Collaboration Prediction in Heterogeneous Information Networks

Shuhong Zhang, Feng Xia, Jun Zhang, Xiaomei Bai, Zhaolong Ning
School of Software, Dalian University of Technology, Dalian 116620, China
Email: shuhong.zhang0@gmail.com, f.xia@ieee.org, junezhang8900@gmail.com, xiaomeibai@outlook.com, zhaolongning@dlut.edu.cn

***Abstract*—To reveal the information hiding in the scholarly big data, relationship analysis among academic entities has been studied from different perspectives in recent years. In this paper, we focus on the problem of collaboration relationship prediction between authors in heterogeneous information networks, and a new method called *MACP*, *i.e.*, Meta path and author Attribute based Collaboration Prediction model, is proposed to solve this problem. We use a two-phase collaboration probability learning approach. First, topological features with author attributes are extracted from the network, and then a supervised learning algorithm is employed to find the best weight associated with each feature to determine the collaboration relationship. We present the experiments on a real information network, namely the APS network, which shows that our proposed model can generate more accurate results compared with the method only considering structural features.**

***Index Terms*—collaboration prediction, meta path, heterogeneous information network.**

1. Introduction

With the explosive growth of research works and publications in recent years, scholarly big data has become a hotspot, and one of the most important components of this issue is the relationship analysis among different academic entities. Various types of relationships between authors deserve to be studied, for example, friendship, co-authorship, advisor-advisee relationship and so on. We focus on the problem of collaboration relationship prediction in this paper, which aims to predict whether two authors that have never collaborated before will build the collaboration relationship sometime in the future, rather than predicting how many times two authors will collaborate in ahead. Learning about the future collaboration relationship of an author is helpful to understand the author's academic circles, whether the author is a cooperated or an independent researcher, and the mechanism behind the collaboration relationship formation.

Collaboration relationship prediction in bibliographic networks aims to estimate the likelihood that the collaboration relationship form between two authors, based on the observations of existing co-author relationships and the attributes of the authors. Most of the existing relationship prediction studies ([1], [2] ,[3]) are conducted on the homogeneous networks, which contain only one type of objects in the networks, such as co-author, citation, as well as co-citation networks. These networks are either extracted from original heterogeneous networks or treat objects and links of different types equally, which can lead to the loss of comprehensive information in the networks. Recent researchers turn to study this problem in the heterogeneous information networks, which consist of multiple types of objects and links. In [4], [5], [6], Sun et al. propose the meta path-based method to analyze the heterogeneous information networks. Along this line, meta path-based methods are leveraged by some other studies ([7], [8], [9], [10]) to mine the comprehensive knowledge in the heterogeneous information networks. These researches focus on the problem of similarity and relevance search, as well as relationship prediction, just on the basis of the features extracted from the connection situation of the heterogeneous information networks. However, the collaboration relationship building between two authors can be influenced by many factors, such as the evolution of the information networks and the object attributes. In this paper, we combine the meta path-based features and some other non-structural attributes to infer the probability of collaboration formation in the future. The contributions of this paper are as follows:

- We study the problem of collaboration prediction in the heterogeneous information networks.
- We propose a new method called *MACP*, which incorporates transitive similarity, temporal dynamics and author attributes into meta path-based topological features, and build a logistic regression model for collaboration prediction.
- Experiments on the real APS heterogeneous information network show that the prediction accuracy can be improved, through considering both the meta path-based and non-structural features together.

The remaining of the paper is organized as follows. We introduce the concepts on heterogeneous information networks and meta paths in Section 2, and then Section 3 describes the proposed features and model in detail. The dataset, experimental results and discussion are provided in Section 4. We concludes the paper and suggest future research in Section 5.

2. Preliminary

In this section, we introduce some basic concepts related to heterogeneous information networks. A heterogeneous information network represents an abstraction of the real world, which either contains multiple types of objects or multiple types of links. It provides us comprehensive information for better understanding the relationships among the objects in reality.

