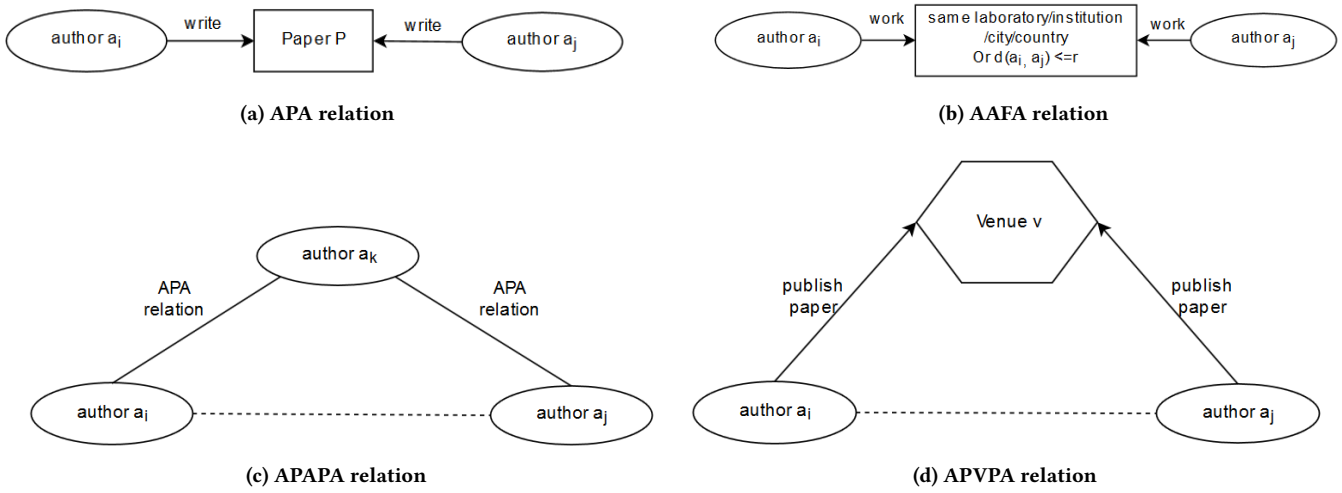
3 PROPOSED APPROACH

In this section, we describe our approach in tackling the problem of predicting co-author link accurately in a given future time interval T_2 based on available factors of the network in the past time interval T_1. We utilize supervised methods for link prediction and feature's exploitation and estimation will be concentrated on this study. In the training stage, the first step we will sample a set of author pairs that haven't co-authored in the past period T_1 and extract features. In the next step, we utilize a machine learning method to built a training model to learn the best coefficients associated with each feature by maximizing the likelihood of relationship formation. In the test stage, we apply the learned coefficients to features of author pairs in the test set and compare predicted accuracy with ground truth.

The feature selection plays a significant role in the decision of algorithm performance of machine learning. Therefore, in this study, we will concentrate to exploit new topological feature and content feature based on textual information. Firstly, our approach is to consider the geographic factor in topological feature extraction. Secondly, we extract the content feature based on textual information from the author's papers using topic modeling. Finally, we combine topological features and content feature in co-author relationship prediction. Features extraction for link prediction will be described details in following subsections.

3.1 Topological Features

In this study, we consider four topological relationships between authors that each relation corresponds to one feature which is utilized in link prediction (see in Table 1). The first three relationships



(a) APA relation **(b) AAFA relation**

(c) APAPA relation **(d) APVPA relation**

Figure 2: Relations in bibliographic network

include APA, APAPA, and APVPA had been considered in previous researches [11, 15, 16], however; there is no study take into account to geographic factor, for instances, two authors have the same laboratory, same research institution, same city, country or geographic distance from their workplaces and so on. It can be said that geographic factor plays a significant role in co-author relationship formation since two researchers work in the same laboratory or research institution have high probability to connect and write paper together or supposing there is short geographic distance from their workplaces, they may have an opportunity to meet the other, connect, discuss and research common problems. Therefore, in this study, besides three existing relations, we consider a new relationship related to geographic factor, namely AAFA. We will describe four relations in detail in subsubsections.

