

# A Survey of Heterogeneous Information Network Analysis

Chuan Shi, *Member, IEEE,*  
 Yitong Li, Jiawei Zhang, Yizhou Sun, *Member, IEEE,*  
 and Philip S. Yu, *Fellow, IEEE*

## Abstract

Most real systems consist of a large number of interacting, multi-typed components, while most contemporary researches model them as homogeneous networks, without distinguishing different types of objects and links in the networks. Recently, more and more researchers begin to consider these interconnected, multi-typed data as heterogeneous information networks, and develop structural analysis approaches by leveraging the rich semantic meaning of structural types of objects and links in the networks. Compared to widely studied homogeneous network, the heterogeneous information network contains richer structure and semantic information, which provides plenty of opportunities as well as a lot of challenges for data mining. In this paper, we provide a survey of heterogeneous information network analysis. We will introduce basic concepts of heterogeneous information network analysis, examine its developments on different data mining tasks, discuss some advanced topics, and point out some future research directions.

## Index Terms

heterogeneous information network, data mining, semi-structural data, meta path

C. Shi, Y.T. Li are with Beijing University of Posts and Telecommunications, Beijing, China. E-mail: shichuan@bupt.edu.cn, liyitong@bupt.edu.cn.

J.W. Zhang and P.S. Yu are with University of Illinois at Chicago, IL, USA. E-mail: jwzhanggy@gmail.com, psyu@uic.edu.

Y.Z. Sun is with Northeastern University, MA, USA. E-mail: yzsun@ccs.neu.edu.

## I. INTRODUCTION

We know that most real systems usually consist of a large number of interacting, multi-typed components [1], such as human social activities, communication and computer systems, and biological networks. In such systems, the interacting components constitute interconnected networks, which can be called information networks without loss of generality [2]. It is clear that information networks are ubiquitous and form a critical component of modern information infrastructure. The information network analysis has gained extremely wide attentions from researchers in many disciplines, such as computer science, social science, physics, and so on. Particularly, the information network analysis has become a hot research topic in data mining and information retrieval fields in the past decades. The basic paradigm is to mine hidden patterns through mining link relations from networked data. The analysis of information network is related to the works in link mining and analysis [3], [4], [5], social network analysis [6], [7], hypertext and web mining [8], network science [9], and graph mining [10].

Most of contemporary information network analyses have a basic assumption: the type of objects or links is unique [2]. That is, the networks are homogeneous containing the same type of objects and links, such as the author collaboration network [11] and the friendship network [12]. These homogeneous networks usually are extracted from real interacting systems by simply ignoring the heterogeneity of objects and links or only considering one type of relations among one type of objects. However, most real systems contain multi-typed interacting components and we can model them as heterogeneous information networks [2] (called HIN or heterogeneous network for short) with different types of objects and links. For example, in bibliographic database, like DBLP [2], papers are connected together via authors, venues and terms; and in Flickr, photos are linked together via users, groups, tags and comments.

Compared to widely-used homogeneous information network, the heterogeneous information network can effectively fuse more information and contain rich semantics in nodes and links, and thus it forms a new development of data mining. More and more researchers have noticed the importance of heterogeneous information network analysis and many novel data mining tasks have been exploited in such networks, such as similarity search [13], [14], clustering [15], and classification [16]. Since Y. Sun, J. Han, et al. proposed the concept of heterogeneous information network in 2009 [17], and the meta path concept subsequently in [14], heterogeneous

information network analysis becomes a hot topic rapidly in the fields of data mining, database, and information retrieval, and a lot of papers have appeared in top conferences and journals of these research fields. In addition, some special workshops on heterogeneous information networks began to be held. For example, the workshop on Heterogeneous Information Network Analysis (HINA) has been held for 3 years in conjunction with IJCAI, and the workshop on Mining Data Semantics (MDS) has also been held for several times.

This paper firstly presents a survey of heterogeneous information network analysis in recent years. Although some articles have introduced the developments of this field [2], [18], [19], they focus on summarizing the works of authors themselves. This paper attempts to clearly introduce basic concepts in heterogeneous network analysis and make a comprehensive investigation on contemporary research developments. Moreover, this paper also discusses some advanced topics and points out several future development directions of this field.

The remaining part is organized as follows. Section II introduces the basic concepts and examples in this field. Section III presents research developments in major data mining tasks, and the advanced topics and future works are introduced in Section IV. Finally, Section V concludes this paper.

## II. BASIC CONCEPTS AND DEFINITIONS

In this section, we introduce some basic concepts in this field, compare the heterogeneous information network with other related concepts, and give some HIN examples.

### A. Basic definitions

An information network represents an abstraction of the real world, focusing on the objects and the interactions among these objects. Formally, we define an information network as follows.

**DEFINITION 1: Information Network** [2], [20]. An information network is defined as a directed graph  $G = (V, E)$  with an object type mapping function  $\varphi : V \rightarrow \mathcal{A}$  and a link type mapping function  $\psi : E \rightarrow \mathcal{R}$ . Each object  $v \in V$  belongs to one particular object type in the object type set  $\mathcal{A}$ :  $\varphi(v) \in \mathcal{A}$ , and each link  $e \in E$  belongs to a particular relation type in the relation type set  $\mathcal{R}$ :  $\psi(e) \in \mathcal{R}$ . If two links belong to the same relation type, the two links share the same starting object type as well as the ending object type.

Different from the traditional network definition, we explicitly distinguish object types and relationship types in information network.

**DEFINITION 2: Heterogeneous/homogeneous information Network.** The information network is called **heterogeneous information network** if the types of objects  $|\mathcal{A}| > 1$  or the types of relations  $|\mathcal{R}| > 1$ ; otherwise, it is a **homogeneous information network**.

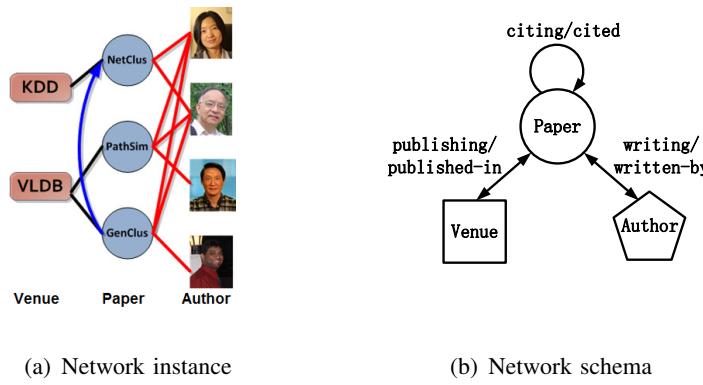


Fig. 1. An example of heterogeneous information network on bibliographic data [2].

**EXAMPLE 1:** Fig. 1 shows a HIN example on bibliographic data [2]. A bibliographic information network, such as the bibliographic network involving computer science researchers derived from DBLP<sup>1</sup>, is a typical heterogeneous network containing three types of information entities: papers, venues, and authors. For each paper, it has links to a set of authors, and a venue, and these links belong to a set of link types.

In order to better understand the object types and link types in a complex heterogeneous information network, it is necessary to provide the meta level (i.e., schema-level) description of the network. Therefore, the concept of network schema is proposed to describe the meta structure of a network.

**DEFINITION 3: Network schema** [2], [20]. The network schema, denoted as  $T_G = (\mathcal{A}, \mathcal{R})$ , is a meta template for an information network  $G = (V, E)$  with the object type mapping  $\varphi : V \rightarrow \mathcal{A}$  and the link type mapping  $\psi : E \rightarrow \mathcal{R}$ , which is a directed graph defined over object types  $\mathcal{A}$ , with edges as relations from  $\mathcal{R}$ .

<sup>1</sup><http://dblp.uni-trier.de/>

The network schema of a heterogeneous information network specifies type constraints on the sets of objects and relationships among the objects. These constraints make a heterogeneous information network semi-structured, guiding the semantics explorations of the network. An information network following a network schema is called a **network instance** of the network schema. For a link type  $R$  connecting object type  $S$  to object type  $T$ , i.e.,  $S \xrightarrow{R} T$ ,  $S$  and  $T$  are the **source object type** and **target object type** of link type  $R$ , which can be denoted as  $R.S$  and  $R.T$ , respectively. The inverse relation  $R^{-1}$  holds naturally for  $T \xrightarrow{R^{-1}} S$ . Generally,  $R$  is not equal to  $R^{-1}$ , unless  $R$  is symmetric.

**EXAMPLE 2:** As described above, Fig. 1(a) demonstrates the real objects and their connections on bibliographic data. Fig. 1(b) illustrates its network schema which describes the object types and their relations in the HIN. Moreover, Fig. 1(a) is a network instance of the network schema Fig. 1(b). In this example, it contains objects from three types of objects: papers ( $P$ ), authors ( $A$ ), and venues ( $V$ ). There are links connecting different types of objects. The link types are defined by the relations between two object types. For example, links existing between authors and papers denote the writing or written-by relations, while those between venues and papers denote the publishing or published-in relations.

Different from homogeneous networks, two objects in a heterogeneous network can be connected via different paths and these paths have different physical meanings. These paths can be categorized as meta paths as follows.

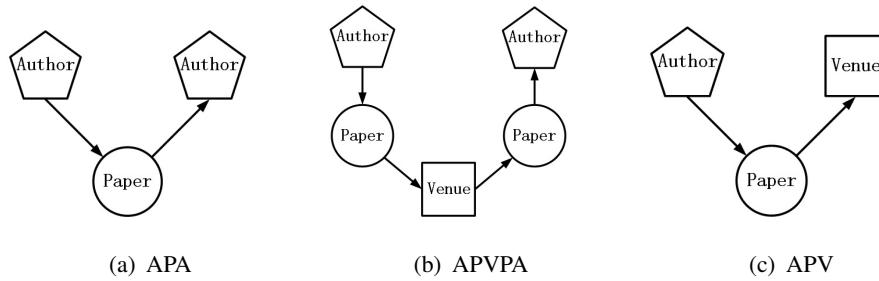


Fig. 2. Examples of meta paths in heterogeneous information network on bibliographic data.

**DEFINITION 4: Meta path** [14]. A meta path  $\mathcal{P}$  is a path defined on a schema  $S = (\mathcal{A}, \mathcal{R})$ , and is denoted in the form of  $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$ , which defines a composite relation  $R = R_1 \circ R_2 \circ \dots \circ R_l$  between objects  $A_1, A_2, \dots, A_{l+1}$ , where  $\circ$  denotes the composition operator on relations.

For simplicity, we can also use object types to denote the meta path if there are no multiple relation types between the same pair of object types:  $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$ . For example, in Fig. 1(a), the relation, authors publishing papers in conferences, can be described using the length-2 meta path  $A \xrightarrow{\text{writing}} P \xrightarrow{\text{written by}} A$ , or *APA* for short. We say a concrete path  $p = (a_1 a_2 \cdots a_{l+1})$  between objects  $a_1$  and  $a_{l+1}$  in network  $G$  is a **path instance** of the relevance path  $\mathcal{P}$ , if for each  $a_i$ ,  $\phi(a_i) = A_i$  and each link  $e_i = \langle a_i, a_{i+1} \rangle$  belongs to the relation  $R_i$  in  $\mathcal{P}$ . It can be denoted as  $p \in \mathcal{P}$ . A meta path  $\mathcal{P}$  is a **symmetric path**, if the relation  $R$  defined by it is symmetric (i.e.,  $\mathcal{P}$  is equal to  $\mathcal{P}^{-1}$ ), such as *APA* and *APVPA*. Two meta paths  $\mathcal{P}_1 = (A_1 A_2 \cdots A_l)$  and  $\mathcal{P}_2 = (B_1 B_2 \cdots B_k)$  are **concatenable** if and only if  $A_l$  is equal to  $B_1$ , and the concatenated path is written as  $\mathcal{P} = (\mathcal{P}_1 \mathcal{P}_2)$ , which equals to  $(A_1 A_2 \cdots A_l B_2 \dots B_k)$ . A simple concatenable example is that *AP* and *PA* can be concatenated to the path *APA*.

TABLE I  
META PATH EXAMPLES AND THEIR PHYSICAL MEANINGS ON BIBLIOGRAPHIC DATA.