Definition 1 (Information Network). *An information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\tau : V \to A$ and a link type mapping function $\phi : E \to R$, where each object $v \in V$ belongs to one particular object type $\tau(v) \in A$, each link type $e \in E$ belongs to a particular relation $\phi(e) \in R$, and if two links belongs to the same relation type, the two links share the same starting object type as well as the ending object type.*

Note that, if a relation exists from type A to type B, denoted as ARB, the inverse relation R^{-1} holds naturally for $BR^{-1}A$. R and its inverse R^{-1} are usually not equal, unless the two types are the same and R is symmetric. When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is classified as heterogeneous information network; otherwise, it is a homogeneous information network.

Furthermore, it is necessary to learn about the meta level (*i.e.*, schema-level) description of a heterogeneous information network, for better understanding the various relationships between different object types and link types. Therefore, we introduce the concept of network schema to describe the meta structure of a network.

Definition 2 (Network Schema). *The network schema, denoted as $T_G = (A, R)$, is a meta template for a heterogeneous network $G = (V, E)$ with the object type mapping $\tau : V \to A$ and the link mapping $\phi : E \to R$, which is a directed graph defined over object types A, with edges as relations from R.*

The network schema of a heterogeneous information network specifies type constraints on the sets of objects and links between the objects, and defines the rules of network formation. It leads to the exploration of the semantics of the network. An information network following a network schema is then called a *network instance* of the network schema.

Example 1 (Bibliographic information network). *A bibliographic information network, such as American Physical Society (APS) [1], is a typical heterogeneous information network. The network schema of APS dataset is shown in Fig. 1. It contains objects from seven types of entities: papers (P), authors (A), affiliations (F), terms (T), subjects (S), periodicals (R), and journals (J) (a journal includes multiple periodicals,* e.g., *PRL including PRL2014, PRL2015). The link types are defined by the relations between two object types. For example, links exist between authors and papers denoting the writing ($write$) or written-by ($write^{-1}$) relations.*

1. http://www.journals.aps.org/datasets

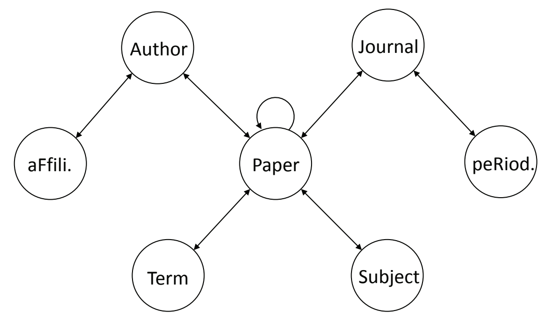

Figure 1: APS network schema

In order to distinguish from path semantics and capture the link type information between two objects, we use the concept of meta path from [4] in a network schema, which is formally defined as follows.

Definition 3 (Meta path). *A meta path P is a path defined on the graph of network schema $T_G = (A, R)$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between type A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.*

The length of P is the number of relations in P. For example, in the APS network schema, the collaboration relation can be described using length-2 meta path $A \xrightarrow{write} P \xrightarrow{write^{-1}} A$, short as APA, if there is no ambiguity. And the reverse meta path, denoted as P^{-1}, defines an inverse relation of the one defined by P. Path instances following meta path P represent concrete link composition between objects. Also, two meta paths are concatenable if and only if the ending object of the first path instance is the same as the starting object of the second one.

3. The MACP Model

In this section, we introduce the MACP model in detail, which includes two stages: (1) the meta path and author attributes based topological feature definition, and (2) the logistic regression-based collaboration prediction model.

3.1. Topological Features in Heterogeneous Networks

Link prediction on the basis of topological features aims to predict the future connectivity through current topological situation of the network. Structural features are extracted from networks, which describe the connectivity properties of object pairs. Also diverse information can be associated with information networks in order to aid to infer the future links, for example, temporal dynamics of the networks, the similarity of pair objects with the same type, and the own attributes attached to the objects.

For collaboration prediction in heterogeneous information networks, we propose three topological features, which combine the basic structural connections with the attached information referred above, which can be extracted from the networks. We first define the basic structural measures of meta path, and the personal attributes of the authors, then we obtain the topological features used for collaboration prediction.

3.1.1. Measures on Meta Paths. Each meta path defines a unique structure for a pair of objects, on behalf of a particular relation. To obtain all the meta paths between two objects, we make use of traversing methods such as analogous breadth-first search algorithm on the APS network schema. For collaboration relation, we extract all the meta paths with the length-constraint, whose start and end type are author, and the target relation is described as APA.

