Author Set Identification via Quasi-Clique Discovery

Yuyan Zheng
Beijing University of Posts and Telecommunications
Beijing, China
zyy0716_source@163.com

Chuan Shi*
Beijing University of Posts and Telecommunications
Beijing, China
shichuan@bupt.edu.cn

Xiangnan Kong
Worcester Polytechnic Institute
MA, USA
xkong@wpi.edu

Yanfang Ye
Case Western Reserve University
OH, USA
yanfang.ye@case.edu

ABSTRACT

Author identification based on heterogeneous bibliographic networks, which is to identify potential authors given an anonymous paper, has been studied in recent years. However, most of the existing works merely consider the relationship between authors and anonymous papers, while ignore the relationships between authors. In this paper, we take the relationships among authors into consideration to study the problem of author set identification, which is to identify an author set rather than an individual author related to an anonymous paper. The proposed problem has important applications to new collaborator discovery and group building. We propose a novel Author Set Identification approach, namely ASI. ASI first extracts a task-guided embedding to learn the low-dimensional representations of nodes in bibliographic network. And then ASI leverages the learned embedding to construct a weighted paper-ego-network, which contains anonymous paper and candidate authors. Finally, converting the optimal author set identification to the quasi-clique discovery in the constructed network, ASI utilizes a local-search heuristic mechanism under the guidance of the devised density function to find the optimal quasi-clique. Extensive experiments on bibliographic networks demonstrate that ASI outperforms the state-of-art baselines in author set identification.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Networks** → *Network structure*; • **Theory of computation** → *Social networks.*

KEYWORDS

Author set identification; Quasi-clique; Heterogeneous network construction; Network embedding; Meta path

*The corresponding author.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357966

ACM Reference Format:
Yuyan Zheng, Chuan Shi, Xiangnan Kong, and Yanfang Ye. 2019. Author Set Identification via Quasi-Clique Discovery. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM'19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357966

1 INTRODUCTION

Heterogeneous information network (HIN) has received increasing attention in the past decade. As a powerful and effective modeling method to fuse complex information, HIN is successfully applied to many data mining tasks [28]. Heterogeneous bibliographic network [30] is a typical example of HIN and it has also received more and more attention in recent years. Various kinds of mining tasks have been studied in bibliographic networks, including relevance search [16, 27], personalized recommendation [19, 25] and so on.

As an important task, the problem of author identification has been extensively studied, which aims to learn a model to rank potential authors for an anonymous paper based on public information. This problem was first exploited in [13] and has been further studied in recent works [6, 22, 40]. These studies can be roughly categorized into two types including supervised learning methods and network embedding methods. The former employs the feature-engineering to predict the correlation between the paper and author, while the latter mainly leverages network structure or semantic content of the paper to learn node representations that can be further used to infer authors.

Although conventional methods for author identification are effective and useful in some real applications, they usually ignore relationship among authors. However, in many scenarios such as finding a potential author group for a given paper, the relationship among authors are very significant. Therefore, we propose to study a new problem called author set identification. We illustrate the problem setting in Fig. 1, in which the heterogeneous bibliographic network and network schema are given as the input, the goal is to learn a model that can identify the optimal author set for a new anonymous paper. Furthermore, Fig. 2 illustrates the difference between the author set identification problem and traditional author identification problem. The problem of author set identification is to acquire an author set related to the anonymous paper, as well as having strong relationship between authors, while the traditional problem only get an author ranking for the anonymous paper.

A basic idea for the problem is to find a set of closely connected authors that are related to an anonymous paper. Therefore, we need

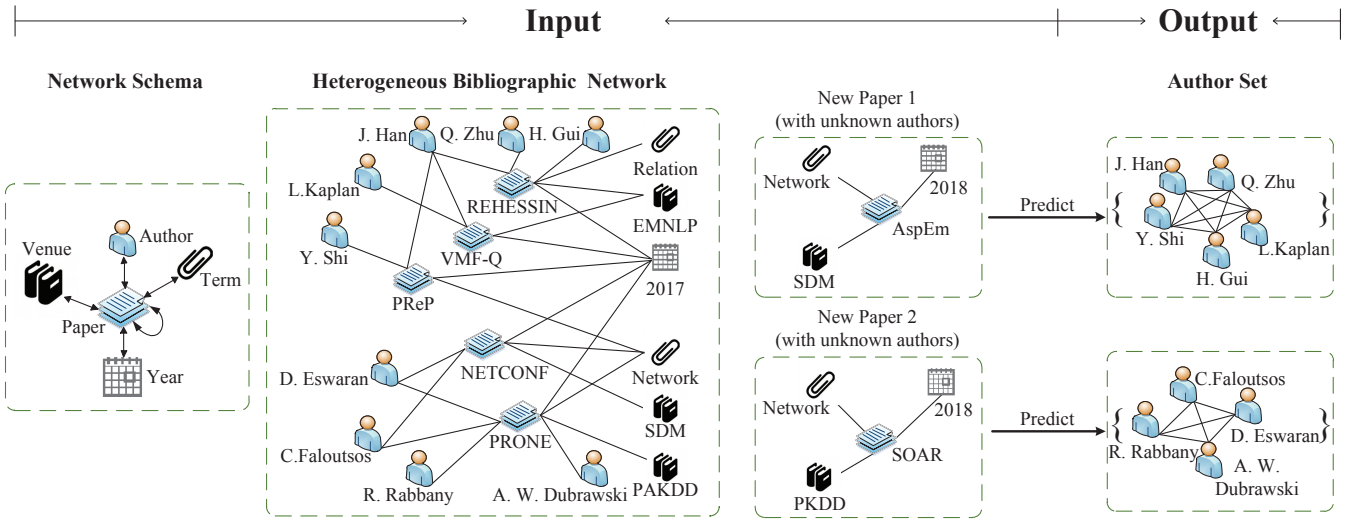


Figure 1: The problem of author set identification in heterogeneous bibliographic networks.

consider the relationship between anonymous paper and authors, as well as relationship between authors. However, it is not a trivial task to take both relationships into account at the same time, which faces two challenges. (1) How can we focus on interactions between anonymous paper and authors, meanwhile preserving rich inherent structural information in heterogeneous bibliographic network. In order to reduce the complexity, we only need to focus on interactions between anonymous paper and authors for our problem settings, while it will lose rich interactions in original network and indirect connections of papers and authors. An effective strategy is desired to solve the dilemma. (2) How can we find an optimal set of closely connected authors that are related to the anonymous paper. It is a NP-hard problem to find an optimal author subset that is related to the anonymous paper. An effective approximate method is needed to solve the NP-hard problem.