Path instance	Meta path	Physical meaning
Sun-NetClus-Han Sun-PathSim-Yu	Author-Paper-Author ( <i>APA</i> )	Authors collaborate on the same paper
Sun-PathSim-VLDB-PathSim-Han Sun-PathSim-VLDB-GenClus-Agarwal	Author-Paper-Venue-Paper-Author ( <i>APVPA</i> )	Authors publish papers on the same venue
Sun-NetClus-KDD Sun-PathSim-VLDB	Author-Paper-Venue ( <i>APV</i> )	Authors publish papers at a venue

EXAMPLE 3: As examples shown in Fig. 2, authors can be connected via meta paths “Author-Paper-Author” (*APA*) path, “Author-Paper-Venue-Paper-Author” (*APVPA*) path, and so on. Moreover, TABLE I shows path instances and semantics of these meta paths. It is obvious that semantics underneath these paths are different. The *APA* path means authors collaborating on the same papers (i.e., co-author relation), while *APVPA* path means authors publishing papers on the same venue. The meta paths can also connect different types of objects. For example, the authors and venues can be connected with the *APV* path, which means authors publishing papers on venues.

The rich semantics of meta path is an important characteristic of HIN. Based on different meta paths, objects have different connection relations with diverse path semantics, which may have an effect on many data mining tasks. For example, the similarity scores among authors evaluated

based on different meta paths are different [14]. Under the *APA* path, the authors co-publishing papers will be more similar, while the authors publishing papers on the same venues will be more similar under the *APVPA* path. Another example is the importance evaluation of objects [21]. The importance of authors under *APA* path has a bias on the authors who write many papers having many authors, while the importance of authors under *APVPA* path emphasizes the authors who publish many papers on those productive conferences. As a unique characteristic and effective semantic capturing tool, meta path has been widely used in many data mining tasks in HIN, such as similarity measure [13], [14], clustering [15], and classification [16].

### *B. Comparisons with related concepts*

With the boom of social network analysis, all kinds of networked data have emerged, and numbers of concepts to model networked data have been proposed. Here we compare heterogeneous network concept with these related concepts.

**Heterogeneous network vs homogeneous network.** Heterogeneous networks include different types of nodes or links, while homogeneous networks only have one type of objects and links. Homogeneous networks can be considered as a special case of heterogeneous networks. Moreover, a heterogeneous network can be converted into a homogeneous network through network projection or ignoring object heterogeneity, while it will make significant information loss. Traditional link mining [22], [23], [24] is usually based on homogeneous network, and many analysis techniques on homogeneous network cannot be directly applied to heterogeneous network.

**Heterogeneous network vs multi-relational network [25].** Different from heterogeneous network, multi-relational network has only one type of objects, but more than one kind of relationship between objects. So multi-relational network can be seen as a special case of heterogeneous network.

**Heterogeneous network vs multi-dimensional/mode network [26].** Tang et al. [26] proposed the multi-dimensional/mode network concept, which has the same meaning with multi-relational network. That is, the network has only one type of objects and more than one kind of relationship between objects. So multi-dimensional/mode network is also a special case of heterogeneous network.

**Heterogeneous network vs composite network [27], [28].** Qiang Yang et al. proposed the

composite network concept [27], [28], where users in networks have various relationships, exhibit different behaviors in each individual network or subnetwork, and share some common latent interests across networks at the same time. So composite network is in fact a multi-relational network, a special case of heterogeneous network.

**Heterogeneous network vs complex network.** A complex network is a network with non-trivial topological features and patterns of connection between its elements that are neither purely regular nor purely random [29]. Such non-trivial topological features include a heavy tail in the degree distribution, a high clustering coefficient, community structure, and hierarchical structure. The studies of complex networks have brought together researchers from many areas including mathematics, physics, biology, computer science, sociology, and others. The studies show that many real networks are complex networks, such as social networks, information networks, technological networks, biological networks, and so on [30]. So we can say that many real heterogeneous networks are complex networks. However, the studies on complex networks usually focus on the structures, functions, and features of networks.

### C. Example datasets of heterogeneous information network

Intuitively, most real systems include multi-typed interacting objects. For example, a social media website (e.g., Facebook) contains a set of object types, such as users, posts, and tags, and a health care system contains doctors, patients, diseases, and devices. Generally speaking, these interacting systems can all be modeled as heterogeneous information networks. Concretely, this kind of networks can be constructed from the following three types of data.

1) Structured data. Structured data stored in database table is organized with entity-relation model. The different-typed entities and their relations naturally construct information networks. For example, the bibliographic data (see the above example) is widely used as heterogeneous information network.

2) Semi-structured data. Semi-structured data is usually stored with XML format. The attributes in XML can be considered as object types, and the object instances can be determined by analyzing the contents of attributes. The connections among attributes construct object relations.

3) Non-structured data. For non-structured data, heterogeneous information networks can also be constructed by objects and relationship extraction. For example, for text data, entity recognition and relation extraction can form the objects and links of HIN.

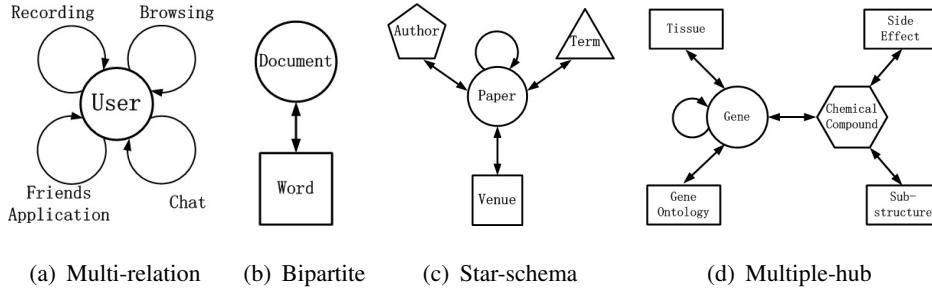


Fig. 3. Network schema of heterogeneous information networks.

Although heterogeneous information networks are ubiquitous, there are not many standard datasets for study. Here we summarize some widely used heterogeneous networks in literals.

**Multi-relational network with single-typed object.** Traditional multi-relational network is a kind of HIN, where there is one type of object and several types of relations among objects. This kind of networks widely exist in social websites, like Facebook and Xiaonei [28]. Fig. 3(a) shows the network schema of such a network [28], where users can be extensively connected with each other through connections, such as recording, browsing, chatting, and sending friends applications.

**Bipartite network.** As a typical HIN, bipartite network is widely used to construct interactions among two types of objects, such as user-item [31], and document-word [32]. Fig. 3(b) shows the schema of a bipartite network connecting documents and words [32]. As an extension of bipartite graphs,  $k$ -partite graphs [33] contain multiple types of objects where links exist among adjacent object types.

**Star-schema network.** Star-schema network is the most popular HIN in this field. In database table, a target object and its attribute objects naturally construct a HIN, where the target object, as the hub node, connects different attribute objects. As an example shown in Fig. 3(c), a bibliographic information network is a typical star-schema heterogeneous network [14], [13], containing different objects (e.g., paper, venue, author, and term) and links among them. Many other datasets can also be represented as star-schema networks, such as the movie data [34], [35] from the Internet Movie Database <sup>2</sup> (IMDB) and the patent data [36] from US patents data

<sup>2</sup>[www.imdb.com/](http://www.imdb.com/)

<sup>3</sup>.

**Multiple-hub network.** Beyond star schema, some networks have more complex structures, which involve multiple hub objects. This kind of networks widely exist in bioinformatics data [37], [38]. A bioinformatics example, shown in Fig. 3(d), includes two hubs: Gene and Chemical compound. Another example can be found in the Douban dataset <sup>4</sup> [39].

Besides these widely used networks, many real systems can also be constructed as more complex heterogeneous networks. In some real applications, users may exist in multiple social networks, and each social network can be modeled as an HIN. Fig. 4(a) shows an example of two heterogeneous social networks (Twitter and Foursquare) [40]. In each network, users are connected with each other through social links, and they are also connected with a set of locations, timestamps and text contents through online activities. Moreover, some users have two accounts in two social networks separately, and they serve as anchor nodes to connect two networks. More generally, some interaction systems are too complex to be modeled as an HIN with a simple network schema. Knowledge graph [41] is such an example. We know that knowledge graph is based on Resource Description Framework (RDF) data, which complies with an  $\langle \text{Subject}, \text{Property}, \text{Object} \rangle$  model. Here “Subject” and “Object” can be considered as objects, and “Property” can be considered as the relation between “Subject” and “Object”. And thus a knowledge graph can be considered as a heterogeneous network, an example is shown in Fig. 4(b). In such a semantic knowledge base, like Yago [42], there are more than 10 million entities (or nodes) of different types, and more than 120 million links among these entities. In such a schema-rich network, it is impossible to depict such network with a simple network schema.

#### D. Why Heterogeneous Information Network Analysis

In the past decades, link analysis has been extensively explored [3]. So many methods have been developed for information network analysis and numerous data mining tasks have been explored in homogeneous networks, such as ranking, clustering, link prediction, and influence analysis. However, due to some unique characteristics (e.g., fusion of more information and

<sup>3</sup><http://www.uspto.gov/patents/>

<sup>4</sup><http://www.douban.com/>

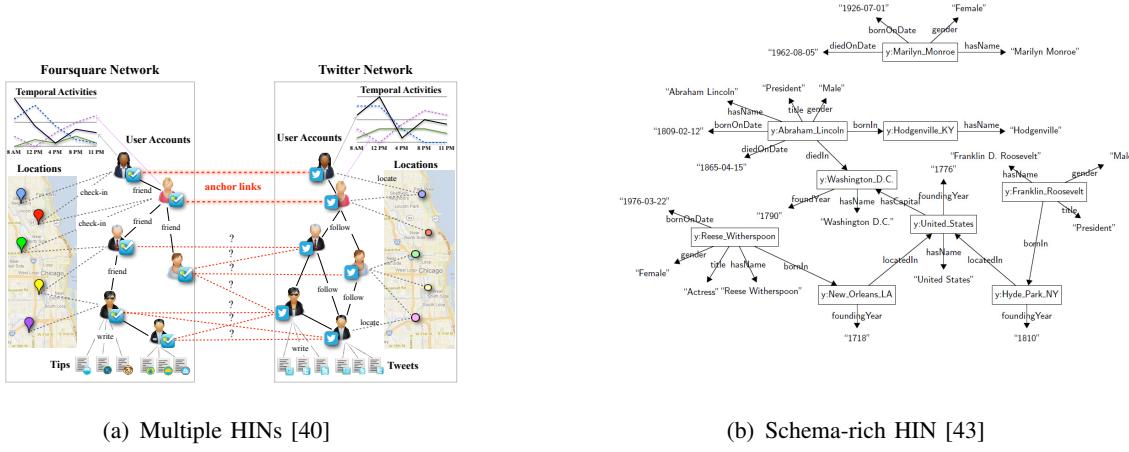


Fig. 4. Two examples of complex heterogeneous information network.

rich semantics) of HIN, most methods in homogeneous networks cannot be directly applied in heterogeneous networks, and it is potential to discover more interesting patterns in this kind of networks.

**It is a new development of data mining.** Early data mining problems focused on analyzing feature vectors of objects. In the late 1990s, with the advent of WWW, more and more data mining researches turned to study links among objects. It is one of the main research directions to mine hidden patterns from feature and link information of objects. In these researches, homogeneous networks are usually constructed from interconnected objects. In recent years, abundant social media emerge, and many different types of objects are interconnected. It is hard to model these interacted objects as homogeneous networks, while it is natural to model different types of objects and relations among them as heterogeneous networks. Particularly, with the rapid increment of user-generated content online, big data analysis is an emergent yet important task to be studied. Variety is one significant characteristic of big data [44]. As a semi-structured representation, heterogeneous information network can be an effective tool to deal with complex big data.

**It is an effective tool to fuse more information.** Compared to homogeneous network, heterogeneous network is natural to fuse more objects and their interactions. In addition, traditional homogeneous networks are usually constructed from single data source, while heterogeneous network can fuse information across multiple data sources. For example, customers use many services provided by Google, such as Google search, G-mail, maps, Google+, etc. So we can

fuse these information with a heterogenous information network, in which customers interact with many different types of objects, such as key words, mails, locations, followers, etc. Broadly speaking, heterogeneous information network can also fuse information cross multiple social network platforms. We know that there are many social network platforms with different objectives, such as Facebook, Twitter, Weixin, and Weibo. Moreover, users often participate in multiple social networks. Since each social network only captures a partial or biased view of a user, we can fuse information across multiple social network platforms with multiple heterogeneous information networks, where each heterogeneous network represents information from one social network with some anchor nodes connecting these networks.