3.1.1 Relation APA. An author a_i have relation APA with author a_j means that a_i and a_j are co-authors. This relation is target relation that we need to predict in future time interval (see figure 2a) based on information in the past time interval.

3.1.2 Relation APAPA. An author a_i have relation APAPA with author a_j means that a_i and a_j have common co-authors. Figure 2c illustrates author a_i and a_j have relation APAPA since they have common co-author a_k. There are several common attributes computed for each node pair (a_i, a_j) from APAPA relation in co-author relationship prediction [7, 8, 15, 18] including:

- Common Neighbours: Common neighbors is defined as the number of common neighbors of two authors a_i and a_j, namely $|\Gamma(a_i) \cap \Gamma(a_j)|$, where $\Gamma(a_i)$ and $\Gamma(a_j)$ are neighbor sets of author a_i, a_j respectively.
- Jaccard's cofficient: Jaccard's cofficient is a normalized measure of common neighbors, namely $\frac{|\Gamma(a_i) \cap \Gamma(a_j)|}{|\Gamma(a_i) \cup \Gamma(a_j)|}$
- Adamic/Adar: Adamic/Adar measure similarity between two nodes by weighting "rarer" common neighbors more heavily, namely $\sum_{a_k \in \Gamma(a_i) \cap \Gamma(a_j)} \frac{1}{log|\Gamma(a_k)|}$
- Path count: Path count measures the number of path instances between two objects following a given meta path, denoted as PC_R where R is relation denoted by the meta path. For APAPA relations, Path count can be calculated by the products of adjacency matrics associated with relation APAPA in the meta path.

3.1.3 Relation APVPA. An author a_i have relation APVPA with author a_j means that a_i and a_j have published their papers in the same conference or journal (see figure 2d). This relation implicitly demonstrates interests in research's field, for instance, author a_i and a_j usually publish their papers into conference NIPS, this is synonymous with they are interested in Neural Information Processing. Moreover, if they participate in the same conferences or workshops, they can opportunity to meet, discuss and connect in research. Therefore, relation APVPA contributes significant feature in co-author relation prediction. In previous research [15], author utilized path count to measure the number of path from a_i to a_j. Besides, we also can utilize follow attributes to compute for each node pair (a_i, a_j) from APVPA relation in co-author relation prediction such as:

- Common Venues: Common venues (conferences/workshops) is defined as the number of common venues of two authors a_i and a_j publish their papers, namely $|V(a_i) \cap V(a_j)|$, where $V(a_i)$ and $V(a_j)$ are sets of venues of a_i, a_j publish their papers respectively.
- Jaccard's cofficient: Jaccard's cofficient is a normalized measure of common venues, namely $\frac{|V(a_i) \cap V(a_j)|}{|V(a_i) \cup V(a_j)|}$

3.1.4 Relation AAFA. This study, we concentrate on exploit new relationship related to geographic factor, namely AAFA. There are numerous ways to define relation AAFA between two authors (see figure 2b), for instance:

- Binary relationship: author a_i and a_j have relation AAFA when they work in the same laboratory or institution, city, country. The similarity about geopraphy between a_i and a_j

can be defined as binary value 0 or 1 where value 1 correspond to a_i and a_j have same laboratory(institution/city/country) or vice versa.

- Metric space: author a_i and a_j have relation AAFA when their workplace distance $d(a_i, a_j) <= r$. The similarity about geopraphy between a_i and a_j can be defined as geographic distance.

3.2 Content Feature

Majority research about co-author relation prediction on both homogeneous and heterogeneous networks considered topological features [1, 6–8, 11, 15–17], rarely mentioned about the paper's content of authors. However, the paper's content plays a significant role in co-author relationship formation since there is a higher probability to contact and connect for two authors research on the same narrow field compare with the dissimilar fields. Although in [6, 17] textual information had been exploited, authors just count the number of common keywords [6] or utilized TFIDF feature vector representation and the cosine measure to compute similarity from the title of papers [17]. Vector Space Model (VSM) [13] is a fundamental technique for textual analysis where each document is represented by a word-frequency vector. Two disadvantages of VSM are the high dimensionality as a result of the high number of unique terms in text corpora and insufficient to capture all semantics. Therefore, in this study, we consider the content feature base on textual information to estimate the author's interest similarity using topic modeling technology. The textual content can be extracted from keywords of papers, titles, abstracts or full text from papers. Topic modeling identifies the distribution of latent topics in the text, which is useful in modeling the interest distribution. Recently, there are dissimilar methods of topic modeling which include Latent Dirichlet Allocation (LDA), Author-Topic Model (ATM), etc. In this study, we will choose LDA and ATM to estimate the topic's probability distribution of authors. We measure the similarity between authors based on their topic's distribution. Experimental results in our previous work [2] demonstrated that probability-based distance is better than vector-based distance. Therefore, it is better if we choose distance measures related to the probability distribution such as KullbackLeibler Divergence, Jensen-Shannon divergence, Hellinger distance, etc.