The most important component of meta path-based features is the measures on different meta paths, once the structures given by meta paths are obtained. Three basic measures on meta paths are defined in [5] as follows.

(1) **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 the relation depicted by the meta path. Path count can be calculated through the products of adjacency matrices associated with each relation in the meta path.

(2) **Normalized path count.** To discount the number of path counts between two objects through the overall connectivity, normalized path count is defined as $NPC_R(a_i,a_j) = \frac{PC_R(a_i,a_j)+PC_{R-1}(a_j,a_i)}{PC_R(a_i,\cdot)+PC_R(\cdot,a_j)}$, where $PC_R(a_i,\cdot)$ denotes the total number of paths following R starting with a_i, and $PC_R(\cdot,a_j)$ depicts the total number of paths following P ending with a_j.

(3) **Symmetric Random Walk.** Random walk measure along a certain meta path, is defined as $RW_R(a_i,a_j) = \frac{PC_R(a_i,a_j)}{PC_R(a_i,\cdot)}$. Symmetric random walk takes the two directional random walk along the meta path into consideration, and can be defined as $SRW_R(a_i,a_j) = RW_R(a_i,a_j) + RW_{R^{-1}}(a_j,a_i)$.

Given a meta path $P = A_1A_2\cdots A_l$, these structural measures can be computed through the commuting matrix M, which is defined as $M = W_{A_1A_2}W_{A_2A_3}\cdots W_{A_{l-1}A_l}$, where $W_{A_iA_j}$ is the adjacency matrix between type A_i and type A_j. M_{ij} denotes the number of paths between objects $a_i \in A_1$ and objects $a_j \in A_l$ following meta path P, $PC_R(a_i,a_j) = M_{ij}$, and $PC_R(a_i,\cdot) = \sum_j M_{ij}$.

3.1.2. Author Attributes. In order to improve the prediction accuracy and generate a more comprehensive feature space, we introduce some attributes attached to the objects, such as author attributes, including activity, influence and collaboration tendency of the starting and ending authors from the given meta path. We incorporate time-aware factor into the author attributes, due to the temporal dynamics of these attributes as authors' publication and citation vary. Then we calculate these attributes during a specific period of time as follows.

(1) **Activity.** In general, authors prefer to make new collaborations with active authors. Thus, we assume that the more active the author is, the more possible the author can collaborate with others. Activity score is defined in [11], which quantifies the activity degree of an author.

$$f_{activity}(a) = \sum_{[t_i,t_{i+1}]}^{i=0\cdots n} N(a)[t_i,t_{i+1}] * \exp(-\delta(t)) \quad (1)$$

where t_0 and t_n denote the start and end time of the data used for computing the attributes, $N(a)[t_i,t_{i+1}]$ is number of publications by author a during the period $[t_i,t_{i+1}]$, $\delta(t) = t_n - t_i$ depicts the number of years from year t_i to year t_n, and $t_{i+1} = t_i + period$ ($period$ represents a sliding window of time, and various values can be assigned, *e.g.* 1,2, or 3 years).

(2) **Influence.** Authors always tend to collaborate with others who are with high influence, and authors' impact is usually displayed by their number of citations. Then we define the influence of an author in Eq.(2) .

$$f_{influence}(a) = \sum_{[t_i,t_{i+1}]}^{i=0\cdots n} C(a)[t_i,t_{i+1}] * \exp(-\delta(t)) \quad (2)$$

where $C(a)[t_i,t_{i+1}]$ denotes the number of citations of author a during the period $[t_i,t_{i+1}]$, and other parameters have the same meaning with those explained in Eq.(1).