**It contains rich semantics.** In heterogeneous networks, different-typed objects and links coexist and they carry different semantic meanings. As a bibliographic example shown in Fig. 1, it includes author, paper, and venue object types. The relation type “Author-Paper” means author writing paper, while the relation type “Paper-Venue” means paper published in venue. Considering the semantic information will lead to more subtle knowledge discovery. For example, in DBLP bibliographic data [14], if you find the most similar authors to “Christos Faloutsos”, you will get his students, like Spiros Papadimitriou and Jimeng Sun, under the *APA* path; while the results are reputable researchers, like Jiawei Han and Rakesh Agrawal, under the *APVPA* path. How to mine interesting patterns with the semantic information is a unique issue in heterogeneous network.

### III. RESEARCH DEVELOPMENTS

Heterogeneous information network provides a new paradigm to manage networked data. Meanwhile, it also introduces new challenges for many data mining tasks. We have analyzed more than 100 papers in this field, and divided them into 7 categories according to their data mining tasks. The proportion of papers belonging to each category is shown in Fig. 5. In this section, we will summarize the developments about these 7 main data mining tasks.

#### A. Similarity Measure

Similarity measure is to evaluate the similarity of objects. It is the basis of many data mining tasks, such as web search, clustering, and product recommendation. Similarity measure has been well studied for a long time. These studies can be roughly categorized into two types:

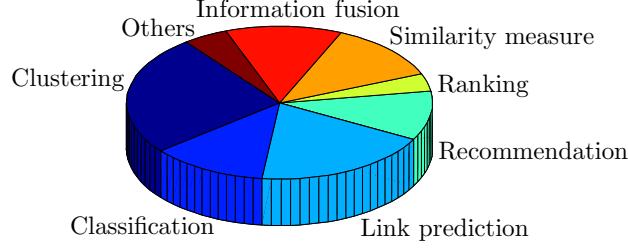


Fig. 5. Paper distribution of heterogeneous information network analysis on different data mining tasks.

TABLE II

TOP-5 MOST SIMILAR AUTHORS TO “CHRISTOS FALOUTSOS” UNDER DIFFERENT META PATHS ON DBLP DATASET.

Rank	Authors	
	APA	APCPA
1	Christos Faloutsos	Christos Faloutsos
2	Spiros Papadimitriou	Jiawei Han
3	Jimeng Sun	Rakesh Agrawal
4	Jia-Yu Pan	Jian Pei
5	Agma J. M. Traina	Charu C. Aggarwal

feature based approaches and link based approaches. The feature based approaches measure the similarity of objects based on their feature values, such as cosine similarity, Jaccard coefficient, and Euclidean distance. The link based approaches measure the similarity of objects based on their link structures in a graph. For example, Personalized PageRank [45] evaluates the probability starting from a source object to a target object by randomly walking with restart, and SimRank [46] evaluates the similarity of two objects by their neighbors’ similarities.

Recently, many researchers begin to consider similarity measure on heterogeneous information networks. Different from similarity measure on homogeneous networks, similarity measure on HIN not only considers structure similarity of two objects but also takes the meta path connecting these two objects into account. As we know, there are different meta paths connecting two objects, and these meta paths contain different semantic meanings, which may lead to different similarities. And thus the similarity measure on HIN is meta path constraint. For example, based on the bibliographic network in Fig. 1, TABLE II shows the most similar authors to Christos

Faloutsos, a well-known expert in data mining, under different meta paths [14]. Based on *APA* path, the most similar authors to Christos are his students (e.g., Spiros Papadimitriou and Jimeng Sun), while the most similar authors are reputable researchers in the same field with Christos under the *APVPA* path (e.g., Jiawei Han and Rakesh Agrawal).

Considering semantics in meta paths constituted by different-typed objects, Sun et al. [14] first propose the path based similarity measure PathSim to evaluate the similarity of same-typed objects based on symmetric paths. Following their work, some researchers [47], [48] extend PathSim by incorporating richer information, such as transitive similarity, temporal dynamics, and supportive attributes. A path-based similarity join method [49] is proposed to return the top  $k$  similar pairs of objects based on user specified join paths. In information retrieval community, Lao and Cohen [50], [51] propose a Path Constrained Random Walk (PCRW) model to measure the entity proximity in a labeled directed graph constructed by the rich metadata of scientific literature.

In order to evaluate the relevance of different-typed objects, Shi et al. [13], [52] propose HeteSim to measure the relevance of any object pair under arbitrary meta path. As an adaption of HeteSim, LSH-HeteSim [53] is proposed to mine the drug-target interaction in heterogeneous biological networks where drugs and targets are connected with complicated semantic paths. In order to overcome the shortcoming of HeteSim in high computation and memory demand, Meng et al. [54] propose the AvgSim measure that evaluates similarity score through two random walk processes along the given meta path and the reverse meta path, respectively. In addition, some methods [55], [56] combine meta path based relevance search with user preference.

More works begin to integrate the network structure and other information to measure similarity of objects in HIN. Combining the influence and similarity information, Wang et al. [57] simultaneously measure social influence and object similarity in a heterogeneous network to produce more meaningful similarity scores. Wang et al. [58] propose a model to learn relevance through analyzing the context of heterogeneous networks for online targeting. Yu et al. [59] predict the semantic meaning based on a user's query in the meta-path-based feature space and learn a ranking model to answer the similarity query. Recently, Zhang et al. [60] propose a similarity measure to compute similarity between centers in an x-star network according to the attribute similarities and the connections among centers.

### B. Clustering

Clustering analysis is the process of partitioning a set of data objects (or observations) into a set of clusters, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. Conventional clustering is based on the features of objects, such as k-means and so on [61]. Recently, clustering based on networked data (e.g., community detection) has been studied a lot. This kind of methods model the data as a homogeneous network, and use the given measure (e.g., normalized cuts [62], and modularity [63]) to divide the network into a series of subgraphs. Many algorithms have been proposed to solve this NP-hard problem, such as spectral method [64], greedy method [65] and sampling technique [66]. Some researches also simultaneously consider objects' link structure and attribute information to increase the accuracy of clustering [67], [68].

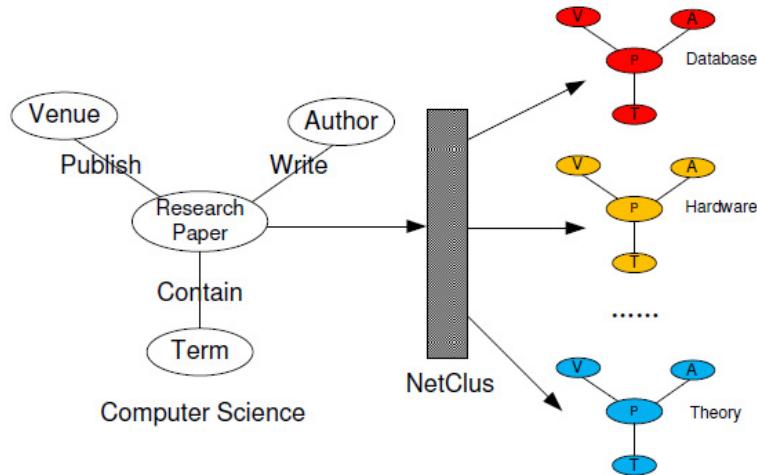


Fig. 6. Clustering on a bibliographic heterogeneous network [20].

Recently, clustering of heterogeneous networks attracts much attention. Compared with homogeneous networks, heterogeneous networks integrate multi-typed objects, which generates new challenges for clustering tasks. On the one hand, multiple types of objects co-existing in a network lead to new clustering paradigms. For example, a cluster may include different types of objects sharing the same topic [20]. Fig. 6 shows the clustering process on a bibliographic heterogeneous network, which splits the original network into several layers (a set of sub-network clusters). For example, a cluster of the database area consists of a set of database authors, conferences, terms,

and papers. In this way, clustering in HIN preserves richer information, but it also faces more challenges. On the other hand, abundant information contained in HIN makes it more convenient to integrate additional information or other learning tasks for clustering. In this section, we will review these works according to the types of integrated information or tasks.

The attribute information is widely integrated into clustering analysis on HIN. Aggarwal et al. [69] use the local succinctness property to create balanced communities across a heterogeneous network. Considering the incompleteness of objects' attributes and different types of links in heterogeneous information networks, Sun et al. [70] propose a model-based clustering algorithm to integrate the incomplete attribute information and the network structure information. Qi et al. [71] propose a clustering algorithm based on heterogeneous random fields to model the structure and content of social media networks with outlier links. Cruz et al. [72] integrate structural dimension and compositional dimension which compose an attributed graph to solve the community detection problem. Recently, a density-based clustering model TCSC [73] is proposed to detect clusters considering the connections in the network and the vertex attributes.

Text information plays an important role in many heterogeneous network studies. Deng et al. [74] introduce a topic model with biased propagation to incorporate heterogeneous information network with topic modeling in a unified way. Furthermore, they [75] propose a joint probabilistic topic model for simultaneously modeling the contents of multi-typed objects of a heterogeneous information network. LSA-PTM [76] is introduced to identify clusters of multi-typed objects by propagating the topics obtained by LSA on the HIN via the links between different objects. Incorporating both the document content and various links in the text related heterogeneous network, Wang et al. [77] propose a unified topic model for topic mining and multiple objects clustering. Recently, CHINC [78] uses general-purpose knowledge as indirect supervision to improve the clustering results.

User guide information is also integrated into clustering analysis. Sun et al. [15] present a semi-supervised clustering algorithm to generate different clustering results with path selection according to user guidance. Luo et al. [79] firstly introduce the concept of relation-path to measure the similarity between same-typed objects and use the labeled information to weight relation-paths, and then propose SemiRPclus for semi-supervised learning in HIN.

Clustering is usually an independent data mining task. However, it can be integrated with other mining tasks to improve performances through mutual enhancing. Recently, ranking-based

clustering on heterogeneous information network has emerged, which shows its advantages on the mutual promotion of clustering and ranking. RankClus [17] generates clusters for a specified type of objects in a bipartite network based on the idea that the qualities of clustering and ranking are mutually enhanced. The following work NetClus [20] is proposed to handle a network with the star-schema. Wang et al. [38] introduce ComClus to promote clustering and ranking performance by applying star schema network with self loop to combine the heterogeneous and homogeneous information. In addition, a general method HeProjI is proposed to do ranking based clustering in heterogeneous networks with arbitrary schema by projecting the network into a sequence of sub-networks [80]. And Chen et al. [81] propose a probabilistic generative model to simultaneously achieve clustering and ranking on a heterogeneous network with arbitrary schema. To make use of both textual information and heterogeneous linked entities, Wang et al. [82] develop a clustering and ranking algorithm to automatically construct multi-typed topical hierarchies. What's more, Qiu et al. [83] propose an algorithm OcdRank to combine overlapping community detection and community-member ranking together in directed heterogeneous social networks.

Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Outlier detection and clustering analysis are two highly related, but different-aimed tasks. To detect outliers, Gupta et al. [84] propose an outlier-aware approach based on joint non-negative matrix factorization to discover popular community distribution patterns. Furthermore, they propose to detect association-based clique outliers in heterogeneous networks given a conjunctive select query [85]. What's more, Zhuang et al. [36] propose an outlier detection algorithm to find subnetwork outliers according to different queries and semantics. Also based on queries, Kuck et al. [86] propose a meta-path based outlierness measure for mining outliers in heterogeneous networks.

In addition, some other information is also integrated. For example, a social influence based clustering framework SI-Cluster is proposed to analyze heterogeneous information networks based on both people's connections and their social activities [87]. Besides the traditional models employed in clustering on HIN, like topic model and spectral clustering, Alqadah et al. [88] propose a novel game theoretic framework for defining and mining clusters in heterogeneous information networks.

### C. Classification

Classification is a data analysis task where a model or classifier is constructed to predict class (categorical) labels. Traditional machine learning has focused on the classification of identically-structured objects satisfying independent identically distribution (IID). However, links exist among objects in many real-world datasets, which makes objects not satisfy IID. So link based object classification has received considerable attention, where a data graph is composed of a set of objects connected to each other via a set of links. Many methods extend traditional classification methods to consider correlations among objects [89], [90]. The link based object classification usually considers that objects and links in the graph are identical respectively. That is, the objects and links among them constitute a homogeneous network.