KullbackLeibler Divergence:

$$d_KL(P||Q) = \sum_{x \in X} P(x) \frac{P(x)}{Q(x)} \tag{1}$$

Jensen-Shannon distance:

$$d_{JS}(P,Q) = \frac{1}{2} \sum_{i=1}^{k} p_i ln \frac{2p_i}{p_i + q_i} + \frac{1}{2} \sum_{i=1}^{k} q_i ln \frac{2q_i}{p_i + q_i} \tag{2}$$

Hellinger distance:

$$d_H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{3}$$

Table 2: Sumnmary statistics for three co-author networks from 1995 to 2015

	Net1	Net2	Net3
Number of Nodes	3079	4321	13605
Number of Edges	10006	13122	29637
Density	0.002	0.0014	0.0003
Diameter	13	17	23
Clustering coefficient	0.769	0.767	0.713
Average degree	6.5	6.074	4.357
Largest component	98.8%	99.7%	59%

4 EXPERIMENTS AND RESULTS

In this section, firstly we illustrate that combination between new topological feature AAFA and existing topological features can improve the co-authorship prediction accuracy compared with the baselines which only using existing topological features. Moreover, we reveal that content feature estimation based on textual information using topic modeling can enhance performance for predicting co-author relation instead of existing techniques such as TFIDF or counting number common keywords. Finally, we demonstrate that the combination of topological features and the content feature in co-author relation prediction following our proposed method can reach the highest result.

4.1 Experiments

4.1.1 Dataset. We utilized dataset "DBLP-SIGWEB.zip" which is derived from September 17, 2015 snapshot of dblp bibliography database. It contains all publications and authors records of 7 ACM SIGWEB conferences:

- ACM conference on Hypertext and social media (HT)
- Joint Conference on Digital Libraries (DL)
- Document Engineering (DocEng)
- Web Science (WebSci)
- Conference on Information and Knowledge and Management (CIKM)
- Conference on Web Science and Data Mining (WSDM)
- User Modeling, Adaptation and Personalization (UMAP) Conference

The dataset also contains the authors, chairs, affiliations and additional metadata information of conferences that are published in ACM digital library.

In this section, we implement experiments on three different networks extracted from the dataset "DBLP-SIGWEB.zip". The first network (Net1) is constructed derived from random 50 authors with degree larger 10. After that, we get co-authors of those 50 authors. Finally, we get more co-authors of all the above authors. The second network (Net2) is constructed similar to Net1, but the difference is to init 50 nodes with the largest publication. Finally, the third network (Net3) is network corresponding with the whole dataset "DBLP-SIGWEB.zip". For Net1 and Net2, we extract related metadata information correspond to authors in the network. Several summary statistics for three co-author networks from 1995 to 2015 are described in the Table 2.

Table 3: Feature's combination

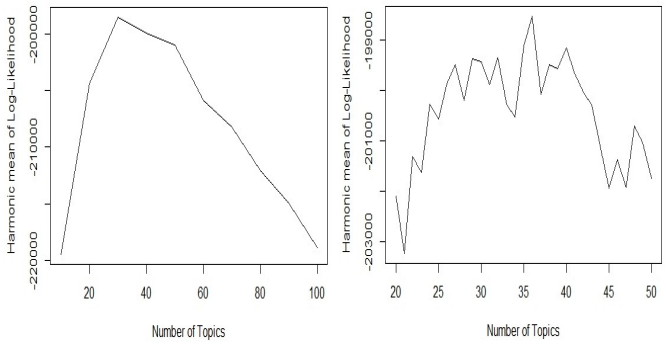
Topological features	Content feature
1. APAPA	8. Number common keywords
2. APVPA	(CK)
3. AAFA	9. TFIDF
4. AAFA + APAPA	10. ATM
5. AAFA + APVPA	11. LDA
6. APAPA + APVPA	
7. AAFA + APAPA + APVPA	
Topological features & Content feature	
12. AAFA + APAPA + APVPA + CK	
13. AAFA + APAPA + APVPA + TFIDF	
14. AAFA + APAPA + APVPA + ATM	
15. AAFA + APAPA + APVPA + LDA	