(3) **Collaboration Tendency.** The authors who have coauthored with numerous authors before are more likely to collaborate with others in the future. Thus we propose the concept of collaboration tendency, collaboration level and collaboration degree to describe the probability of an author to be a collaborator. Collaboration level represents the ratio of collaborated papers in all the papers published by the author, and collaboration degree denotes average author numbers of each paper written by the author. In fact, most of the papers are collaborated work, hence collaboration level always approaches 1. We define that collaboration tendency is the product of collaboration level and collaboration degree, shown in Eq.(3).

$$f_{colten}(a) = \sum_{[t_i,t_{i+1}]}^{i=0\cdots n} \frac{Co(a)[t_i,t_{i+1}]}{N(a)[t_i,t_{i+1}]} * \frac{D(a)[t_i,t_{i+1}]}{N(a)[t_i,t_{i+1}]} * \exp(-\delta(t)) \quad (3)$$

where $Co(a)[t_i,t_{i+1}]$ represents the number of collaborated papers of author a during the period $[t_i,t_{i+1}]$, $D(a)[t_i,t_{i+1}]$ denotes the number of distinguished co-authors of author a during the period $[t_i,t_{i+1}]$, and other parameters are the same with the those explained in Eq.(1).

Normalization: To scale the activity, influence and collaboration tendency of authors as $[0,1]$, we normalize the values as follows:

$$f_{attribute}(a) = \frac{f_{attribute}(a) - \min_{a_i\in A}(f_{attribute}(a_i))}{\max_{a_i\in A}(f_{attribute}(a_i)) - \min_{a_i\in A}(f_{attribute}(a_i))} \quad (4)$$

where $f_{attribute}(a)$ represents the author's own attributes, which are activity $f_{activity}(a)$, influence $f_{influence}(a)$ or collaboration tendency $f_{colten}(a)$ of author a. These attributes are all positively correlated with collaboration building. We treat the three attributes as a vector, and obtain the norm as an argument to represent collaboration probability of an author.

3.1.3. Topological Features. Diverse information are associated with information networks, and various attributes can be attached to the nodes or links in a heterogeneous information network. For example, temporal information is often associated with links to reflect the dynamics of an information network. Also, the similarity between two objects is helpful to build collaboration, and the inherent attributes of an author have an influence on the collaboration formation. Therefore, we combine the three factors with the structural meta path-based measures as topological features to predict collaboration, which are path count with temporal dynamics, normalized path count with transitive similarity, and symmetric random walk with author attributes.

(1) **Path count with temporal dynamics.** Heterogeneous information networks evolve over time, and two objects are more probable to establish collaboration relationship if there are more recent connections between them. The newly building meta path instances give more contribution for collaboration in the future. Therefore, we differentiate the impacts of paths formed at different timestamps [7].

Given a meta path $P = A_1A_2 \cdots A_l$, its commuting matrix is $M_P = M^t_{P_1} M^t_{P_2} \cdots M^t_{P_g}$, where $M^t_{P_i}$ is the commuting matrix for meta path P_i with temporal information incorporated. $\sum_{i=1}^{g} l(P_i) = l(P)$, where $l(P_i)$ is the length of short meta path P_i. P_i is a meta path on which an event happens in a particular timestamp. For example, it can be APR in bibliographic networks which represents an author published one paper in a periodical in a particular year. $M^t_{P_i} = M_{P_i} \cdot T_{P_i}$, where M_{P_i} is path count matrix on P_i, and T_{P_i} is the temporal matrix on P_i, with each element representing the weight of path between the start object $x \in A_s(P_i)$ and end object $y \in A_e(P_i)$, where $A_s(P_i)$ is the start object type of meta path P_i and $A_e(P_i)$ is the end object type of P_i. The weight in temporal matrix is assigned according to the timestamp of the path formation. We use a common time decay function to determine the weights, as $f(t) = \alpha^{(t_1 - t)} (t_0 \le t \le t_1)$, where t_0 and t_1 represent the start and end time of data, and $\alpha (0 < \alpha < 1)$ is a variable.

(2) **Normalized path count with transitive similarity.** The meta path instances consisting of more similar objects with the same type are more likely to build collaboration in ahead. Thus, we put different weights on the meta path instance considering the transitive similarity between the start and end objects following the path [7].