In this paper, we propose a novel author set identification approach called ASI. Specifically, in order to focus on interactions between papers and authors and preserve rich structural information in original heterogeneous bibliographic network, ASI first constructs a weighted paper-ego-network, which only contains the anonymous paper and authors and corresponding relations (paper-author and author-author). When construct the network, ASI proposes a task-guided embedding method called TaskGE to learn the low-dimensional representations of nodes, and then applies these embeddings to determine the weights of edges in the constructed network. Thus, the constructed network only contains the anonymous paper and authors, and preserves rich structure information through embeddings learned from the original network. Furthermore, we introduce the concept of quasi-clique in dense subgraph and convert the optimal author set identification to the quasi-clique discovery in the weighted paper-ego-network. We design the local-search heuristic method under the guidance of a novel density function to find the optimal quasi-clique (author set). We summarize our contributions as follows.

We propose to study a problem of author set identification, which could bring crucial implications to many applications, such as reviewer recommendation or new collaborators discovery.

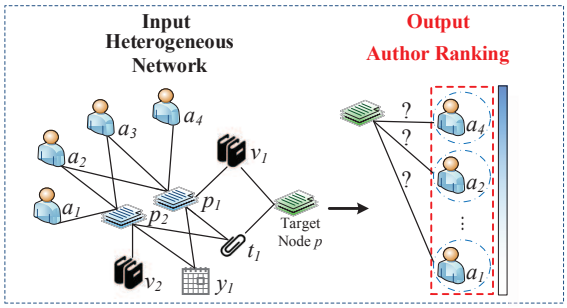
We propose a novel method ASI for this problem. ASI first constructs a weighted paper-ego-network, in which the weights of edges are determined by the proposed embedding method TaskGE. Then we introduce the quasi-clique and propose an approach of local-search heuristic under the guidance of the designed novel density function to find the optimal author set in constructed network.

We conduct extensive experiments on bibliography network to evaluate the performance of ASI. The results demonstrate the superiority of ASI by comparing with the state-of-art baselines. What's more, ASI can also automatically determine the number of authors for a given paper.

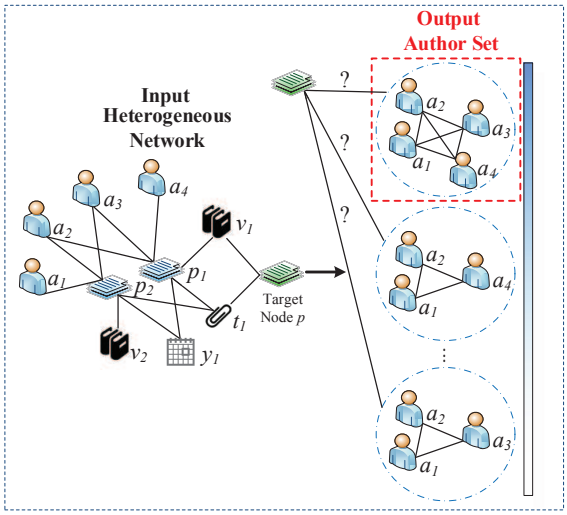
2 RELATED WORK

Heterogeneous bibliographic network mining has attracted large amounts of attention in recent years. Many of these works are devoted to the research of mining problem, such as collaborator recommendation [18, 34] and author identification [6, 40]. Although the problem of author identification has been studied in [6, 22], they only focus on the identification of individual author and neglect the relationship between authors. We consider these two aspects and propose to study the problem of author set identification in bibliographic network. We not only discover the possible author of a new paper but also find a set of collaborator so as to better accomplish a paper. Meanwhile these researchers also may be the potential future collaborator.

The proposed method ASI for the problem includes the following two phases of paper-ego-network construction and optimal quasi-clique discovery. In the first phase, we employ the embedding to construct the network. As we have known, network embedding is an effective method to learn the low-dimensional representation of network and has recently been a hot research topic. Motivated by



(a) Author identification problem [5]



(b) Author set identification problem (this paper)

Figure 2: The illustration of the comparison between our problem and author identification problem.

language modelling [21], some random walk-based methods, like DeepWalk [24] and node2vec [12], have been proposed to learn the representation of nodes in network. Besides, LINE [33] is presented as an efficient network embedding method. There are also some deep neural network based models for network embedding [37]. After that, in order to tackle network heterogeneity, many researchers make many attempts. metapath2vec [9] leverages meta paths to guide the random walk process to learn network representations. PTE [32] partitions an HIN into multiple bipartite graphs and performs network embedding individually. HERec [26] designs a meta-path based random walk strategy to generate meaningful node sequences for HIN embedding. RHINE [20] considers the structural characteristics of heterogeneous relations and proposes a novel relation structure-aware HIN embedding model. Wang et al. [38] make the first effort toward HIN embedding in hyperbolic spaces. However, most of these methods are general-purpose embedding that are independent of tasks. We design an embedding method relying on specific-task to learn the low-dimensional vectors, which can be further used to construct the weighted paper-ego-network.

In constructed network, we introduce the quasi-clique to find the optimal author set, which is a significant concept in dense subgraph discovery. As is known to all, a great deal of subgraph mining approaches have been investigated, such as dense subgraph mining [5, 10], quasi-cliques mining [23, 36], k-plexes mining [2, 8]. Dense subgraph discovery has also been studied from a wide variety of perspectives. By solving a parametric maximum-flow problem, the problem of densest subgraph can be solved in polynomial time [11]. Asashiro et al. [1] propose a greedy algorithm with 1/2-approximation in linear time. After that, there emerge many researches about variants of the densest subgraph like DkS, which discover a densest subgraph of k vertices [3]. Also, there are many works focusing on alternative density functions of dense subgraph. For instance, Tsourakakis et al. [36] present a general framework of density function based on the concept of quasi-clique. Sozio et al. [29] are concentrated in the monotone constraints of minimum degree density. In addition, dense subgraph has also shown its good performance in various applications such as fraud detection [14] and community detection [5]. To the best of our knowledge, however, it is the first attempt to employ dense subgraph to address the problem of author set identification.