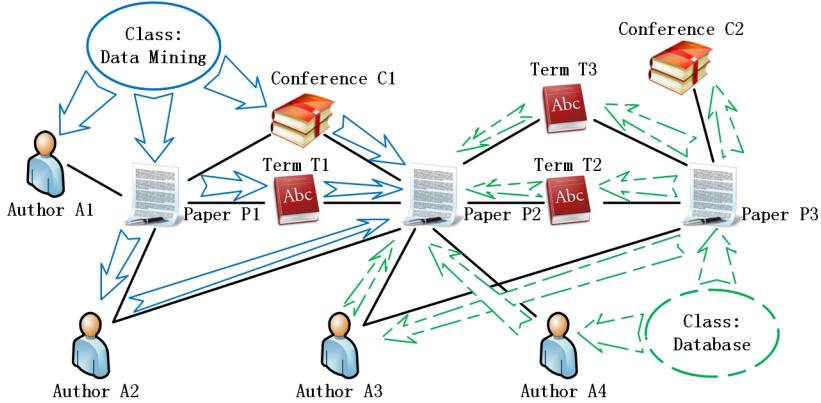


Fig. 7. Classification process on a bibliographic heterogeneous network [91].

Different from traditional classification researches, the classification problems studied in HIN have some new characteristics. First, the objects contained in HIN are different-typed, which means we can classify multiple types of objects simultaneously. Second, label knowledge can spread through various links among different-typed objects. Taking Fig. 7 as an example [91], four types of objects (paper, author, conference and term) are interconnected by multi-typed links. The classification process can be intuitively viewed as a process of knowledge propagation throughout the network, where the arrows indicate possible knowledge flow. In this HIN condition, the label of objects is decided by the effects of different-typed objects along different-typed links.

Many works extend traditional classification to heterogeneous information networks. Some

works extend transductive classification task, which is to predict labels for the given unlabeled data. For example, GNetMine [91] is proposed to model the link structure in information networks with arbitrary network schema and arbitrary number of object/link types. Recently, Luo et al. propose HetPathMine [92] to cluster with small labeled data on HIN through a novel meta path selection model, and Jacob et al. [93] propose a method to label nodes of different types by computing a latent representation of nodes in a space where two connected nodes tend to have close latent representations. Some works also extend inductive classification that is to construct a decision function in the whole data space. For example, Rossi et al. [94] use a bipartite heterogeneous network to represent textual document collections and propose IMBHN algorithm to induce a classification model assigning weights to textual terms.

Multi-label classification is prevalent in many real-world applications, where each example can be associated with a set of multiple labels simultaneously [37]. This kind of classification tasks are also extended to HIN. Angelova et al. [95] introduce a multi-label graph-based classification model for labeling heterogeneous networks by modeling the mutual influence between nodes as a random walk process. Kong et al. [37] use multiple types of relationships mined from the linkage structure of HIN to facilitate the multi-label classification process. Zhou et al. [96] propose an edge-centric multi-label classification approach considering both the structure affinity and the label vicinity.

As a unique characteristic, meta path is widely used in classification on HIN. Meta paths are usually used for feature generation in many methods, such as GNetMine [91] and HetPathMine [92]. Moreover, Kong et al. [16] introduce the concept of meta-path based dependencies among objects to study the collective classification problem.

Similar to clustering problem, classification is also integrated with other data mining tasks on HIN. Ranking-based classification is to integrate classification and ranking in a simultaneous, mutually enhancing process. Ji et al. [16] propose a ranking-based classification framework, RankClass, to perform more accurate analysis. As an extension of RankClass, Chen et al. [97] propose the F-RankClass for a unified classification framework that can be applied to binary or multi-class classification of unimodal or multimodal data. Some methods also integrate classification with information propagation. For example, Jendoubi et al. [98] classify the social message based on its spreading in the network and the theory of belief functions.

#### D. Link prediction

Link prediction is a fundamental problem in link mining that attempts to estimate the likelihood of the existence of a link between two nodes, based on observed links and the attributes of nodes. Link prediction is often viewed as a simple binary classification problem: for any two potentially linked objects, predict whether the link exists (1) or not (0). One kind of approach is to make this prediction entirely based on structural properties of the network. Liben-Nowell and Kleinberg [99] present a survey of predictors based on different graph proximity measures. Another kind of approach is to make use of attribute information for link prediction. For example, Popescul et al. [100] introduce a structured logistic regression model that can make use of relational features to predict the existence of links.

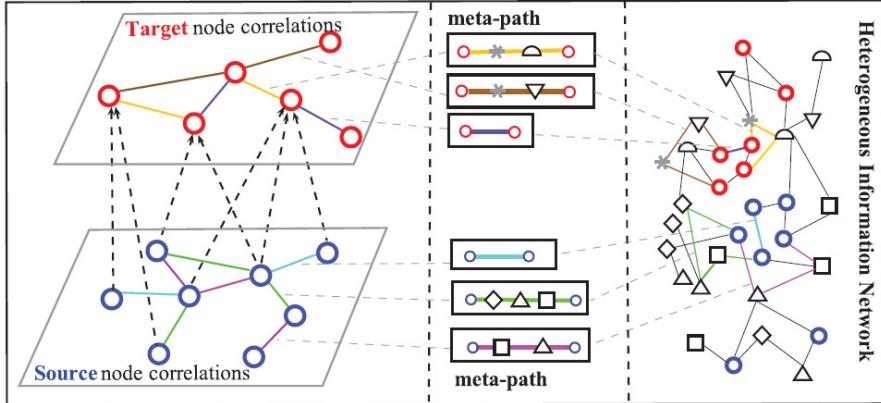


Fig. 8. Collective link prediction in heterogeneous information network [101].

Link prediction in an HIN has been an important research topic for recent years, which has the following characteristics. First, the links to be predicted are of different types, since objects in HIN are connected with different types of links. Second, there are dependencies existing among multiple types of links. So link prediction in an HIN needs to collectively predict multiple types of links by capturing the diverse and complex relationships among different types of links and leveraging the complementary prediction information. For example, Fig. 8 shows a collective prediction problem of multiple types of links in HIN [101]. Based on different meta paths, two objects have different link relations, and these relations have mutual effects.

Utilizing the meta path, many works employ a two-step process to solve link prediction

problem in HIN. The first step is to extract meta path based feature vectors, while the second step is to train a regression or classification model to compute the existence probability of a link [102], [103], [104], [101], [105]. For example, Sun et al. [102] propose PathPredict to solve the problem of co-author relationship prediction through meta path based feature extraction and logistic regression-based model. Zhang et al. [106] use meta path based features to predict organization chart or management hierarchy. Utilizing diverse and complex linkage information, Cao et al. [101] design a relatedness measure to construct the feature vectors of links and propose an iterative framework to predict multiple types of links collectively. In addition, Sun et al. [105] model the distribution of relationship building time with the use of the extracted topological features to predict when a certain relationship will be formed.

Probabilistic models are also widely applied for link prediction tasks in HIN. Yang et al. [25] propose a probabilistic method MRIP which models the influence propagating between heterogeneous relationships to predict links in multi-relational heterogeneous networks. Also, the TFGM model [107] defines a latent topic layer to bridge multiple networks and designs a semi-supervised learning model to mine competitive relationships across heterogeneous networks. Dong et al. [108] develop a transfer-based ranking factor graph model that combines several social patterns with network structure information for link prediction and recommendation. Matrix factorization is another common tool to handle link prediction problems. For example, Huang et al. [109] develop the joint manifold factorization (JMF) method to perform trust prediction with the ancillary rating matrix via aggregating heterogeneous social networks.

The approaches mentioned above mainly focus on link prediction on one single heterogeneous network. Recently, Zhang et al. [40], [110], [111] propose the problem of link prediction across multiple aligned heterogeneous networks. A two-phase link prediction method is put forward in [40]. The first phase is to extract heterogeneous features from multiple networks, while the second phase is to infer anchor links by formulating it as a stable matching problem. In addition, Zhang et al. [110] propose SCAN-PS to solve the social link prediction problem for new users using the “anchors”. Furthermore, they propose the TRAIL [111] method to predict social links and location links simultaneously. Also aimed at the cold start problem of new users, Liu et al. [112] propose the aligned factor graph model for user-user link prediction problem by utilizing information from another similar social network. In order to identify users from multiple heterogeneous social networks and integrate different networks, an energy-based model

COSNET [113] is proposed by considering both local and global consistency among multiple networks.

Most of the available works on link prediction are designed for static networks, however, the problem of dynamic link prediction is also very important and challenging. Taking into account both the dynamic and heterogeneous nature of web data, Zhao et al. [114] propose a general framework to characterize and predict community members from the evolution of heterogeneous web data. In order to solve the problem of dynamic link inference in temporal and heterogeneous information networks, Aggarwal et al. [115], [116] develop a two-level scheme which makes efficient macro- and micro-decisions for combining the topology and type information.

### E. Ranking

Ranking is an important data mining task in network analysis, which evaluates object importance or popularity based on some ranking functions. Many ranking methods have been proposed in homogeneous networks. For example, PageRank [117] evaluates the importance of objects through a random walk process, and HITS [118] ranks objects using the authority and hub scores. These approaches only consider the same type of objects in homogeneous networks.

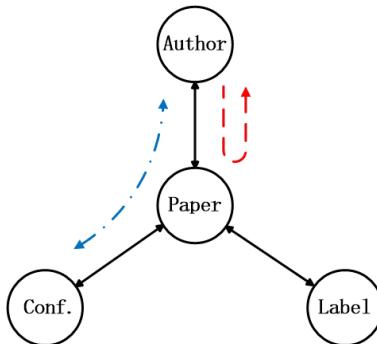


Fig. 9. An example of ranking in bibliographic heterogeneous network [21].

Ranking in heterogeneous information networks is an important and meaningful task, but faces several challenges. First, there are different types of objects and relations in HIN, and treating all objects equally will mix different types of objects together. Second, different types of objects and relations in HIN carry different semantic meanings, which may lead to different ranking results. Taking the bibliographic heterogeneous network in Fig. 9 as an example, ranking on authors

may have different results under different meta paths [21], since these meta paths will construct different link structures among authors. Moreover, the rankings of different-typed objects have mutual effects. For example, reputable authors usually publish papers on top conferences.

The co-ranking problem on bipartite graphs has been widely explored in the past decades. For example, Zhou et al. [119] co-rank authors and their publications by coupling two random walk processes, and co-HITS [120] incorporates the bipartite graph with the content information and the constraints of relevance. Soulier et al. [121] propose a bi-type entity ranking algorithm to rank jointly documents and authors in a bibliographic network regarding a topical query by combining content-based and network-based features. There are also some ranking works on multi-relational network. For example, MultiRank [122] is proposed to determine the importance of both objects and relations simultaneously for multi-relational data, and HAR [123] is proposed to determine hub and authority scores of objects and relevance scores of relations in multi-relational data for query search. These two methods focus on the same type of objects with multi-relations. Recently, Huang et al. [124] integrate both formal genre and inferred social networks with tweet networks to rank tweets. Although this work makes use of various types of objects in heterogeneous networks, it still ranks one type of objects.

Considering the characteristics of meta path on HIN, some works propose path based ranking methods. For example, Liu et al. [125] develop a publication ranking method with pseudo relevance feedback by leveraging a number of meta paths on the heterogeneous bibliographic graph. Applying the tensor analysis, Li et al. [21] propose HRank to simultaneously evaluate the importance of multiple types of objects and meta paths.

Ranking problem is also extended to HIN constructed by social media network. For image search in social media, Tsai et al. [126] propose SocialRank which uses social hints for image search and ranking in social networks. To identify high quality objects (questions, answers, and users) in Q&A systems, Zhang et al. [127] devise an unsupervised heterogeneous network based framework to co-rank multiple objects in Q&A sites. For heterogeneous cross-domain ranking problem, Wang et al. [128] propose a general regularized framework to discover a latent space for two domains and minimize two weighted ranking functions simultaneously in the latent space. Considering the dynamic nature of literature networks, a mutual reinforcement ranking framework is proposed to rank the future popularity of new publications and young researchers simultaneously [129].

### F. Recommendation

Recommender systems help consumers to make product recommendations that are likely to be of interest to the user such as books, movies, and restaurants. It uses a broad range of techniques from information retrieval, statistics, and machine learning to search for similarities among items and customer preferences. Traditional recommender systems normally only utilize the user-item rating feedback information for recommendation. Collaborative filtering is one of the most popular techniques, which includes two types of approaches: memory-based methods and model-based methods. Recently, matrix factorization has shown its effectiveness and efficiency in recommender systems, which factorizes the user-item rating matrix into two low rank user-specific and item-specific matrices, and then utilizes the factorized matrices to make further predictions [130]. With the prevalence of social media, more and more researchers study social recommender system, which utilizes social relations among users [131], [132].