Given a meta path $P = A_1A_2 \cdots A_l$, where there exist many objects with the same type as the start and the end objects. Its commuting matrix is $M_P = M^s_{P_1} M^s_{P_2} \cdots M^s_{P_d}$, where $M^s_{P_i}$ is the commuting matrix for short symmetric meta path P_i with transitive similarity integrated. $\sum_{i=1}^{d} l(P_i) = l(P)$, where $l(P_i)$ is the length of short meta path P_i. P_i is a meta path of which the start and end objects with the same type are transitively similar. $M^s_{P_i} = M_{P_i} \cdot S_{P_i}$, where M_{P_i} is the path count matrix on P_i, and S_{P_i} is the transitive similarity matrix on P_i, with each element representing the similarity between the start object $x \in A_s(P_i)$ and end object $y \in A_e(P_i)$, where $A_s(P_i)$ is the start object type of meta path P_i and $A_e(P_i)$ is the end object type of P_i. We utilize a meta path-based peer similarity measure *PathSim* to obtain the transitive similarity between two objects of the same type x and y following a symmetric meta path P, and the definition [4] is: $s(x, y) = \frac{2 \times PC_R(x,y)}{PC_R(x,x) + PC_R(y,y)}$. Then we obtain the normalized path count with transitive similarity leveraging the modified normalized path count formula, with path count with transitive similarity instead of the original path count in the formula. In the feature matrix of normalized path count with transitive similarity, each element is gained from the above commuting matrix with transitive similarity M_P.

(3) **Symmetric random walk with author attributes.** Authors' attributes react the probability of building collaboration relationship with others to some degree, such as the influence, activity and collaboration tendency of authors described in detail in 3.1.2. Hence, we apply author attributes to the start and end objects with author type following a certain meta path.

Given a meta path P, the symmetric random walk with author attributes is defined as: $SRW_{AA_R}(x, y) = RW_R(x, y) * AA_x + RW_{R_{-1}}(y, x) * AA_y$, where AA_x and AA_y represent the attribute of the start and end objects with author type, and are calculated from the norm of the attribute tri-vector above.

For each meta path, we apply any topological measure on it, and acquire a unique topological feature, which composes the topological feature sets. Then we use hybrid topological features in the set to predict the collaboration relationship.

3.2. Collaboration Prediction Model

We introduce the collaboration relation prediction model which models the probability of future co-authorship between pair authors as a function of the topological features between them. We extract the topological features from the training author pairs, and build prediction model to obtain the best coefficients associated with each feature.

The problem of collaboration prediction can be considered as a binary classification model, which contains future collaboration or not. We choose the standard binary classifier, the logistic regression with L_2 regularization as the prediction model, and employ gradient descent algorithm as the optimization method to obtain the best weight of each feature, in order to maximize the likelihood of collaboration formation.

4. Experiments

In this section, we show that our proposed collaboration prediction model (*MACP*) can improve the co-authorship

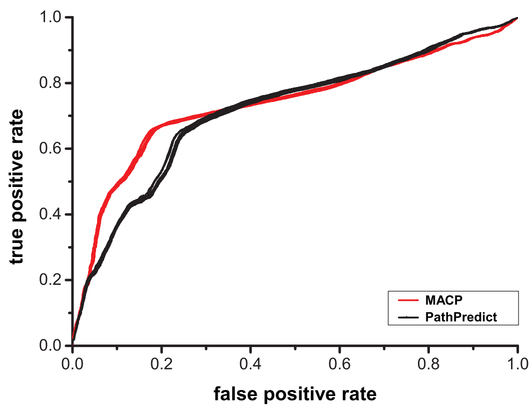

Figure 2: ROC Curve

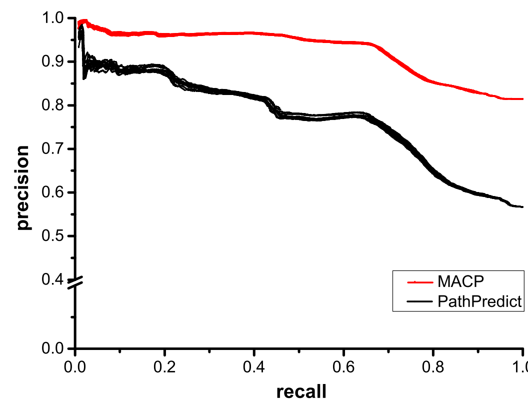

Figure 3: Precision-recall Curve

prediction accuracy, in comparison with other two models, one of which uses the same topological feature space in Support Vector Machine (SVM) classification algorithm (*MACP-SVM*), and the other employs basic structural features in the same logistic regression algorithm, called *PathPredict* proposed in [5].