The prototypical dense graph is the clique (nodes in clique all connected to each other), but, discovering the largest clique is inapproximable and clique is in practice too strict to miss a single edge in an otherwise dense subgraph. Extracting the densest subgraph tends to favor large subgraphs with small edge density and large diameter. Hence, we select the method of extracting quasi-clique to find optimal author set, which can discover subgraph of much higher quality than densest subgraph. Also, this method is very suitable for our scenario, just as it can find compact, dense and smaller diameter's subgraph, which is a desirable property in our task.

3 PRELIMINARY

In this section, we introduce some basic concepts and formalize the problem of author set identification in the bibliographic network.

Definition 3.1. **Heterogeneous Information Network (HIN)** [31]. An HIN is defined as a directed graph $G = {}^{1}V; E^{o}$, where V and E denote node set and edge set, respectively. There exists a node type mapping function $\varphi : V \; ! \; A$ and an edge type mapping function $\psi : E \; ! \; R$, where A and R are node type set and edge type set, respectively, and $jAj + jRj > 2$.

The bibliographic network can be seen as an HIN. Fig. 1 illustrates an example of bibliographic network and its corresponding network schema, which is a meta template for an HIN and illustrates the node types and their interaction relations. We can see that the bibliographic network contains five types of nodes, that is, author (A), paper (P), venue (V), term (T) and year (Y), and multiple semantic relations (e.g., writing relations between authors and papers, published-by relations between papers and venues, and citation relations between papers).

In HIN, two nodes can be connected via different semantic paths and the physical meaning of different paths is distinct from one another. These paths can be defined as meta-paths.

Definition 3.2. **Meta Path** [31]. Meta path is a path in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$ (abbreviated as $A_1 A_2 \cdots A_{l+1}$), where A_i denotes node type and R_i means edge type. It is a sequence of node types and edge types between nodes, which describes a compositional relation between two given node types.

In this paper, we study the novel problem of author set identification in bibliographic network, which can be defined as follows.

Definition 3.3. **Author Set Identification Problem**. Given a bibliographic network $G = (V, E)$, which includes a set of papers and papers' relevant information (i.e., authors, venues, terms and year), the goal is to design a method to acquire an author set S_A^* from C_A for a new anonymous paper p, such that S_A^* is the optimal set to collaborate on the paper p among all subsets of C_A, where $C_A = \{a_1, a_2, \ldots, a_m\}$ denotes the set of all candidate authors.

In order to find the optimal author set, we introduce the concept of quasi-clique, which can be defined as follows.

Definition 3.4. **Quasi-Clique** [36]. A set of nodes S is an α-quasi-clique if $e(S) \geq \alpha \binom{|S|}{2}$, i.e., if the edge density of the subgraph induced by S exceeds a threshold parameter $\alpha \in [0, 1]$. The edge density is defined as $e(S)/\binom{|S|}{2}$, where $e(S)$ is the size of edges in the subgraph induced by S.

4 THE PROPOSED METHOD

In this section, we present the proposed method that leverages quasi-clique for *A*uthor *S*et *I*dentification, called **ASI**. The overall architecture of ASI is shown in Fig. 3. Given a heterogeneous bibliographic network and an anonymous paper (Fig. 3(a)), we first construct a weighted paper-ego-network for each anonymous paper (Fig. 3(b)), and then find the optimal quasi-clique with constraint (OQCC) in the weighted paper-ego-network (Fig. 3(c)). In the following, we will clarify the basic idea and specific details about these two phases.

We aim to find a set of closely connected authors that are related to the anonymous paper. However, how can we pay attention to interactions between anonymous paper and authors, meanwhile preserving rich inherent structural information in heterogeneous bibliographic network. We consider to construct a weighted paper-ego-network only containing the anonymous paper and authors. In order to preserve rich structural information in bibliographic network, we propose the task-guided embedding method to learn vector representations of nodes, which can be further used to determine the weights of edges through proper distance function for constructing the weighted paper-ego-network. Then we transform the author set identification into the problem of quasi-clique extraction with constraint. Finally, We propose an approach of local-search heuristic under the guidance of designed novel density function, so as to discover the optimal quasi-clique in the constructed network.

4.1 Weighted Paper-Ego-Network Construction

In order to reduce the complexity and merely focus on interactions between anonymous paper and authors, meanwhile preserve rich inherent structural information in heterogeneous bibliographic network. We just need to focus on two kinds of relationships, including relationship between the anonymous paper and author, as well as relationship between authors. Therefore, we consider to construct a weighted paper-ego-network, which only contains anonymous paper p and candidate authors. The key of constructing the network is to determine the weights of edges between anonymous paper and authors, and edges between authors. For edges between authors, we propose the task-guided embedding (TaskGE) to learn the low-dimensional representations of nodes, which preserves rich structure information in original network and can be further used to determine the weight of edges between authors. Since the feature representation for the anonymous paper is unknown, we first employ the weighted combination of feature vectors of its observed neighbors in the network to calculate its vector. Then we can easily determine the edges between anonymous paper and authors based on the computed representation of the anonymous paper and the vectors of authors obtained in TaskGE.

Specifically, for each anonymous paper p, we denote the constructed weighted paper-ego-network by $G_p = (V, E, W)$, where V is a set of nodes , E is a set of edges and W is a set of the weight on each edge. V includes two types of nodes, that is, anonymous paper p and candidate authors. Correspondingly, E contains two types of edges, namely, the edge between paper p and any candidate author a, and the edge between any two candidate authors a_1, a_2. We denote the weights of these two types of edges by w_{pa} and $w_{a_1 a_2}$, respectively. Different from existing general-purpose embedding, our embedding method is totally dependent of specific-task. We exploit two unique characteristics or significant aspects of author set identification task. One is the proximity between anonymous paper and authors, we model it as paper-author-aware embedding. The other is the strong relationship between authors, we model it as author-author-aware embedding.