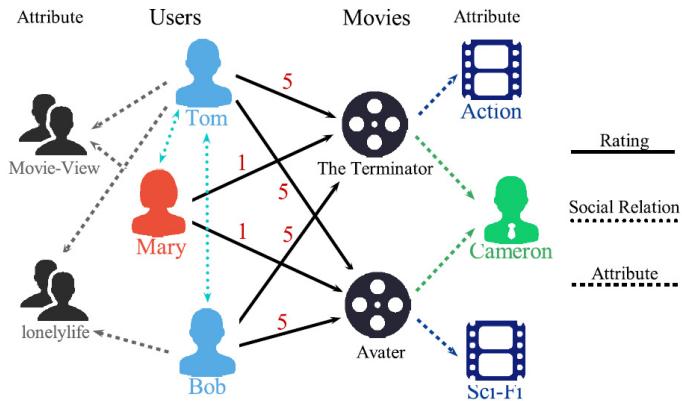


Fig. 10. An example of heterogeneous information network for movie recommendation [39].

Recently, some researchers have begun to be aware of the importance of heterogeneous information for recommendations. The comprehensive information and rich semantics of HIN make it promising to generate better recommendations. Fig. 10 shows such an example in movie recommendation [39]. The HIN not only contains different types of objects (e.g., users and movies) but also illustrates all kinds of relations among objects, such as viewing information, social relations, and attribute information. Constructing heterogeneous networks for recommendation can effectively fuse all kinds of information, which can be potentially utilized for

recommendation. Moreover, the objects and relations in the networks have different semantics, which can be explored to reveal subtle relations among objects.

Meta path is well used to explore the semantics and extract relations among objects. Shi et al. [34] implement a semantic-based recommendation system HeteRecom, which employs the semantics information of meta path to evaluate the similarities between movies. Furthermore, considering the attribute values, such as rating score on links, they model the recommender system as a weighted HIN and propose a semantic path based personalized recommendation method SemRec [39]. In order to take full advantage of the relationship heterogeneity, Yu et al. [35], [133] introduce meta-path-based latent features to represent the connectivity between users and items along different types of paths, and then define recommendation models at both global and personalized levels with Bayesian ranking optimization techniques. Also based on meta path, Burke et al. [134] present an approach for recommendation which incorporates multiple relations in a weighted hybrid.

A number of approaches employ heterogeneous information network to fuse various kinds of information. Utilizing different contexts information, Jamali et al. [31] propose a context-dependent matrix factorization model which considers a general latent factor for every entity and context-dependent latent factors for every context. Using user implicit feedback data, Yu et al. [35], [133] solve the global and personalized entity recommendation problem. Based on related interest groups, Ren et al. [135] propose a cluster-based citation recommendation framework to predict each query's citations in bibliographic networks. Similarly, Wu et al. [136] exploit graph summarization and content-based clustering for media recommendation with the interest group information. Based on multiple heterogeneous network features, Yang et al. [137] model multiple features into a unified framework with a SVM-Rank based method. In addition, using multiple types of relations, Luo et al. [138] propose a social collaborative filtering algorithm.

#### *G. Information Fusion*

Information fusion denotes the process of merging information from heterogeneous sources with differing conceptual, contextual and typographical representations. Due to the availability of various data sources, fusing these scattered distributed information sources has become an important research problem. In the past decades, dozens of papers have been published on this topic in many traditional data mining areas, e.g., data schemas integration in data ware-

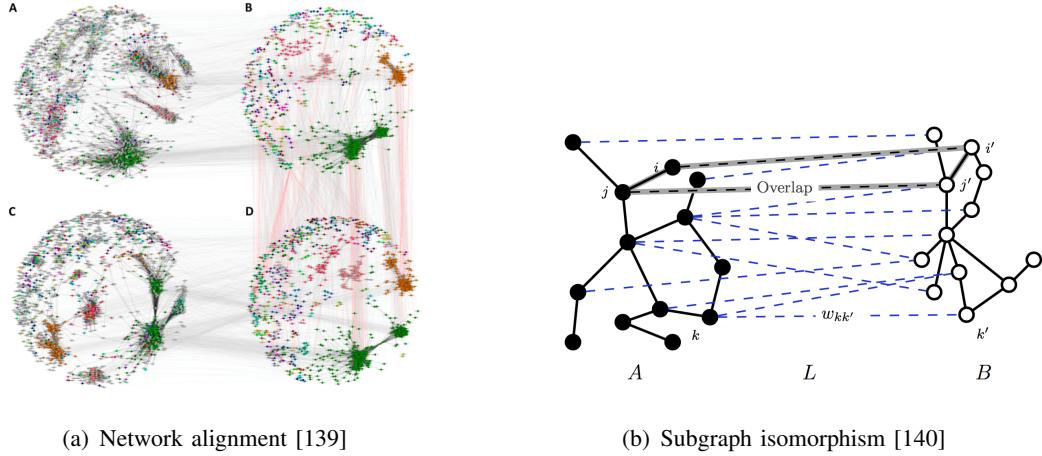


Fig. 11. An example of information network alignment.

house [141], protein-protein interaction (PPI) networks and gene regulatory networks matching in bioinformatics [142], [143], [144], [145], and ontology mapping in web semantics [146]. Nowadays, with the surge of HIN, information fusion across multiple HINs has become a novel yet important research problem. By fusing information from different HINs, we can obtain a more comprehensive and consistent knowledge about the common information entities shared in different HINs, including their structures, properties, and activities.

To fuse the information in multiple HINs, an important prerequisite will be to align the HINs via the shared common information entities, which can be users in social networks, authors in bibliographical networks, and protein molecules in biological networks. For instance, Fig. 4(a) is about the alignment of two heterogeneous social networks via the shared users, and Fig. 11(a) shows an example about alignment of two biological networks via molecules of common properties. Perfect HIN alignment is a challenging problem as the underlying subgraph isomorphism problem, as shown in Fig. 11(b), is actually NP-complete [147]. Meanwhile, based on the structure and attribute information available in HINs, a large number of approximated HIN alignment algorithms have been proposed so far. Enlightened by the homogeneous network alignment method in [148], Koutra et al. [149] propose to align two bipartite graphs with a fast network alignment algorithm. Zafarani et al. [87] propose to match users across social networks based on various node attributes, e.g., username, typing patterns and language patterns etc. Kong et al. [40] formulate the heterogeneous social network alignment problem as an anchor link

prediction problem. A two-step supervised method MNA is proposed in [40] to infer potential anchor links across networks with heterogeneous information in the networks. However, social networks in the real world are actually mostly partially aligned and lots of users are not anchor users. Zhang et al. have proposed the partial network alignment methods based on supervised learning setting and PU learning setting in [150] and [151] respectively. In addition to these pairwise social network alignment problems, multiple (more than two) social networks can be aligned simultaneously. Zhang et al. [152] discover that the inferred cross-network mapping of entities in social network alignment should meet the transitivity law and has an inherent one-to-one constraint. A new multiple social alignment framework is introduced in [152] to minimize the alignment costs and preserve the transitivity law and one-to-one constraint on the inferred mappings.

By fusing multiple HINs, the heterogeneous information available in each network can be transferred to other aligned networks and lots of application problems on HIN, e.g., link prediction and friend recommendation [153], [151], community detection [154], information diffusion [155], will benefit from it a lot.

Via the inferred mappings, Zhang et al. propose to transfer heterogeneous links across aligned networks to improve quality of predicted links/recommended friends [153], [151]. For new networks [111] and new users [110] with little social activity information, the transferred information can greatly overcome the cold start problem when predicting links for them. What's more, information about the shared entities across aligned networks can provide us with a more comprehensive knowledge about the community structures formed by them. By utilizing the information across multiple aligned networks, Zhang et al. [156] propose a new model to refine the clustering results of the shared entities with information in other aligned networks mutually. Jin et al. [157] propose a scalable framework to study the synergistic partitioning of multiple aligned large-scale networks, which takes the relationships among different networks into consideration and tries to maintain the consistency on partitioning the same nodes of different networks into the same partitions. Zhang et al. [154] study the community detection in emerging networks with information transferred from other aligned networks to overcome the cold start problem. In addition, by fusing multiple heterogeneous social networks, users in networks will be extensively connected with each other via both intra-network connections (e.g., friendship connections among users) and inter-network connections (i.e., the inferred mappings across

networks). As a result, information can reach more users and achieve broader influence across the aligned social networks. Zhan et al. propose a new model to study the information diffusion process across multiple aligned networks in [155].

#### *H. Other applications*

Besides the tasks discussed above, there are many other applications in heterogeneous networks, such as influence propagation and privacy risk problem. To quantitatively learn influence from heterogeneous networks, Liu et al. [158] first use a generative graphical model to learn the direct influence, and then use propagation methods to mine indirect and global influence. Using meta paths, Zhan et al. [155] propose a model M&M to solve the influence maximization problem in multiple partially aligned heterogeneous online social networks. For privacy risk in anonymized HIN, Zhang et al. [159] present a de-anonymization attack that exploits the identified vulnerability to prey upon the risk. Aiming at the inferior performances of unsupervised text embedding methods, Tang et al. [160] propose a semi-supervised representation learning method for text data, in which labeled information and different levels of word co-occurrence information are represented as a large-scale heterogeneous text network.

#### *I. Application Systems*

Besides many data mining tasks explored on heterogeneous network, some demo systems have designed prototype applications on HIN. Through employing a path-based relevance measure to evaluate the relevance between any-typed objects and capture the subtle semantic containing in meta path, Shi et al. [34] implement a HeteRecom system for semantic recommendation. Yu et al. [161] demonstrate a prototype system on query-driven discovery of semantically similar substructures in heterogeneous networks. Danilevsky et al. [162] present the AMETHYST system for exploring and analyzing a topical hierarchy constructed from an HIN. In LikeMiner system, Jin et al. [163] introduce a heterogeneous network model for social media with ‘likes’, and propose ‘like’ mining algorithms to estimate representativeness and influence of objects. Meanwhile, they design SocialSpamGuard [164], a scalable and online social media spam detection system for social network security. Taking DBLP as an example, Tao et al. [165] construct a Research-Insight system to demonstrate the power of database-oriented information network analysis including ranking, clustering, classification, recommendation, and prediction.

Furthermore, they construct a semi-structured news information network NewsNet and develop a NewsNetExplorer system [166] to provide a set of news information network exploration and mining functions.

Some real application systems also have been designed. One of the most famous works is ArnetMiner<sup>5</sup> [167], which offers comprehensive search and mining services for academic community. ArnetMiner not only provides abundant online academic services but also offers ideal test platform for heterogenous information network analysis. PatentMiner<sup>6</sup> [168] is another application which is a general topic-driven framework for analyzing and mining heterogeneous patent networks.

#### IV. ADVANCED TOPICS

Although many data mining tasks have been exploited in heterogeneous information network, it is still a young and promising research field. Here we illustrate some advanced topics, including challenging research issues and unexplored tasks, and point out some potential future research directions.

##### *A. More complex network construction*

There is a basic assumption in contemporary researches that a heterogeneous information network to be investigated is well-defined, and objects and links in the network are clean and unambiguous. However, it is not the case in real applications. In fact, constructing heterogeneous information network from real data often faces challenges.

If the networked data are structured data, like relational database, it may be easy to construct a heterogeneous information network with well-defined schema, such as DBLP network [14] and Movie network [34], [35]. However, even in this kind of heterogeneous network, objects and links can still be noisy. (1) Objects in a network may not exactly correspond to entities in real world, such as duplication of name [169] in bibliography data. That is, one object in a network may refer to multiple entities, or different objects may refer to the same entity. We can integrate entity resolution [170] with network mining to clean objects or links beforehand. For example, Shen et

<sup>5</sup><http://aminer.org/>

<sup>6</sup><http://pminer.org/home.do?m=home>

al. [171] propose a probabilistic model SHINE to link named entity mentions detected from the unstructured Web text with their corresponding entities existing in a heterogeneous information network. Ren et al. [172] propose a relation phrase-based entity recognition framework, called ClusType. The framework runs data-driven phrase mining to generate entity mention candidates and relation phrases, and enforces the principle that relation phrases should be softly clustered when propagating type information in a heterogeneous network constructed by argument entities. (2) Relations among objects may not be explicitly given or not complete sometimes, e.g., the advisor-advisee relationship in the DBLP network [173]. Link prediction [99] can be employed to fill out the missing relations for comprehensive networks. (3) Objects and links may not be reliable or trustable, e.g., the inaccurate item information in an E-commerce website and conflicting information of certain objects from multiple websites. So it is the first step to clean and integrate networked data for high-quality network construction, such as trustworthiness modeling [174], [175] and spam detection [176].