4.1.2 Experiment Setting. We consider two time intervals for the network, according to the publication year associated to each paper: $T_1 = [1995, 2010]$ and $T_2 = [2011, 2015]$. For tranning stage, we utilize T_1 as past time interval and T_2 as the future time interval. We consider an author pair (a_i, a_j) where a_i, a_j be called source author and target author respectively. Firstly, we find all the source authors that they have relationships building with existing authors in the future time interval of T_2, and use these new relationships as positive training pairs. We also sample an equal-sized of negative pairs. Therefore, in the training dataset, the size of the positive pairs is balanced with negative pairs.

For our experiments, we utilize classification methods as the prediction model. We perform experiments with different sets of features and evaluate the incremental performance improvement. These feature's combination is shown in Table 3.

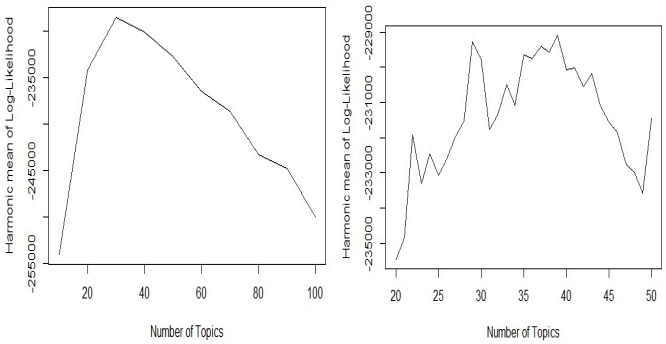
In the first group of topological features, we chose Common Neighbours to calculate similarity for feature APAPA and Common Venues for APVPA. For feature AAFA, we constructed relationship AAFA as a binary relationship based on the same laboratory information. We implemented seven experiments with combination from three relations APAPA, APVPA, and AAFA. The first three experiments are experiments correspond to single topological features. The experiments 4 and 5 are combinations between one old feature APAPA and APVPA with new feature AAFA. Moreover, we experimented 6 with the composition of two old feature APAPA and APVPA and the amalgamation of those features with new feature AAFA in experiment 7. We utilized the results of experiments 1, 2 and 6 as baselines to compare with experiments 4, 5 and 7 respectively, from that point on the significance of feature AAFA.

For content feature, we collected textual information from keywords of author's papers in the past time interval T_1 since in scientific publication, keywords play a significant role in illustrating the specific domain of researchers works. To estimate the topic's probability distribution of authors using topic modeling, we need to estimate the number of the topic in the corpus. Firstly, we defined the number of the topic for the whole corpus based on the Harmonic mean of Log-Likelihood (HLK) [3]. We calculated HLK with the number of topics in the range [10, 100] with sequence 10. We realized that the best number of topics is in the range [20, 50] for Net1 and Net2 (Figure 3a and 4a), [30, 60] (Figure 5a) for Net3. Therefore, we ran HLK again with sequence 1 and obtained the



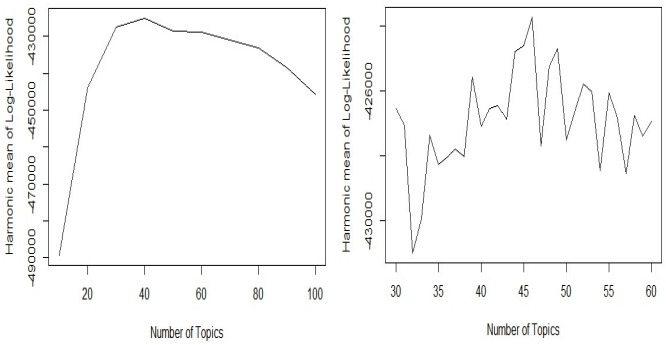
(a) ♯ topics ∈ [10, 100], seq 10 **(b) ♯ topics ∈ [20, 50], seq 1**