4.1.1 Paper-Author-Aware Embedding. Intuitively, for a given paper p, the relevance score of p and any one a of its true authors should be ξ larger than that of p and other author a' who is not the author of p. If not, a loss penalty will incur. Here, we employ the hinge loss [40] to define a general function to model the relationship between paper and author as follows:

$$L_{R_{PVA}} = \sum_{r \in P_{PVA}} E_{<p,a,a'>} [\xi + f(p,a) - f(p,a')]_+, \quad (1)$$

where $[x]_+ = max(x, 0)$ is the standard hinge loss, ξ is the safety margin size [4]. $< p, a, a' >$ denotes the triples <paper, positive author, negative author>. r and P_{PVA} denote any meta path and the set of meta paths between paper and author, respectively. Generally, we can add any proper meta paths between paper and author to P_{PVA} for leveraging multiple information. Actually, there exist multiple indirect relations besides the direct relation between paper and author. For example, $P_{PVA} = \{PA, PTPA\}$ means we not only consider the direct author but also take the potential authors into account. Correspondingly, $P_{PVA} = \{PA\}$ means we only consider the direct author of paper. $f(p,a)$ stands for the metric between paper p and author a. As demonstrated by CML [15], distance metric [39] satisfies better triangle inequality and transition property than inner-product, we use the euclidean distance to define the metric:

$$f(p,a) = \|X_p - X_a\|_2^2, \quad (2)$$

where X_p and X_a are the embedding vectors of p and a, respectively. For a new anonymous paper, we adopt similar approach to Chen et

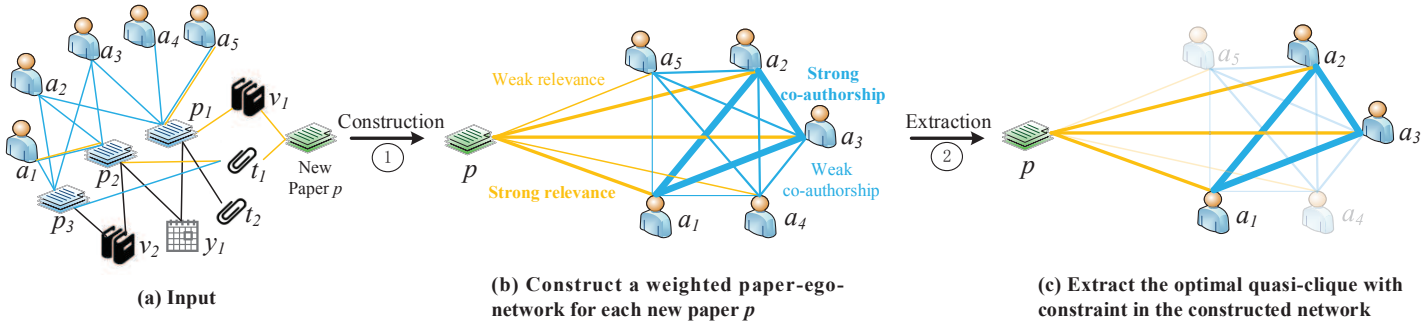


Figure 3: The overall architecture of proposed method ASI (weighted paper-ego-network construction and optimal quasi-clique with constraint extraction)

al. [6] to calculate its vector representation. That is, the embedding of a paper is represented as the weighted combination of the vectors of observed different types of neighbors in the network as follows:

$$X_p = \sum_{t=1}^{\tilde{O}} w_t X_p^t; \quad (3)$$

where n is the number of neighbors' types of paper p, X_p^t is the mean of vectors of the t-th node type, $X_p^t = \frac{1}{|N_p^t|} \sum_{i \in N_p^t} X_i$, N_p^t denotes the set of nodes of the t-th type. In this paper, we do not employ the reference type of nodes due to the lack of citation data.

4.1.2 Author-Author-Aware Embedding. L_{RPvA} models the relationship between paper and author, in this subsection, we will consider how to model the relationship between authors. It is reasonable that there should be strong relationship between co-authors. In other words, the relevance score between co-authors should be larger than that of authors who have never collaborated with each other. Similarly, there might exist some potential co-authorship between authors implicitly indicated by meta paths like $APTPA$. Therefore, we also define a general function to formulate the triple relation $< a, a^+, a >$.

$$L_{RAvA} = \sum_{r \in P_{AvA}}^{\tilde{O}} E(a, a^+, a^0) \cdot \xi + f(a, a^+) - f(a, a^0) + \gamma; \quad (4)$$

where a^+ means any co-author of a, a denotes any author who has never cooperated with a. f is the metric function which has been introduced in subsection 4.1.1. r denotes any meta path between authors. P_{AvA} is the set of meta path between authors. $P_{AvA} = \{APA\}$ means we only consider existing co-authors.

4.1.3 Regularization. Recently, Cogswell et al. [7] propose a new regularization technique called covariance regularization, which is initially used to reduce the correlation between activations in a deep neural network. Afterwards, Hsieh et al. [15] find that it is useful in de-correlating the dimensions. As covariances can be seen as a measure of linear redundancy between dimensions, this loss of covariance regularization essentially tries to prevent each dimension from being redundant. Therefore, we employ loss of covariance regularization as follows:

$$L_{reg} = \frac{1}{N} \|C\|_f - \|diag(C)\|_2^2; \quad (5)$$

where $\|\cdot\|_f$ is the Froeninus norm, C is covariance matrix between all pairs of dimensions i and j, $C_{ij} = \frac{1}{N} \sum_{k=1}^{N} (X_k^i - u_i)(X_k^j - u_j)$, $u_i = \frac{1}{N} \sum_{k=1}^{N} X_k^i$, X_k^i denotes the i-th dimension of embedding vector of node k.

Finally, we combine three parts above to get the unified objective function for task-guided embedding as follows:

$$L = L_{RPvA} + \gamma L_{RAvA} + \lambda L_{reg}; \quad (6)$$

where L_{reg} is the regularization term for avoiding over-fitting, λ controls penalty of regularization, γ is a harmonic factor to balance two components. In this paper, we only consider the direct relation PA in L_{RPvA} and APA in L_{RAvA}.

To minimize L, we design a sampling based mini-batch Adam optimizer [17]. To get the training triples $< p, a, a^0 >$ and $< a, a^+, a >$, we draw positive samples according to the proportion of path instances of different meta paths. This sampling strategy can avoid the problem of under-sampling for relations with a large number of links or over-sampling for those with a small number of links. For each sampled positive example $< p, a >$, we first fix vertex p and the corresponding relation. Then we randomly generate negative vertex a^0 which has not the same relation with p to construct training triples $< p, a, a^0 >$. Similarly, we can fix a and corresponding relation to acquire training triples $< a, a^+, a >$.

Given the low-dimension representation learned above, we can easily calculate w_{pa} and $w_{a_1 a_2}$ using distance function such as cosine.