If the networked data are unstructured data, such as text data, multimedia data and multilingual data, it becomes more challenging to construct qualified heterogeneous information networks. In order to construct high-quality HINs, information extraction, natural language processing, and many other techniques should be integrated with network construction. Mining quality phrases is a critical step to form entities of networks from text data. Kishky et al. [177] propose a computationally efficient and effective model ToPMine, which first executes a phrase mining framework to segment a document into single and multi-word phrases, and then employs a new topic model that operates on the induced document partition. Furthermore, Liu et al. [178] propose an effective and scalable method SegPhrase+ that integrates quality phrases extraction with phrasal segmentation. Relationship extraction is another important step to form links among objects in network. Wang et al. [173] mine hidden advisor-advisee relationships from bibliographic data, and they further infer hierarchical relationships among partially ordered objects with heterogeneous attributes and links [179]. Broadly speaking, we can also extract entity and relationship to construct heterogeneous network from multimedia data and multilingual data, as we have done on text data.

### B. More powerful mining methods

For ubiquitous heterogeneous information networks, numbers of mining methods have been proposed on many data mining tasks. As we have said, heterogeneous information networks have two important characteristics: complex structure and rich semantics. According to these two characteristics, we summarize the contemporary works and point out future directions.

*1) Network Structure:* In heterogeneous network, objects can be organized in different forms. Bipartite graph is widely used to organize two types of objects and the relations among them [31], [32], [17]. As an extension of bipartite graphs,  $K$ -partite graphs [33] are able to represent multiple types of objects. Recently, heterogeneous networks are usually organized as star-schema networks, such as bibliographic data [20], [14], [13] and movie data [34], [35]. To combine the heterogeneous and homogeneous information, star schema with self loop is also proposed [38]. Different from only one hub object type existing in star schema network, some networked data have multiple hub object types, e.g., the bioinformatics data [80]. For this kind of networks, Shi et al. [80] propose a HeProjI method which projects a general heterogeneous network into a sequence of sub-networks with bipartite or star-schema structure.

In real applications, the networked data are usually more complex and irregular. Some real networks may contain attribute values on links, and these attribute values may contain important information. For example, users usually rate movies with a score from 1 to 5 in movie recommender system, where the rating scores represent users' attitudes to movies, and the "author of" relation between authors and papers in bibliographic networks can take values (e.g., 1, 2, 3) which means the order of authors in the paper. In this kind of applications, we need to consider the effect of attribute values on the weighted heterogeneous information network [39]. There are some time-series data, for example, a period of biographic data and rating information of users and movies. For this kind of data, we need to construct dynamic heterogeneous network [180] and consider the effect of time factor. In some applications, one kind of objects may exist in multiple heterogeneous networks [40], [110]. For example, users usually co-exist in multiple social networks, such as Facebook, Google+, and Twitter. In this kind of applications, we need to align users in different networks and effectively fuse information from different networks [150], [151], [152]. More broadly, many networked data are difficult to be modeled with heterogeneous network with a simple network schema. For example, in RDF data, there are so many types of

objects and relations, which cannot be described with network schema [181], [78]. Many research problems arise with this kind of schema-rich HINs, for example, management of objects and relations with so many types and automatic generation of meta paths. As the real networked data become more complex, we need to design more powerful and flexible heterogeneous networks, which also provides more challenges for data mining.

2) *Semantic Mining*: As a unique characteristic, objects and links in HIN contain rich semantics. Meta path can effectively capture subtle semantics among objects, and many works have exploited the meta path based mining tasks. For example, in similarity measure task, object pairs have different similarities under different meta paths [14], [13]; in recommendation task, different items will be recommended under different paths [39]. In addition, meta path is also widely used for feature extraction. Object similarity can be measured under different meta paths, which can be used as feature vectors for many tasks, such as clustering [15], link prediction [101], and recommendation [133].

However, some researchers have noticed the shortcomings of meta path. In some applications, meta path fails to capture more subtle semantics. For example, the “Author-Paper-Author” path describes the collaboration relation among authors. However, it cannot depict the fact that Philip S. Yu and Jiawei Han have many collaborations in data mining field but they seldom collaborate in information retrieval field. In order to overcome the shortcoming existing in meta path, Shi et al. [21] propose the constrained meta path concept, which can confine some constraints on objects. Taking Fig. 3(c) as an example, the constrained meta path  $APA|P.L = "Data\ Mining"$  represents the co-author relation of authors in data mining field through constraining the label of papers with “Data Mining”. Moreover, Liu et al. [125] propose the concept “restricted meta-path” which enables in-depth knowledge mining on the heterogeneous bibliographic networks by allowing restrictions on the node set. In addition, traditional HIN and meta path do not consider the attribute values on links, while weighted links are very common in real applications. Examples include rating scores between users and items in recommender system and the order of authors in papers in bibliographic network. Taking Fig. 10 as an example, the rating relation between users and movies can take scores from 1 to 5. Under the meta path “User-Movie-User”, Tom has the same similarity with Mary and Bob, but we can find that they may have totally different tastes due to different rating scores. Shi et al. [39] propose weighted meta path to consider attribute values on links and more subtly capture path semantics through distinguishing different

link attribute values. As an effective semantic capture tool, meta path has shown its power in semantic capture and feature selection. However, it may be coarse in some applications, so we need to extend traditional meta path for more subtle semantic capture. Broadly speaking, we can also design new, and more powerful semantic capture tools.

More importantly, the meta path approach faces challenges on path selection and their weight importances. How can we select meta paths in real applications? Theoretically, there are infinite meta paths in an HIN. In contemporary works, the network schema of HIN is usually small and simple, so we can assign some short and meaningful meta paths according to domain knowledge and experiences. Sun et al. [14] have validated that the long meta paths are not meaningful and they fail to produce good similarity measures. However, there is no work to study the effect of long meta paths on other mining tasks. In addition, there are so many meta paths even for short paths in some complex networks, like RDF network. It is a critical task to automatically extract meta paths in this condition. Recently, Meng et al. [181] study how to discover meta paths automatically which can best explain the relationship between node pairs. Another important issue is to automatically determine the weights of meta paths. Some methods have been proposed to explore this issue. For example, Lao et al. [50] employ a supervised method to learn weights, and Sun et al. [15] combine meta-path selection and user-guided information for clustering. Some interesting works are still worth doing. The ideal path weights learned should embody the importance of paths and reflect users' preferences. However, the similarity evaluations based on different paths have significant bias, which may make path weights hard to reflect path importances. So prioritized path weights are needed. In addition, if there are numerous meta paths in real applications (e.g., RDF network), the path weight learning will be more important and challenging.

In Fig. 12, we summarize some typical works in the HIN field from two perspectives: network structure and semantic exploration. We respectively select several typical works from six mining tasks mentioned above, and put these works in a coordinate according to network structure and semantics explored in these works. Note that we denominate those un-named methods with the first letter of keywords in the title, such as UGES [59] and CPIH [103]. Along the X-axis, the network structure becomes more complex, and semantics information becomes richer along the Y-axis. For example, RankClus [17] is designed for bipartite networks and only captures link semantics (different-typed links contain different semantics). While PathSim [14] can deal

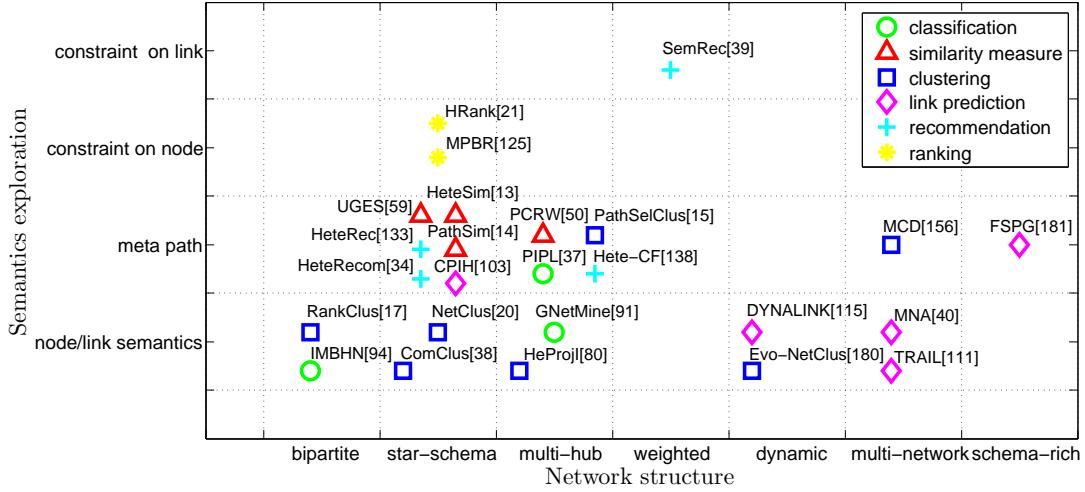


Fig. 12. Summarization of typical works on HIN according to network structure and semantic exploration.

with more complex star-schema networks and use meta path to mine deeper semantics. Further, SemRec [39] adds constraints to links to explore more subtle semantic information in a weighted HIN. From the figure, we can also find that most contemporary works focus on simple network structures (e.g., bipartite or star schema networks) and primary semantic exploration (e.g., meta path). In the future, we can exploit more complex heterogeneous networks with more powerful semantics capture tools.

### C. Bigger networked data

In order to illustrate the benefits of HIN, we need to design data mining algorithms on big networked data in wider domains. The variety is an important characteristic of big data. HIN is a powerful tool to handle the variety of big data, since it can flexibly and effectively integrate varied objects and heterogeneous information. However, it is non-trivial work to build a real HIN based analysis system. Besides research challenges mentioned above, such as network construction, it will face many practical technique challenges. Real HIN is huge, even dynamic, so it usually cannot be contained in memory and cannot be handled directly. We know that a user at a time could be only interested in a tiny portion of nodes, links, or sub-networks. Instead of directly mining the whole network, we can mine hidden but small networks “extracted” dynamically from

some existing networks, based on user-specified constraints or expected node/link behaviors. How to discover such hidden networks and mine knowledge (e.g., clusters, behaviors, and anomalies) from such hidden but non-isolated networks could be an interesting but challenging problem.

Most of contemporary data mining tasks on HIN only work on small dataset, and fail to consider the quick and parallel process on big data. Some research works have begun to consider the quick computation of mining algorithms on HIN. For example, Sun et al. [14] design a co-clustering based pruning strategy to fasten the processing speed of PathSim. Lao et al. [51] propose the quick computation strategies of PCRW, and Shi et al. [52], [54] also consider the quick/parallel computation of HeteSim. In addition, cloud computing also provides an option to handle big networked data. Although parallel graph mining algorithms [182] and platforms [183] have been proposed, parallel HIN analysis methods face some unique challenges. For example, the partition of HIN needs to consider the overload balances of computing nodes, as well as balances of different-typed nodes. Moreover, it is also challenging to mine integrated path semantics in partitioned subgraphs.

#### *D. More applications*

Due to unique characteristics of HIN, many data mining tasks have been explored on HIN, which are summarized as above. In fact, more data mining tasks can be studied on HIN. Here we introduce two potential applications.

The online analytical processing (OLAP) has shown its power in multidimensional analysis of structured, relational data [184]. The similar analysis can also be done, when we view a heterogeneous information network from different angles and at different levels of granularity. Taking a bibliographic network as an example, we can observe the change of published papers on a conference in the time or district dimension, when we designate papers and conferences as the object types and publish relations as the link type. Some preliminary studies have been done on this issue. Zhao et al. [185] introduce graph cube to support OLAP queries effectively on large multidimensional networks; Li et al. [186] design InfoNetOLAPer to provide topic-oriented, integrated, and multidimensional organizational solutions for information networks. Yin et al. [187] have developed a novel HMGraph OLAP framework to mine multi-dimensional heterogeneous information networks with more dimensions and operations. These works consider link relation as a measure. However, they usually ignore semantic information in heterogeneous

networks determined by multiple nodes and links. So the study of online analytical processing of heterogeneous information networks is still worth exploring.

Information diffusion is a vast research domain and has attracted research interests from many fields, such as physics, biology, etc. Traditional information diffusion is studied on homogeneous networks [188], where information is propagated in one single channel. However, in many applications, pieces of information or diseases are propagated among different types of objects. For example, diseases could propagate among people, different kinds of animals and food, via different channels. Few works explore this issue. Liu et al. [189] propose a generative graphical model which utilizes the heterogeneous link information and the textual content associated with each node to mine topic level direct influence. In order to capture better spreading models that represent the real world patterns, it is desirable to pay more attention to the study of information diffusion in heterogeneous information networks.