Figure 3: Log-likelihood for Net1



(a) ♯ topics ∈ [10, 100], seq 10 **(b) ♯ topics ∈ [20, 50], seq 1**

Figure 4: Log-likelihood for Net2



(a) ♯ topics ∈ [10, 100], seq 10 **(b) ♯ topics ∈ [20, 50], seq 1**

Figure 5: Log-likelihood for Net3

best is 36 for Net1 (Figure 3b), 39 for Net2 (Figure 4b) and 46 for Net3 (Figure 5b). After defining the number of topics, we estimated the topic's probability distribution of authors using ATM and LDA. There are already several available packages for topic modeling

including *topicmodels* or *lda* in R, or *Gensim* [1] in Python. In this study, we chose Gensim for training the topic modeling. Finally, we implemented experiments 10, 11 with the feature is the similarity between author pairs based on their topic's probability distribution. In this study, we chose Hellinger distance to measure the distance between the two topic's distribution probabilities. Besides, we conducted experiments 8 and 9 as baselines respect with two existing methods are to counting the number of common keywords and TFIDF as in [6, 17].

Finally, we implemented experiments from 12 to 15 correspond to the amalgamation of three topological features (AAFA, APAPA, APVPA) and content feature with three estimating methods including number common keywords, TFIDF and topic modeling (ATM and LDA).

There exist various classification algorithms for supervised learning. Although their performances are comparable, some usually work better than others for a specific dataset or domain. In this research, we experimented with three different classification algorithms including Support Vector Machine (SVM, Linear Kernel), Decision Tree (DT) and Random Forest (RF).

4.2 Results

Experimental results demonstrated performance of our classifiers using different sets of features showed in Table 4, 5 and 6. Firstly, classification results on Net1 are shown in Table 4. For all three classification algorithms, we can see that when combining one existing topological feature (APAPA or APVPA) with new feature AAFA, the accuracy of classification outperforms comparison with just using one of them. Moreover, the amalgamation of both existing feature APAPA and APVPA with new feature AAFA obtained higher accuracy compare with just utilizing APAPA and APVPA. Particularly, Support Vector Machine and Random Forest reached better accuracy compared with Decision Tree. This illustrates the significance of new topological feature AAFA in contribution to improve the performance of co-author relationship prediction. On the other hand, for classification with the content feature, Support Vector Machine and Random Forest express that estimating textual information from keywords using topic modeling bring higher performance compared to two old methods TFIDF and number common keywords. Especially, Random Forest with LDA reached the highest accuracy. Finally, the highest accuracy is reached by Random Forest with combination between three topological features and content feature using topic modeling with LDA. Figure 6 shows the performance of our Random Forest classifiers on Net1 using different combinations of features.

Next, second experimental results on Net2 and Net3 are shown in Table 5 and 6 respectively. Similar to classification results with topological features on Net1, results of classification on Net2 and Net3 continue to demonstrate the importance of new topological feature AAFA when combining it with two existing topological features APAPA and APVPA on all different classification methods. On the other hand, for content feature, the use of topic modeling with LDA in Random Forest bring high efficiency on Net3 while both ATM and LDA in Support Vector Machine and Random Forest give effective performances in Net2 compare with two existing methods

[1] https://pypi.python.org/pypi/gensim

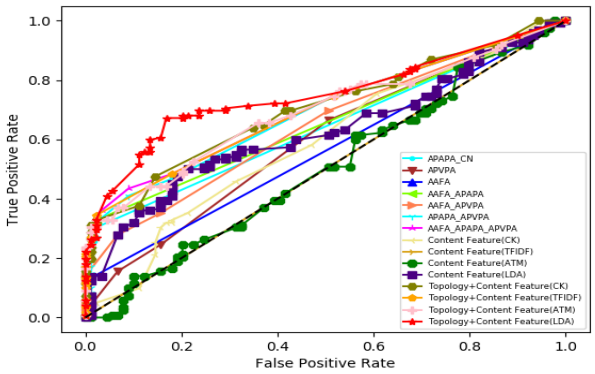


Figure 6: ROC curve of Random Forest classifier on different feature sets on Net1