4.2 Optimal Quasi-Clique with Constraint Extraction in Weighted Paper-Ego-Network

For each new paper p, we construct a weighted paper-ego-network $G_p = (V, E, W)$. In order to find the optimal author set for the given paper p in G_p, we propose a new method called OQCCE which is an adaptation of the local-search heuristic by Tsourakakis et al. [36]. The algorithm selects p as initial set. Then under the guidance of designed novel density function, algorithm iterates two phases of adding or removing the designated nodes until the quasi-clique with maximum density function is discovered. What's more, the novel density function considers two kinds of heterogeneous relationships, including the close relationship between the anonymous paper and author, as well as relationship between authors.

In specific, we regard the node p as constraint, which means that the extracted subgraph must contain node p. In [36], there is only one type of edge. However, there exist two types of edges in weighted paper-ego-network. The simplest method is to assign equal significance to two types of edges. In fact, the importance may vary. Therefore, we introduce a variable β to adjust the importance of two types of edges. Meanwhile, we also adapt the density function to accommodate the weighted network. Accordingly, the proposed novel density function can be defined as follows:

$$\partial_{\alpha;\beta}(S) = \beta \sum_{(i,j)\in D_{PA}} w_{ij} + \sum_{(k,l)\in D_{AA}} w_{kl} - \alpha \binom{|S|}{2}; \quad (7)$$

where S represents a subset of vertices of network G_p having $S \subseteq V$, $|S|$ denotes the number of nodes in the subgraph induced by S, w_{ij} is the weight of edge between nodes i and j in the subgraph induced by S. D_{PA} represents the set of edges between given paper p and candidate authors in the subgraph induced by S. Likewise, D_{AA} represents the set of edges between authors in the subgraph induced by S. β controls the importance of paper-author edge. α is a constant. The first two parts in Eq. 7 favors subgraphs with abundant edges while the third part penalizes large subgraphs.

Based on the proposed density function above, next we will describe how to find the optimal quasi-clique with constraint in G_p. The outline of our algorithm, OQCCE, is shown as Algorithm 1.

Algorithm 1: OQCCE

Input : Weighted paper ego network $G_p = (V, E, W)$; maximum number of iterations I_{max}; the constrained node p

Output : A subset of nodes $S \subseteq V$ and $p \in S$

```
S ← {p}, b1 ← TRUE, i ← 1;
while b1 and i ≤ Imax do
    b2 ← TRUE;
    while b2 do
        if there exists u ∈ V \ S and ∂α;β(S ∪ {u}) ≥ ∂α;β(S) then
            S ← S ∪ {u};
        else
            b2 ← FALSE;
    if there exists u ∈ S and u ≠ p and ∂α;β(S \ {u}) ≥ ∂α;β(S) then
        S ← S \ {u};
    else
        b1 ← FALSE;
    i ← i + 1;
```

The algorithm firstly selects constrained node p as the initial set. Then it traverses all nodes one by one and adds u to S if $\partial_{\alpha;\beta}(S \cup \{u\})$ improves. Afterward, the algorithm traverses every vertex v in S and remove v if $\partial_{\alpha;\beta}(S \setminus \{v\})$ enhances. Note that we cannot remove constrained node p during the period of removal. The algorithm repeats these two phases of addition and removal until an optimum is reached or the number of iterations exceeds I_{max}.

Table 1: Statistics of the two datasets.

Dataset	# papers	# authors	# terms	# venues
AMiner-I	8821	12660	12467	5
AMiner-II	35349	36247	31446	14

Table 2: The features extracted for supervised learning methods.

No.	Feature description
1	Total number of papers
2	Number of different venues
3	Number of different years
4	Number of references the author cited before
5	Ratio of references the author cited before
6	Number of author's citations in the references
7	Ratio of author's citations in the references
8	Number of references written by the author
9	Ratio of references written by the author
10	Ratio of author's papers in the references
11	Number of shared word
12	Ratio of shared words
13	Whether the author attend the venue before
14	Number of times the author attend the venue before
15	Ratio of times the author attend the venue before
16	Number of papers author published in the last 3 years
17	Ratio of papers author published in the last 3 years

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 Datasets. AMiner [35] is a classical academic network, which contains millions of author and paper information from major computer science venues for more than 50 years. We utilize part of the AMiner dataset[1] from 1954 to 2015. Specifically, we extract two subsets with different scale, denoted by AMiner-I and AMiner-II. AMiner-I is a small subset data of some important venues in data mining area, which includes 5 venues, namely KDD, ICDM, SDM, CIKM and PKDD. AMiner-II is a large subset of four areas, including Artificial Intelligence (AI), Data Mining (DM), Databases (DB), and Information System (IS). For each area, we choose some important venues[2] which have influential publications. The detailed descriptions of the two datasets are shown in Table 1.

5.1.2 Baselines. In order to examine the effectiveness of our approach, we compare against the following three kinds of representative methods:

Similarity measure. We design two kinds of similarity measure methods based on meta paths $PTPA$ and $PCPA$, which can indirectly connect the new paper and candidate authors with term or venue. Then we rank candidate authors according to the similarity scores between candidate authors and the new paper. Here, we adopt the number of path instances as the similarity score.

Feature method. Following the work of Chen et al. [6] , we extracted 17 features for each paper-author pair shown in Table 2. We choose LR, SVM and Bayesian as learning algorithms.

HetNetE. HetNetE is a recent model proposed in [6] for author identification problem. It first learns the low-dimensional feature vectors of nodes using the task-specific and general network

[1] https://www.aminer.cn/citation

[2] AI: ICML, AAAI, IJCAI, NIPS. DM: KDD, WSDM, ICDM, PKDD. DB: SIGMOD, VLDB, ICDE. IS: SIGIR, CIKM.

embedding method. Then it predicts author of the given paper with the learned embedding vectors.

5.1.3 Parameter Settings. For our method ASI, we set the embedding dimension d to 128, the size of negative samples to 2, the margin ξ to 2, the learning rate to 0.00001, the batch size to 200, the regularization penalty λ to 10, the trade-off factor γ to 1.0, α to 0.01, β to 0.1. For HetNetE and Feature method, we choose the optimal parameter. Three meta paths APC, APW, APP are jointly used in HetNetE. In addition, for fairness comparison, we do not adopt the reference types of nodes when computing the embedding vectors of papers due to the lack of most citations in HetNetE and ASI.

5.1.4 Evaluation Metrics. We adopt $Precision$ (P), $Recall$ (R), $F1$ score, $Jaccard$ index (J), MAP (mean average precision) and $RMSE$ as evaluation metrics.