## V. CONCLUSION

There is a surge on heterogenous information network analysis in recent years because of rich structural and semantic information in this kind of networks. This paper provides an extensive survey in this rapidly expanding field. We present the recent developments of different data mining tasks on heterogenous information network, along with future development directions. Hopefully, this survey will give researchers an understanding of the fundamental issues and a good starting point to work on this field.

## REFERENCES

- [1] J. Han, "Mining heterogeneous information networks by exploring the power of links," in *Discovery Science*, 2009, pp. 13–30.
- [2] Y. Sun and J. Han, "Mining heterogeneous information networks: a structural analysis approach," *SIGKDD Explorations*, vol. 14, no. 2, pp. 20–28, 2012.
- [3] L. Getoor and C. P. Diehl, "Link mining: a survey," *SIGKDD Explorations*, vol. 7, no. 2, pp. 3–12, 2005.
- [4] D. Jensen and H. Goldberg, *AAAI Fall Symposium on AI and Link Analysis*. AAAI Press, 1998.
- [5] R. Feldman, "Link analysis: Current state of the art," *Tutorial at the KDD*, vol. 2, 2002.
- [6] S. Wasserman, *Social network analysis: Methods and applications*. Cambridge university press, 1994.
- [7] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of information Science*, vol. 28, no. 6, pp. 441–453, 2002.
- [8] S. Chakrabarti *et al.*, *Mining the Web: Analysis of hypertext and semi structured data*. Morgan Kaufmann, 2002.
- [9] T. G. Lewis, *Network science: Theory and applications*. John Wiley & Sons, 2011.
- [10] D. J. Cook and L. B. Holder, "Graph-based data mining," *IEEE Intelligent Systems*, vol. 15, no. 2, pp. 32–41, 2000.
- [11] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *KDD*, 2010, pp. 243–252.
- [12] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *KDD*, 2010, pp. 393–402.
- [13] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu, "Relevance search in heterogeneous networks," in *EDBT*, 2012, pp. 180–191.
- [14] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," in *VLDB*, 2011, pp. 992–1003.
- [15] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," in *KDD*, 2012, pp. 1348–1356.
- [16] X. Kong, P. S. Yu, Y. Ding, and D. J. Wild, "Meta path-based collective classification in heterogeneous information networks," in *CIKM*, 2012, pp. 1567–1571.
- [17] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "RankClus: integrating clustering with ranking for heterogeneous information network analysis," in *EDBT*, 2009, pp. 565–576.
- [18] Y. Sun and J. Han, *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers, 2012.
- [19] Y. Sun and J. Han, "Meta-path-based search and mining in heterogeneous information networks," *Tsinghua Science and Technology*, vol. 18, no. 4, pp. 329–338, 2013.
- [20] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *KDD*, 2009, pp. 797–806.
- [21] Y. Li, C. Shi, S. Y. Philip, and Q. Chen, "Hrank: A path based ranking method in heterogeneous information network," in *WAIM*, 2014, pp. 553–565.
- [22] J. Tang, H. Gao, X. Hu, and H. Liu, "Exploiting homophily effect for trust prediction," in *WSDM*, 2013, pp. 53–62.
- [23] I. Konstas, V. Stathopoulo, and J. M. Jose, "On social networks and collaborative recommendation," in *SIGIR*, 2009, pp. 195–202.
- [24] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM*, 2003, pp. 556–559.

- [25] Y. Yang, N. V. Chawla, Y. Sun, and J. Han, "Predicting links in multi-relational and heterogeneous networks," in *ICDM*, 2012, pp. 755–764.
- [26] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, "Community evolution in dynamic multi-mode networks," in *KDD*, 2008, pp. 677–685.
- [27] E. Zhong, W. Fan, J. Wang, L. Xiao, and Y. Li, "Comsoc: Adaptive transfer of user behaviors over composite social network," in *KDD*, 2012, pp. 696–704.
- [28] E. Zhong, W. Fan, Y. Zhu, and Q. Yang, "Modeling the dynamics of composite social networks," in *KDD*, 2013, pp. 937–945.
- [29] J. Kim and T. Wilhelm, "What is a complex graph?" *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 11, pp. 2637–2652, 2008.
- [30] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.
- [31] M. Jamali and L. Lakshmanan, "Heteromf: recommendation in heterogeneous information networks using context dependent factor models," in *WWW*, 2013, pp. 643–654.
- [32] B. Long, Z. M. Zhang, and P. S. Yu, "Co-clustering by block value decomposition," in *KDD*, 2005, pp. 635–640.
- [33] B. Long, X. Wu, Z. Zhang, and P. S. Yu, "Unsupervised learning on k-partite graphs," in *KDD*, 2006, pp. 317–326.
- [34] C. Shi, C. Zhou, X. Kong, P. S. Yu, G. Liu, and B. Wang, "Heterocom: a semantic-based recommendation system in heterogeneous networks," in *KDD*, 2012, pp. 1552–1555.
- [35] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han, "Recommendation in heterogeneous information networks with implicit user feedback," in *RecSys*, 2013, pp. 347–350.
- [36] H. Zhuang, J. Zhang, G. Brova, J. Tang, H. Cam, X. Yan, and J. Han, "Mining query-based subnetwork outliers in heterogeneous information networks," in *ICDM*, 2014, pp. 1127–1132.
- [37] X. Kong, B. Cao, and P. S. Yu, "Multi-label classification by mining label and instance correlations from heterogeneous information networks," in *KDD*, 2013, pp. 614–622.
- [38] R. Wang, C. Shi, P. S. Yu, and B. Wu, "Integrating clustering and ranking on hybrid heterogeneous information network," in *PAKDD*, 2013, pp. 583–594.
- [39] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu, "Semantic path based personalized recommendation on weighted heterogeneous information networks," in *CIKM*, 2015.
- [40] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *CIKM*, 2013, pp. 179–188.
- [41] A. Singhal, "Introducing the knowledge graph: things, not strings," *Official Google Blog*, 2012.
- [42] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, 2007, pp. 697–706.
- [43] L. Zou, M. T. Özsü, L. Chen, X. Shen, R. Huang, and D. Zhao, "gstore: a graph-based SPARQL query engine," *VLDB Journal*, vol. 23, no. 4, pp. 565–590, 2014.
- [44] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.
- [45] G. Jeh and J. Widom, "Scaling personalized web search," in *WWW*, 2003, pp. 271–279.
- [46] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *KDD*, 2002, pp. 538–543.
- [47] J. He, J. Bailey, and R. Zhang, "Exploiting transitive similarity and temporal dynamics for similarity search in heterogeneous information networks," *Database Systems for Advanced Applications*, vol. 8422, pp. 141–155, 2014.

- [48] L. Hou U., K. Yao, and H. Mak, "Pathsimext: Revisiting pathsim in heterogeneous information networks," in *WAIM*, 2014, pp. 38–42.
- [49] Y. Xiong, Y. Zhu, and P. S. Yu, "Top-k similarity join in heterogeneous information networks," *IEEE Transactions on Knowledge & Data Engineering*, vol. 27, no. 6, pp. 1710–1723, 2015.
- [50] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine Learning*, vol. 81, no. 2, pp. 53–67, 2010.
- [51] N. Lao and W. Cohen, "Fast query execution for retrieval models based on path-constrained random walks," in *KDD*, 2010, pp. 881–888.
- [52] C. Shi, X. Kong, Y. Huang, S. Y. Philip, and B. Wu, "Hetesim: A general framework for relevance measure in heterogeneous networks," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 10, pp. 2479–2492, 2014.
- [53] C. Li, J. Sun, Y. Xiong, and G. Zheng, "An efficient drug-target interaction mining algorithm in heterogeneous biological networks," in *PAKDD*, 2014, pp. 65–76.
- [54] X. Meng, C. Shi, Y. Li, and B. W. Lei Zhang, "Relevance measure in large-scale heterogeneous networks," in *APWeb*, 2014, pp. 636–643.
- [55] S. Bu, X. Hong, Z. Peng, and Q. Li, "Integrating meta-path selection with user-preference for top-k relevant search in heterogeneous information networks," in *CSCWD*, 2014, pp. 301–306.
- [56] M. Zhu, T. Zhu, Z. Peng, G. Yang, Y. Xu, S. Wang, X. Wang, and X. Hong, "Relevance search on signed heterogeneous information network based on meta-path factorization," in *WAIM*, 2015, pp. 181–192.
- [57] G. Wang, Q. Hu, and P. S. Yu, "Influence and similarity on heterogeneous networks," in *CIKM*, 2012, pp. 1462–1466.
- [58] C. Wang, R. Raina, D. Fong, D. Zhou, J. Han, and G. J. Badros, "Learning relevance from heterogeneous social network and its application in online targeting," in *SIGIR*, 2011, pp. 655–664.
- [59] X. Yu, Y. Sun, B. Norick, T. Mao, and J. Han, "User guided entity similarity search using meta-path selection in heterogeneous information networks," in *CIKM*, 2012, pp. 2025–2029.
- [60] M. Zhang, H. Hu, Z. He, and W. Wang, "Top-k similarity search in heterogeneous information networks with x-star network schema," *Expert Systems with Applications*, vol. 42, no. 2, pp. 699–712, 2015.
- [61] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [62] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
- [63] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.
- [64] U. V. Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [65] K. Wakita and T. Tsurumi, "Finding community structure in mega-scale social networks," in *WWW*, 2007, pp. 1275–1276.
- [66] M. Sales-Pardo, R. Guimera, A. Moreira, and L. Amaral, "Extracting the hierarchical organization of complex systems," *Proceedings of the National Academy of Sciences*, vol. 104, no. 39, pp. 15 224–15 229, 2007.
- [67] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," in *VLDB*, 2009, pp. 718–729.
- [68] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *KDD*, 2009, pp. 927–936.
- [69] C. Aggarwal, Y. Xie, and P. Yu, "Towards community detection in locally heterogeneous networks," in *SDM*, 2011, pp. 391–402.

- [70] Y. Sun, C. Aggarwal, and J. Han, “Relation strength-aware clustering of heterogeneous information networks with incomplete attributes,” in *VLDB*, 2012, pp. 394–405.
- [71] G. J. Qi, C. C. Aggarwal, and T. S. Huang, “On clustering heterogeneous social media objects with outlier links,” in *WSDM*, 2012, pp. 553–562.
- [72] J. D. Cruz, C. Bothorel, and F. Poulet, “Integrating heterogeneous information within a social network for detecting communities,” in *ASONAM*, 2013, pp. 1453–1454.
- [73] B. Boden, M. Ester, and T. Seidl, “Density-based subspace clustering in heterogeneous networks,” in *ECML/PKDD*, 2014, pp. 149–164.
- [74] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin, “Probabilistic topic models with biased propagation on heterogeneous information networks,” in *KDD*, 2011, pp. 1271–1279.
- [75] H. Deng, B. Zhao, and J. Han, “Collective topic modeling for heterogeneous networks,” in *SIGIR*, 2011, pp. 1109–1110.
- [76] Q. Wang, Z. Peng, F. Jiang, and Q. Li, “Lsa-ptm: A propagation-based topic model using latent semantic analysis on heterogeneous information networks,” in *WAIM*, vol. 7923, 2013, pp. 13–24.
- [77] Q. Wang, Z. Peng, S. Wang, P. S. Yu, Q. Li, and X. Hong, “clutm: Content and link integrated topic model on heterogeneous information networks,” in *WAIM*, 2015, pp. 207–218.
- [78] C. Wang, Y. Song, A. El-Kishky, D. Roth, M. Zhang, and J. Han, “Incorporating world knowledge to document clustering via heterogeneous information networks,” in *KDD*, 2015, pp. 1215–1224.
- [79] C. Luo, W. Pang, and Z. Wang, “Semi-supervised clustering on heterogeneous information networks,” *Advances in Knowledge Discovery and Data Mining*, vol. 8444, pp. 548–559, 2014.
- [80] C. Shi, R. Wang, Y. Li, P. S. Yu, and B. Wu, “Ranking-based clustering on general heterogeneous information networks by network projection,” in *CIKM*, 2014, pp. 699–708.
- [81] J. Chen, W. Dai, Y. Sun, and J. Dy, “Clustering and ranking in heterogeneous information networks via gamma-poisson model,” in *SDM*, 2015, pp. 425–432.
- [82] C. Wang, M. Danilevsky, J. Liu, N. Desai, H. Ji, and J. Han, “Constructing topical hierarchies in heterogeneous information networks,” in *ICDM*, 2013, pp. 767–776.
- [83] C. Qiu, W. Chen, T. Wang, and K. Lei, “Overlapping community detection in directed heterogeneous social network,” in *WAIM*, 2015, pp. 490–493.
- [84] M. Gupta, J. Gao, and J. Han, “Community distribution outlier detection in heterogeneous information networks,” in *ECML*, 2013, pp. 557–573.
- [85] M. Gupta, J. Gao, X. Yan, H. Cam, and J. Han, “On detecting association-based clique outliers in heterogeneous information networks,” in *ASONAM*, 2013, pp. 108–115.
- [86] J. Kuck, H. Zhuang, X. Yan, H. Cam, and J. Han, “Query-based outlier detection in heterogeneous information networks,” in *EDBT*, 2015, pp. 325–336.
- [87] Y. Zhou and L. Liu, “Social influence based clustering of heterogeneous information networks,” in *KDD*, 2013, pp. 338–346.
- [88] F. Alqadah and R. Bhatnagar, “A game theoretic framework for heterogenous information network clustering,” in *KDD*, 2011, pp. 795–804.
- [89] S. Chakrabarti, B. Dom, and P. Indyk, “Enhanced hypertext categorization using hyperlinks,” *ACM SIGMOD Record*, vol. 27, no. 2, pp. 307–318, 1998.