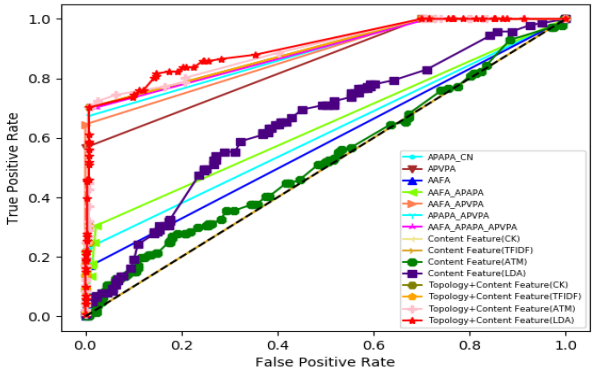


Figure 7: ROC curve of Random Forest classifier on different feature sets on Net2

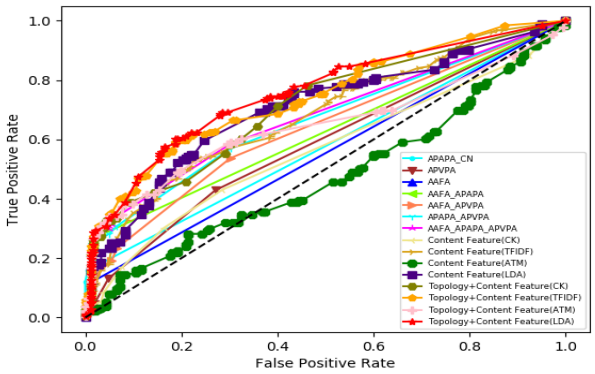


Figure 8: ROC curve of Random Forest classifier on different feature sets on Net3

number common keywords and TFIDF. Finally, the highest accuracy of co-author relationship prediction on Net2 is obtained by Random

Table 4: Results of Classification on Net1

Features	Prediction Accuracy					
	SVM		DT		RF	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
APAPA	0.588	0.641	0.588	0.641	0.588	0.641
APVPA	0.592	0.594	0.588	0.583	0.592	0.594
AAFA	0.55	0.561	0.493	0.561	0.493	0.561
AFAA + APAPA	**0.602**	**0.654**	**0.602**	**0.655**	**0.602**	**0.655**
AFAA + APVPA	**0.611**	**0.640**	**0.602**	**0.630**	**0.611**	**0.645**
APAPA + APVPA	0.616	0.695	0.607	0.689	0.626	0.693
AAFA + APAPA + APVPA	**0.635**	**0.703**	**0.611**	**0.695**	**0.635**	**0.703**
CK	0.536	0.595	0.526	0.579	0.559	0.591
TFIDF	0.422	0.5	0.422	0.5	0.422	0.5
ATM	0.5	0.5	0.555	0.566	0.479	0.504
LDA	**0.588**	**0.659**	0.531	0.533	**0.611**	**0.627**
AAFA + APAPA + APVPA + CK	0.635	0.71	0.597	0.647	0.602	0.705
AAFA + APAPA + APVPA + TFIDF	0.635	0.703	0.611	0.695	0.626	0.700
AAFA + APAPA + APVPA + ATM	0.635	0.714	0.545	0.537	0.592	0.693
AAFA + APAPA + APVPA + LDA	**0.668**	**0.75**	0.588	0.587	**0.687***	**0.75***