P. It reflects the accuracy of returned author set, which can be defined as the ratio of the true authors in the returned author set. $P = \frac{|S_A^0 \setminus S_A|}{|S_A^0|}$, where S_A^0 denotes the returned author set or the returned top-k author set in $P@k$. S_A means the true author set.
R. It shows the ratio of returned true authors in the whole true author set. It can be computed as follows: $R = \frac{|S_A^0 \setminus S_A|}{|S_A|}$, where S_A^0 and S_A have the same meanings introduced above.
$F1$. It is the harmony average of P and R, which is defined as: $F1 = \frac{2PR}{P+R}$.
$Jaccard$ index. It measures similarity between two sets and is formulated as: $J = \frac{|S_A^0 \setminus S_A|}{|S_A^0 \cup S_A|}$, which means the ratio of the intersection and the union of two sets.
MAP. It is computed as mean of AP at different k for a paper. $AP = \frac{\sum_{i=1}^{k} p@i \cdot rel_i}{\#\text{ of correct author}}$, where rel_i equals 1 if the result at rank i is correct author and 0 otherwise.
$RMSE$. It is a measure of difference between the number of authors returned by model and the number of true authors. $RMSE = \sqrt{\frac{1}{|m|}(|S_A^0| - |S_A|)^2}$, where m is the number of test papers, S_A^0 and S_A are the number of returned author and true author, respectively.

5.2 Comparisons and Analysis

To evaluate the performance, we regard papers published before 2014 as training set and papers published in 2014 and 2015 as test set. Since it is time consuming to rank all candidate authors for each anonymous paper in the evaluation procedure, following the strategy in [6], for each paper in the test set, we randomly sample some negative authors and obtain 100 candidate authors in all. Then, we rank the 100 candidate authors consisting of the positive and sampled negative authors for each paper. For our method ASI, we also select the same 100 candidate authors to construct the weighted paper ego network for each test paper. The final results are averaged over all the test papers for each evaluation metric.

We report the results of performance comparison in tables 3, 4. As we can observe, (1) Our method ASI achieves better performance than all baselines on all measures except R and MAP. It improves the performance by more than 15% on P, J and $F1$ averagely. Although ASI does not achieve the best performance on R, it is also near the best value. (2) ASI can automatically confirm the appropriate number of authors for a given paper, which can be clearly demonstrated by the lowest value on metric $RMSE$. In a word, ASI not only can discover a set of authors with strong relationship but also can determine the proper number of authors for an anonymous paper. (3) To our surprise, the similarity measure method based on PTPA has very good performance, which indicates that the term has a significant role in finding author set for a given paper.

5.3 Parameter Sensitivity

In this section, we conduct experiments to investigate the sensitivity of different parameters in our method ASI, i.e., the embedding dimension d, training times l and control factor β in density function. We investigate how a specific parameter influence the performance of ASI by changing its value and fixing the others. The result is shown in Figure 4.

From the result, we can observe that P, R, J, $F1$ and MAP increase firstly and then slightly decrease with the increment of the dimension d. The trend of $RMSE$ is the opposite. When d is around 128, ASI achieves the best performance. The performance becomes stable when the training times l is above 50. The optimal value of β is around 0.1.

6 CASE STUDY

We present the case study to show the performance difference between ASI and three selected baselines, i.e., PTPA (best similarity-based method), SVM (best feature-based method) and HetNetE. Table 5 lists the top 10 ranked authors for two papers, which are published in KDD 2014 and CIKM 2014, respectively.

From table 5, we can see that the PTPA method returns some authors who have similar research themes with the given paper. Although many of them are not the correct authors, they still can be considered as the potential authors. HetNetE achieves relatively better performance than other baselines. ASI can automatically find an author set for a new paper, which does not require to specify the number of authors. This number is usually unknown a prior and is difficult to estimate. All in all, ASI not only has better effectiveness than other methods but also has the ability of determining number of authors for a given paper.

7 CONCLUSION

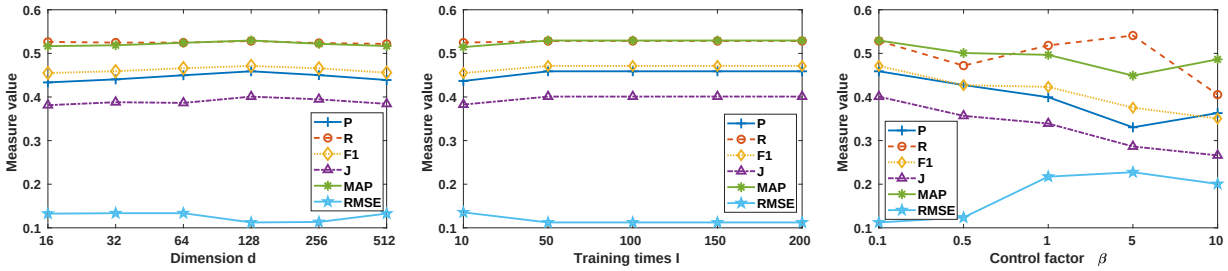
In this paper, we propose to study the problem of author set identification in heterogeneous bibliography network. A novel approach called ASI is presented to solve the problem. ASI includes two phases of weighted paper-ego-network construction and OQCC extraction. In the first phase, in order to determine the weights of edges in constructed network, we present the task-guided embedding (TaskGE) to learn the vector representations of nodes. TaskGE not only is totally task-specific but also takes full advantage of inherent structural information of bibliography network. In the second phase, we introduce the concept of quasi-clique and propose a local-search heuristic approach under the guidance of designed novel density function so as to find the optimal author set in weighted paper-ego-network. The experiments on academic dataset demonstrate that ASI outperforms state-of-the-art baselines and can automatically determine the number of authors. As future

Table 3: Results of effectiveness experiments on AMiner-I. We use bold to mark the best performance for each comparison. " indicates higher is better, # indicates lower is better. "Avg." means the average rank of different methods.