The APS heterogeneous bibliographic information network is used for experiments. We choose two journals with low impact factor, *PRA (2.991 in 2013)* and *PRB (3.664 in 2013)*, and one journal with high impact factor, *PRL (7.728 in 2013)*. Two time intervals are considered for the network, according to the publication year of each paper: $T_0 = [2000, 2006]$, $T_1 = [2007, 2013]$. We use T_0 as past time interval, and T_1 as future time interval. In the training stage, we find author pairs as positive samples, which have no relationship in the past time interval and build collaboration in the future time interval. We also choose an equal size of negative pairs, to balance the size of positive and negative samples. The source authors are comprised of the authors who have published more than 10 papers in the past time interval. We confine the target authors that are relatively close to the source authors, to avoid the excessive computing between the unrelated authors.

TABLE 1: Comparison of different prediction models

Model	Accuracy	Precision	Recall	F1-socre	AUC
MACP	0.743	0.845	0.704	0.768	0.777
MACP-SVM	0.692	0.777	0.640	0.702	0.717
PathPredict	0.705	0.787	0.657	0.717	0.732

In order to measure the prediction accuracy, we use tenfold stratified cross-validation to assess the quality of each method. Common metrics are used to evaluate the prediction result, which include accuracy , precision, recall, f1 score, the precision-recall curve, receiver operating characteristic curve (ROC curve), and the area under ROC curve (AUC). We use the meta paths with length constraint 5 and different measures on each meta path to conduct the experiments. We choose two contrastive methods, which are SVM learning method with the same topological features as our *MACP* model, and the *PathPredict* model with PC, NPC, RW and SRW features described in 3.1.1.

The numerical metrics of different prediction models are shown in Table 1, which shows that our proposed model obtain the best prediction quality. The *MACP-SVM* model has quite lower prediction quality due to the extremely slow convergence speed and limiting iterations compared to the logistic regression algorithm, which is fast converged in the experiments. The ROC curve and precision-recall curve of the *MACP* and *PathPredict* model are displayed in Fig. 2, Fig.3, which exhibits *MACP* model beats the *PathPredict* models in average, because the topological feature space in *MACP* model is the combination of the structural features and the attached attributes of the objects and links in the networks, whereas the features in *PathPredict* model is just the basic structural features. Different feature spaces lead to the distinguished prediction precision. It turns out that considering the attached information of objects and links is helpful to improve the prediction quality, such as the transitive similarity, temporal dynamics and author attributes. And the author attributes such as influence, activity and collaboration tendency have an effect on the collaboration relationship formation to some degree. These features can help to improve prediction accuracy as supplements to meta path-based feature space.

5. Conclusion and Future Work

In this paper, we study the problem of collaboration relationship prediction in heterogeneous information networks.The presented MACP model is utilized to solve this problem, which defines the meta path and author attribute based topological features, and builds logistic regression-based prediction model. Experiments on APS dataset show that the prediction accuracy has been improved by considering the structural topology and author attributes together. In the future, we attempt to combine the authors' relationships in both academic and social networks to analyze more extensive collaboration relationships between authors.

References

[1] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *Proceedings of Advances in neural information processing systems*, 2003.

[2] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[3] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 243–252.

[4] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," in *Proceedings of the VLDB Endowment*, 2011, pp. 992–1003.

[5] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2011, pp. 121–128.

[6] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: relationship prediction in heterogeneous information networks," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 663–672.

[7] J. He, J. Bailey, and R. Zhang, "Exploiting transitive similarity and temporal dynamics for similarity search in heterogeneous information networks," in *Proceedings of Database Systems for Advanced Applications*, 2014, pp. 141–155.

[8] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu, "Relevance search in heterogeneous networks," in *Proceedings of the 15th international conference on extending database technology*, 2012, pp. 180–191.

[9] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks." in *SDM*, 2012, pp. 1119–1130.

[10] C. Luo, R. Guan, Z. Wang, and C. Lin, "Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks," in *Advances in Information Retrieval*, 2014, pp. 210–221.

[11] T. Huynh, A. Takasu, T. Masada, and K. Hoang, "Collaborator recommendation for isolated researchers," in *Proceedings of IEEE International Conference on Advanced Information Networking and Applications (AINA)*, 2014, pp. 639–644.