Table 5: Results of Classification on Net2

Features	Prediction Accuracy					
	SVM		DT		RF	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
APAPA	0.581	0.611	0.581	0.611	0.581	0.611
APVPA	0.654	0.5	0.85	0.849	0.85	0.849
AAFA	**0.705**	**0.579**	**0.705**	**0.579**	**0.705**	**0.579**
AFAA + APAPA	**0.744**	**0.642**	**0.744**	**0.641**	**0.744**	**0.641**
AFAA + APVPA	**0.705**	**0.584**	**0.877**	**0.876**	**0.877**	**0.876**
APAPA + APVPA	0.725	0.568	0.882	0.883	0.882	0.881
AAFA + APAPA + APVPA	**0.744**	**0.7**	**0.892**	**0.893**	**0.892**	**0.887**
CK	0.565	0.603	**0.654**	**0.5**	0.654	0.5
TFIDF	0.654	0.5	**0.654**	**0.5**	0.654	0.5
ATM	0.654	0.5	0.6	0.547	**0.649**	**0.528**
LDA	**0.661**	**0.653**	0.570	0.526	**0.656**	**0.649**
AAFA + APAPA + APVPA + CK	0.627	0.656	**0.892**	**0.893**	0.754	0.891
AAFA + APAPA + APVPA + TFIDF	0.744	0.7	**0.892**	**0.893**	0.754	0.891
AAFA + APAPA + APVPA + ATM	**0.754**	**0.615**	0.833	0.822	**0.892***	**0.898***
AAFA + APAPA + APVPA + LDA	0.727	0.731	0.799	0.794	**0.843**	**0.907***

Table 6: Results of Classification on Net3

Features	Prediction Accuracy					
	SVM		DT		RF	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
APAPA	0.584	0.577	0.584	0.577	0.584	0.577
APVPA	0.581	0.586	0.581	0.586	0.581	0.586
AAFA	0.546	0.553	0.546	0.553	0.546	0.553
AFAA + APAPA	**0.619**	**0.625**	**0.619**	**0.624**	**0.619**	**0.624**
AFAA + APVPA	**0.616**	**0.637**	**0.611**	**0.634**	**0.616**	**0.637**
APAPA + APVPA	0.632	0.666	0.619	0.66	0.632	0.667
AAFA + APAPA + APVPA	**0.632**	**0.679**	**0.632**	**0.668**	**0.641**	**0.678**
CK	0.565	0.603	0.568	0.551	0.57	0.563
TFIDF	**0.665**	**0.704**	**0.592**	**0.587**	0.643	0.688
ATM	0.511	0.488	0.508	0.508	0.497	0.476
LDA	0.635	0.682	0.584	0.583	**0.662**	**0.704**
AAFA + APAPA + APVPA + CK	0.627	0.656	**0.646**	**0.698**	0.632	0.712
AAFA + APAPA + APVPA + TFIDF	**0.678**	**0.749**	0.622	0.619	0.7	0.744
AAFA + APAPA + APVPA + ATM	0.632	0.679	0.6	0.6	0.63	0.648
AAFA + APAPA + APVPA + LDA	0.662	0.733	0.630	0.630	**0.7***	**0.756***

Forest with composition between three topological features and content feature using topic modeling with ATM while LDA for Net3. Figure 7 and 8 demonstrate the performance of our Random Forest classifiers on Net2 and Net3 using different combinations of features respectively.

In short, experimental results demonstrate that new topological feature related to geographic factor can contribute to improving the performance of co-author link prediction. Moreover, the utilize of topic modeling to estimate for the content feature can bring effective performance in link prediction compare with two existing

methods are number common keywords and TFIDF. Particularly, the incorporation between topological features and content feature following our approach can obtain the highest performance.

In these experiments, we see that Random Forest outperforms compare with Support Vector Machine and Decision Tree. Perhaps, Support Vector Machine may be less sensitive to the choice of input parameters than Random Forest. Besides, Random Forests are typically more accurate than single decision trees since they consist of multiple single trees each based on a random sample of the training data. There are two reasons why Random Forests outperform single decision trees including trees are diverse and are unpruned. Each random forest tree is learned on a random sample, and at each node, a random set of features are considered for splitting. Therefore, this mechanism creates a diversity of trees. Moreover, while a single decision tree is often pruned, a random forest tree is fully grown and unpruned, and so naturally, the feature space is split into more and smaller regions.

5 CONCLUSION

In this research, we propose a new approach for co-author relationship prediction in a bibliographic network by exploiting geographic factor and latent topic information. The supervised method is utilized for link prediction where we combine different features with the different measuring coefficient. We concentrate on features selection for link prediction since it makes a significant contribution to prediction performance. Firstly, we concentrate to exploit a new topological feature based on relation related to the geographic factor. Besides, we discover content feature based on textual information using topic modeling. Finally, we combine topological features and content feature in co-author relationship prediction. Experimental results illustrated that the presence of a new topological feature related to geographic factor with existing topological features contributes to improving the accuracy of co-author link prediction. Moreover, utilizing topic modeling for estimating content feature from textual information can enhance the accuracy of co-author relationship prediction compare with using existing techniques. Especially, the highest accuracy can be reached from the combination of topological features and content feature following our proposed approach. In future works, we will conduct experiments on datasets where relation AAFA is considered in metric space and the bigger textual information set such as from abstracts or the whole of papers.