Methods			Evaluation						Avg.
			P (")	R (")	J (")	F1 (")	MAP (")	RMSE (#)	
Top-5	Similarity measure	PTPA	0.2716 (2)	0.5007 (7)	0.2310 (2)	0.3356 (2)	**0.6109** (1)	0.1714 (2)	2.67
		PCPA	0.2098 (7)	0.3937 (11)	0.1680 (7)	0.2614 (7)	0.4718 (9)	0.1714 (2)	7.16
	Feature method	LR	0.2160 (5)	0.3915 (12)	0.1827 (6)	0.2657 (4)	0.4834 (7)	0.1714 (2)	6.00
		SVM	0.2493 (3)	0.4562 (9)	0.2154 (4)	0.3081 (3)	0.5451 (3)	0.1714 (2)	4.00
		Bayesian	0.2209 (4)	0.4075 (10)	0.1888 (5)	0.2733 (5)	0.4951 (6)	0.1714 (2)	5.33
	HetNetE		0.2123 (6)	0.3870 (13)	0.1669 (8)	0.2616 (6)	0.4571 (11)	0.1714 (2)	7.66
Top-10	Similarity measure	PTPA	0.1555 (9)	0.5779 (2)	0.1454 (10)	0.2365 (9)	0.5897 (2)	0.5023 (3)	5.83
		PCPA	0.1388 (11)	0.5066 (5)	0.1257 (13)	0.2110 (11)	0.4517 (12)	0.5023 (3)	9.10
	Feature method	LR	0.1358 (13)	0.5005 (8)	0.1270 (12)	0.2059 (13)	0.4664 (10)	0.5023 (3)	9.83
		SVM	0.1629 (8)	**0.5988** (1)	0.1538 (9)	0.2477 (8)	0.5296 (4)	0.5023 (3)	5.50
		Bayesian	0.1364 (12)	0.5010 (6)	0.1277 (11)	0.2069 (12)	0.4767 (8)	0.5023 (3)	8.67
	HetNetE		0.1506 (10)	0.5347 (3)	0.2269 (3)	0.2275 (10)	0.4435 (13)	0.5023 (3)	7.00
ASI			**0.4589** (1)	0.5284 (4)	**0.4009** (1)	**0.4712** (1)	0.5295 (5)	**0.1123** (1)	2.00

Table 4: Results of effectiveness experiments on AMiner-II. We use bold to mark the best performance for each comparison. " indicates higher is better, # indicates lower is better. "Avg." means the average rank of different methods.

Methods			Evaluation						Avg.
			P (")	R (")	J (")	F1 (")	MAP (")	RMSE (#)	
Top-5	Similarity measure	PTPA	0.3391 (2)	0.5899 (6)	0.2886 (2)	0.4108 (2)	0.7165 (3)	0.2880 (2)	2.83
		PCPA	0.3287 (3)	0.5743 (8)	0.2776 (4)	0.3986 (3)	0.6595 (6)	0.2880 (2)	4.33
	Feature method	LR	0.3113 (4)	0.5400 (9)	0.2645 (5)	0.3769 (4)	0.6605 (5)	0.2880 (2)	4.83
		SVM	0.2202 (7)	0.4553 (12)	0.1674 (11)	0.2803 (9)	**0.9948** (1)	0.2880 (2)	7.00
		Bayesian	0.2964 (5)	0.5144 (10)	0.2491 (6)	0.3587 (5)	0.6458 (8)	0.2880 (2)	6.00
	HetNetE		0.2645 (6)	0.4561 (11)	0.2078 (7)	0.3191 (6)	0.6021 (12)	0.2880 (2)	7.33
Top-10	Similarity measure	PTPA	0.1927 (8)	**0.6624** (1)	0.1795 (8)	0.2884 (7)	0.6913 (4)	0.8536 (3)	5.16
		PCPA	0.1913 (9)	0.6531 (2)	0.1778 (9)	0.2860 (8)	0.6363 (10)	0.8536 (3)	6.83
	Feature method	LR	0.1857 (10)	0.5779 (7)	0.1729 (10)	0.2775 (10)	0.6382 (9)	0.8536 (3)	8.16
		SVM	0.1101 (13)	0.4553 (12)	0.0943 (13)	0.1702 (13)	0.9948 (1)	0.8536 (3)	5.00
		Bayesian	0.1786 (11)	0.6157 (4)	0.1661 (12)	0.2673 (11)	0.6227 (11)	0.8536 (3)	8.66
	HetNetE		0.1720 (12)	0.6350 (3)	0.2858 (3)	0.2564 (12)	0.5602 (13)	0.8536 (3)	7.66
ASI			**0.5981** (1)	0.6019 (5)	**0.4943** (1)	**0.5720** (1)	0.6566 (7)	**0.2058** (1)	2.66



(a) Dimension (b) Training times (c) Control factor

Figure 4: The performance of ASI with different parameter settings.

Table 5: Top ranked authors for two query papers. We list top-10 authors for three baselines–PTPA, SVM and HetNetE, while we list the extracted optimal author set with the automatically determined number of authors for ASI, which is unordered.

(a) Paper title: who are experts specializing in landscape photography?: analyzing topic-specific authority on content sharing services. (KDD 2014)

Ground-truth	PTPA	SVM	HetNetE	ASI
B. Bi	**J. Cho**	L. Denoyer	T. Calders	**J. Cho**
B. Kao	S. Geva	P. Mirajkar	G. Das	S. Geva
C. Wan	L. Denoyer	S. Geva	**B. Kao**	**B. Bi**
J. Cho	P. Mirajkar	**J. Cho**	C. J. Hsieh	**B. Kao**
	A. Seetharaman	C. J. Hsieh	J. Vaidya	P. Mirajkar
	R. Zhang	J. Vaidya	K. M. Borgwardt	L. Denoyer
	O. Maimon	K. M. Borgwardt	C. H. Park	
	J. L. Huang	B. Wang	M. Gori	
	G. Giannakopoulos	C. H. Park	A. Laurent	
	P. Fournier-Viger	B. Li	A. S. Varde	

(b) Paper title: similarity search using concept graphs. (CIKM 2014)

Ground-truth	PTPA	SVM	HetNetE	ASI
R. Agrawal	**S. Gollapudi**	C. S. Perng	**R. Agrawal**	**S. Gollapudi**
S. Gollapudi	D. McLeod	D. McLeod	N. Pissinou	**K. Kenthapadi**
A. Kannan	**A. Kannan**	**S. Gollapudi**	**S. Gollapudi**	C. S. Perng
K. Kenthapadi	C. S. Perng	K. Lee	C. S. Perng	D. Yang
	S. Ofek-Koifman	S. Ofek-Koifman	S. Bressan	**A. Kannan**
	T. Raeder	F. Coetzee	**A. Kannan**	**R. Agrawal**
	Conrad Murphy	Deepa Paranjpe	Toon Calders	
	S. H. Wu	M. A. Hasan	M. A. Hasan	
	K. Satou	**A. Kannan**	J. Chen	
	C. Siefkes	T. Raeder	**K. Kenthapadi**	

work, we will consider how to combine multiple information such as unstructured semantic content to improve the performance. We will also consider leveraging more information such as text to acquire the representation of papers. In addition, we will extend our approach to other applications like actor set identification for a given movie.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (No. 61772082, 61702296, 61806020), the National Key Research and Development Program of China (2017YFB0803304), the Beijing Municipal Natural Science Foundation (4182043), and the CCF-Tencent Open Fund. This work is partially supported by the NSF under grants CNS-1618629, CNS-1814825, CNS-1845138 and OAC-1839909, the NIJ 2018-75-CX-0032. This work is also supported in part by China Scholarship Council.