REFERENCES

[1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the Web. *Social Networks* 25, 3 (July 2003), 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1

[2] Quang Vu Bui, Karim Sayadi, Soufian Ben Amor, and Marc Bui. 2017. Combining Latent Dirichlet Allocation and K-Means for Documents Clustering: Effect of Probabilistic Based Distance Measures. In *Intelligent Information and Database Systems (Lecture Notes in Computer Science)*, Ngoc Thanh Nguyen, Satoshi Tojo, Le Minh Nguyen, and Bogdan Trawiński (Eds.). Springer International Publishing, 248–257.

[3] Wray Buntine. 2009. Estimating Likelihoods for Topic Models. In *Advances in Machine Learning (Lecture Notes in Computer Science)*, Zhi-Hua Zhou and Takashi Washio (Eds.). Springer Berlin Heidelberg, 51–64.

[4] Fei Gao, Katarzyna Musial, Colin Cooper, and Sophia Tsoka. 2015. Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics. *Sci. Program.* 2015 (Jan. 2015), 1:1–1:1. https://doi.org/10.1155/2015/172879

[5] Lise Getoor and Christopher P. Diehl. 2005. Link mining: a survey. *SIGKDD Explorations* 7 (2005), 3–12. https://doi.org/10.1145/1117454.1117456

[6] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. [n.d.]. Link Prediction using Supervised Learning. ([n. d.]), 10.

[7] David Liben-Nowell and Jon Kleinberg. 2003. The Link Prediction Problem for Social Networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03)*. ACM, New York, NY, USA, 556–559. https://doi.org/10.1145/956863.956972 event-place: New Orleans, LA, USA.

[8] Ryan N. Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. 2010. New Perspectives and Methods in Link Prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. ACM, New York, NY, USA, 243–252. https://doi.org/10.1145/1835804.1835837 event-place: Washington, DC, USA.

[9] Linyuan Lu and Tao Zhou. 2011. Link Prediction in Complex Networks: A Survey. *Physica A: Statistical Mechanics and its Applications* 390, 6 (March 2011), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027 arXiv: 1010.0725.

[10] David M. Blei, Andrew Y. Ng, and Michael Jordan. 2001. Latent Dirichlet Allocation. In *The Journal of Machine Learning Research*, Vol. 3. 601–608.

[11] Alexandrin Popescul, Rin Popescul, and Lyle H. Ungar. 2003. *Statistical Relational Learning for Link Prediction.*

[12] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The Author-Topic Model for Authors and Documents. In *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*. Canada, 487–494. https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1145&proceeding_id=20

[13] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24, 5 (Jan. 1988), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

[14] Virinchi Srinivas and Pabitra Mitra. 2016. Applications of Link Prediction. In *Link Prediction in Social Networks: Role of Power Law Distribution*, Virinchi Srinivas and Pabitra Mitra (Eds.). Springer International Publishing, Cham, 57–61. https://doi.org/10.1007/978-3-319-28922-9_5

[15] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han. 2011. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*. 121–128. https://doi.org/10.1109/ASONAM.2011.112

[16] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. 2003. Link Prediction in Relational Data. In *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'03)*. MIT Press, Cambridge, MA, USA, 659–666. http://dl.acm.org/citation.cfm?id=2981345.2981428 event-place: Whistler, British Columbia, Canada.

[17] Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. 2007. Local Probabilistic Models for Link Prediction. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, Omaha, NE, USA, 322–331. https://doi.org/10.1109/ICDM.2007.108

[18] Yang Yang, Ryan N. Lichtenwalter, and Nitesh V. Chawla. 2015. Evaluating Link Prediction Methods. *Knowledge and Information Systems* 45, 3 (Dec. 2015), 751–782. https://doi.org/10.1007/s10115-014-0789-0 arXiv: 1505.04094.