REFERENCES

[1] Yuichi Asahiro, Kazuo Iwama, Hisao Tamaki, and Takeshi Tokuyama. 2000. Greedily finding a dense subgraph. *Journal of Algorithms* 34, 2 (2000), 203–221.
[2] Devora Berlowitz, Sara Cohen, and Benny Kimelfeld. 2015. Efficient Enumeration of Maximal k-Plexes. In *Acm Sigmod International Conference on Management of Data*.
[3] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. 2010. Detecting high log-densities: an O (n 1•4) approximation for densest k-subgraph. In *Proceedings of TOC*. ACM, 201–210.
[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*. 2787–2795.
[5] Jie Chen and Yousef Saad. 2012. Dense subgraph extraction with application to community detection. *IEEE Transactions on Knowledge and Data Engineering* 24, 7 (2012), 1216–1230.
[6] Ting Chen and Yizhou Sun. 2017. Task-guided and path-augmented heterogeneous network embedding for author identification. In *Proceedings of WSDM*. ACM, 295–304.
[7] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. 2015. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068* (2015).
[8] Alessio Conte, Donatella Firmani, Caterina Mordente, Maurizio Patrignani, and Riccardo Torlone. 2017. Fast Enumeration of Large k-Plexes. In *Acm Sigkdd International Conference on Knowledge Discovery Data Mining*.
[9] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of SIGKDD*. ACM, 135–144.
[10] Christos Giatsidis, Dimitrios M Thilikos, and Michalis Vazirgiannis. 2013. D-cores: measuring collaboration of directed graphs based on degeneracy. *Knowledge and information systems* 35, 2 (2013), 311–343.
[11] Andrew V Goldberg. 1984. *Finding a maximum density subgraph*. University of California Berkeley, CA.
[12] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of SIGKDD*. ACM, 855–864.
[13] Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: author identification using only citations. *Acm Sigkdd Explorations Newsletter* 5, 2 (2003), 179–184.
[14] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. 2016. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of SIGKDD*. ACM, 895–904.
[15] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative metric learning. In *Proceedings of WWW*. 193–201.
[16] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. 2016. Meta structure: Computing relevance in large heterogeneous information networks. In *Proceedings of SIGKDD*. ACM, 1595–1604.
[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[18] Jing Li, Feng Xia, Wei Wang, Zhen Chen, Nana Yaw Asabere, and Huizhen Jiang. 2014. Acrec: a co-authorship based random walk model for academic

collaboration recommendation. In *Proceedings of WWW*. ACM, 1209–1214.
[19] Xiaozhong Liu, Yingying Yu, Chun Guo, and Yizhou Sun. 2014. Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. In *Proceedings of CIKM*. ACM, 121–130.
[20] Yuanfu Lu, Chuan Shi, Linmei Hu, and Zhiyuan Liu. 2019. Relation Structure-Aware Heterogeneous Information Network Embedding. In *Proceedings of AAAI*.
[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[22] Mathias Payer, Ling Huang, Neil Zhenqiang Gong, Kevin Borgolte, and Mario Frank. 2015. What you submit is who you are: a multimodal approach for deanonymizing scientific publications. *IEEE Transactions on Information Forensics and Security* 10, 1 (2015), 200–212.
[23] Jian Pei, Daxin Jiang, and Aidong Zhang. 2005. On mining cross-graph quasi-cliques. In *Proceedings of SIGKDD*. ACM, 228–238.
[24] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of SIGKDD*. ACM, 701–710.
[25] Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 2014. Cluscite: Effective citation recommendation by information network-based clustering. In *Proceedings of SIGKDD*. ACM, 821–830.
[26] Chuan Shi, Binbin Hu, Xin Zhao, and Philip Yu. 2018. Heterogeneous Information Network Embedding for Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2018).
[27] Chuan Shi, Xiangnan Kong, Yue Huang, S Yu Philip, and Bin Wu. 2014. HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. *IEEE Transactions on Knowledge and Data Engineering* 26, 10 (2014), 2479–2492.
[28] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2017. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2017), 17–37.
[29] Mauro Sozio and Aristides Gionis. 2010. The community-search problem and how to plan a successful cocktail party. In *Proceedings of SIGKDD*. ACM, 939–948.
[30] Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3, 2 (2012), 1–159.
[31] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment* 4, 11 (2011), 992–1003.
[32] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of SIGKDD*. ACM, 1165–1174.
[33] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of WWW*. 1067–1077.
[34] Jie Tang, Sen Wu, Jimeng Sun, and Hang Su. 2012. Cross-domain collaboration recommendation. In *Proceedings of SIGKDD*. ACM, 1285–1293.
[35] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of SIGKDD*. ACM, 990–998.
[36] Charalampos Tsourakakis, Francesco Bonchi, Aristides Gionis, Francesco Gullo, and Maria Tsiarli. 2013. Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees. In *Proceedings of SIGKDD*. ACM, 104–112.
[37] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *Proceedings of SIGKDD*. ACM, 1225–1234.
[38] Xiao Wang, Yiding Zhang, and Chuan Shi. 2019. Hyperbolic Heterogeneous Information Network Embedding. In *Proceedings of AAAI*.
[39] Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, Feb (2009), 207–244.
[40] Chuxu Zhang, Chao Huang, Lu Yu, Xiangliang Zhang, and Nitesh V Chawla. 2018. Camel: Content-Aware and Meta-path Augmented Metric Learning for Author Identification. In *Proceedings of WWW*. 709–718.