- [90] J. Lafferty, A. McCallum, and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *ICML*, 2001, pp. 282–289.
- [91] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, “Graph regularized transductive classification on heterogeneous information networks,” in *ECML/PKDD*, 2010, pp. 570–586.
- [92] C. Luo, R. Guan, Z. Wang, and C. Lin, “Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks,” *Advances in Information Retrieval*, vol. 8416, pp. 210–221, 2014.
- [93] Y. Jacob, L. Denoyer, and P. Gallinari, “Learning latent representations of nodes for classifying in heterogeneous social networks,” in *WSDM*, 2014, pp. 373–382.
- [94] R. G. Rossi, T. de Paulo Faleiros, A. de Andrade Lopes, and S. O. Rezende, “Inductive model generation for text categorization using a bipartite heterogeneous network,” in *ICDM*, 2012, pp. 1086–1091.
- [95] R. Angelova, G. Kasneci, and G. Weikum, “Graffiti: graph-based classification in heterogeneous networks,” in *WWW*, 2012, pp. 139–170.
- [96] Y. Zhou and L. Liu, “Activity-edge centric multi-label classification for mining heterogeneous information networks,” in *KDD*, 2014, pp. 1276–1285.
- [97] S. D. Chen, Y. Y. Chen, J. Han, and P. Moulin, “A feature-enhanced ranking-based classifier for multimodal data and heterogeneous information networks,” in *ICDM*, 2013, pp. 997–1002.
- [98] S. Jendoubi, A. Martin, L. Lietard, and B. B. Yaghlane, “Classification of message spreading in a heterogeneous social network,” in *IPMU*, 2014, pp. 66–75.
- [99] 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.
- [100] A. Popescul and L. H. Ungar, “Statistical relational learning for link prediction,” in *IJCAI workshop on learning statistical models from relational data*, 2003.
- [101] B. Cao, X. Kong, and P. S. Yu, “Collective prediction of multiple types of links in heterogeneous information networks,” in *ICDM*, 2014, pp. 50–59.
- [102] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, “Co-author relationship prediction in heterogeneous bibliographic networks,” in *ASONAM*, 2011, pp. 121–128.
- [103] X. Yu, Q. Gu, M. Zhou, and J. Han, “Citation prediction in heterogeneous bibliographic networks,” in *SDM*, 2012, pp. 1119–1130.
- [104] J. Chen, H. Gao, Z. Wu, and D. Li, “Tag co-occurrence relationship prediction in heterogeneous information networks,” in *ICPADS*, 2013, pp. 528–533.
- [105] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, “When will it happen?: relationship prediction in heterogeneous information networks,” in *WSDM*, 2012, pp. 663–672.
- [106] J. Zhang, S. Y. Philip, and Y. Lv, “Organizational chart inference,” in *KDD*, 2015, pp. 1435–1444.
- [107] Y. Yang, J. Tang, J. Keomany, Y. Zhao, J. Li, Y. Ding, T. Li, and L. Wang, “Mining competitive relationships by learning across heterogeneous networks,” in *CIKM*, 2012, pp. 1432–1441.
- [108] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, “Link prediction and recommendation across heterogeneous social networks,” in *ICDM*, 2012, pp. 181–190.
- [109] J. Huang, F. Nie, H. Huang, and Y. C. Tu, “Trust prediction via aggregating heterogeneous social networks,” in *CIKM*, 2012, pp. 1774–1778.

- [110] J. Zhang, X. Kong, and P. S. Yu, "Predicting social links for new users across aligned heterogeneous social networks," in *ICDM*, 2013, pp. 1289–1294.
- [111] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *WSDM*, 2014, pp. 303–312.
- [112] F. Liu and S.-T. Xia, "Link prediction in aligned heterogeneous networks," in *PAKDD*, 2015, pp. 33–44.
- [113] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "Cosnet: Connecting heterogeneous social networks with local and global consistency," in *KDD*, 2015, pp. 1485–1494.
- [114] Q. Zhao, S. S. Bhowmick, X. Zheng, and K. Yi, "Characterizing and predicting community members from evolutionary and heterogeneous networks," in *CIKM*, 2008, pp. 309–318.
- [115] C. C. Aggarwal, Y. Xie, and S. Y. Philip, "On dynamic link inference in heterogeneous networks," in *SDM*, 2012, pp. 415–426.
- [116] C. C. Aggarwal, Y. Xie, and P. S. Yu, "A framework for dynamic link prediction in heterogeneous networks," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 7, no. 1, pp. 14–33, 2014.
- [117] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Technical report, Stanford University Database Group, 1998.
- [118] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," in *SODA*, 1999, pp. 668–677.
- [119] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *ICDM*, 2007, pp. 739–744.
- [120] H. Deng, M. R. Lyu, and I. King, "A generalized co-hits algorithm and its application to bipartite graphs," in *KDD*, 2009, pp. 239–248.
- [121] L. Soulier, L. B. Jabeur, L. Tamine, and W. Bahsoun, "On ranking relevant entities in heterogeneous networks using a language-based model," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, pp. 500–515, 2013.
- [122] M. K. NG, X. Li, and Y. Ye, "Multirank: Co-ranking for objects and relations in multi-relational data," in *KDD*, 2011, pp. 1217–1225.
- [123] X. Li, M. K. Ng, and Y. Ye, "Har: Hub, authority and relevance scores in multi-relational data for query search," in *SDM*, 2012, pp. 141–152.
- [124] H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. K. Le, T. F. Abdelzaher, J. Han, A. Leung, J. P. Hancock *et al.*, "Tweet ranking based on heterogeneous networks," in *COLING*, 2012, pp. 1239–1256.
- [125] X. Liu, Y. Yu, C. Guo, and Y. Sun, "Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation," in *CIKM*, 2014, pp. 121–130.
- [126] M. H. Tsai, C. Aggarwal, and T. Huang, "Ranking in heterogeneous social media," in *WSDM*, 2014, pp. 613–622.
- [127] J. Zhang, X. Kong, L. Jie, Y. Chang, and S. Y. Philip, "Ncr: A scalable network-based approach to co-ranking in question-and-answer sites," in *CIKM*, 2014, pp. 709–718.
- [128] B. Wang, J. Tang, W. Fan, S. Chen, C. Tan, and Z. Yang, "Query-dependent cross-domain ranking in heterogeneous network," *Knowledge and information systems*, vol. 34, no. 1, pp. 109–145, 2013.
- [129] S. Wang, S. Xie, X. Zhang, Z. Li, P. S. Yu, and X. Shu, "Future influence ranking of scientific literature," in *SIAM*, 2014, pp. 749–757.
- [130] N. Srebro, T. Jaakkola *et al.*, "Weighted low-rank approximations," in *ICML*, 2003, pp. 720–727.
- [131] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," in *SIGIR*, 2009, pp. 203–210.

- [132] X. Yang, H. Steck, and Y. Liu, “Circle-based recommendation in online social networks,” in *KDD*, 2012, pp. 1267–1275.
- [133] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han, “Personalized entity recommendation: A heterogeneous information network approach,” in *WSDM*, 2014, pp. 283–292.
- [134] R. Burke, F. Vahedian, and B. Mobasher, “Hybrid recommendation in heterogeneous networks,” in *UMAP*, 2014, pp. 49–60.
- [135] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, “Cluscite: Effective citation recommendation by information network-based clustering,” in *KDD*, 2014, pp. 821–830.
- [136] J. Wu, L. Chen, Q. Yu, P. Han, and Z. Wu, “Trust-aware media recommendation in heterogeneous social networks,” *World Wide Web*, vol. 18, no. 1, pp. 139–157, 2015.
- [137] C. Yang, J. Sun, J. Ma, S. Zhang, G. Wang, and Z. Hua, “Scientific collaborator recommendation in heterogeneous bibliographic networks,” in *HICSS*, 2015, pp. 552–561.
- [138] C. Luo, W. Pang, Z. Wang, and C. Lin, “Hete-cf: Social-based collaborative filtering recommendation using heterogeneous relations,” in *ICDM*, 2014, pp. 917–922.
- [139] S. P. Ficklin and F. A. Feltus, “Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice,” *Plant Physiology*, vol. 156, no. 3, pp. 1244–1256, 2011.
- [140] A. M. Khan, D. F. Gleich, A. Pothen, and M. Halappanavar, “A multithreaded algorithm for network alignment via approximate matching,” in *High Performance Computing, Networking, Storage and Analysis*, 2012, pp. 1–11.
- [141] S. Melnik, H. Garcia-Molina, and E. Rahm, “Similarity flooding: A versatile graph matching algorithm and its application to schema matching,” in *ICDE*, 2002, pp. 117–128.
- [142] M. Kalaev, V. Bafna, and R. Sharan, “Fast and accurate alignment of multiple protein networks,” in *Research in Computational Molecular Biology*, 2008, pp. 246–256.
- [143] Y.-K. Shih and S. Parthasarathy, “Scalable global alignment for multiple biological networks,” *BMC bioinformatics*, vol. 13, no. Suppl 3, p. S11, 2012.
- [144] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, “Isorankn: spectral methods for global alignment of multiple protein networks,” *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.
- [145] R. Singh, J. Xu, and B. Berger, “Pairwise global alignment of protein interaction networks by matching neighborhood topology,” in *Research in computational molecular biology*, 2007, pp. 16–31.
- [146] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, “Ontology matching: A machine learning approach,” in *Handbook on ontologies*, 2004, pp. 385–403.
- [147] G. W. Klau, “A new graph-based method for pairwise global network alignment,” *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S59, 2009.
- [148] S. Umeyama, “An eigendecomposition approach to weighted graph matching problems,” *IEEE TPAMI*, vol. 10, no. 5, pp. 695–703, 1988.
- [149] D. Koutra, H. Tong, and D. Lubensky, “Big-align: Fast bipartite graph alignment,” in *ICDM*, 2013, pp. 389–398.
- [150] J. Zhang, W. Shao, S. Wang, X. Kong, and P. Yu, “Partial network alignment with anchor meta path and truncated generalized stable matching,” in *IRI*, 2015.
- [151] J. Zhang and P. Yu, “Integrated anchor and social link predictions across social networks,” in *IJCAI*, 2015.
- [152] J. Zhang and P. S. Yu, “Multiple anonymized social networks alignment,” in *ICDM*, 2015.
- [153] J. Zhang, P. S. Yu, and Z.-H. Zhou, “Meta-path based multi-network collective link prediction,” in *KDD*, 2014, pp. 1286–1295.

- [154] J. Zhang and P. Yu, "Community detection for emerging networks," in *SDM*, 2015.
- [155] Q. Zhan, J. Zhang, S. Wang, S. Y. Philip, and J. Xie, "Influence maximization across partially aligned heterogeneous social networks," in *PAKDD*, 2015, pp. 58–69.
- [156] J. Zhang and P. Yu, "McD: Mutual clustering across multiple heterogeneous networks," in *IEEE BigData Congress*, 2015.
- [157] S. Jin, J. Zhang, P. S. Yu, S. Yang, and A. Li, "Synergistic partitioning in multiple large scale social networks," in *IEEE BigData*, 2014, pp. 281–290.
- [158] L. Liu, J. Tang, J. Han, and S. Yang, "Learning influence from heterogeneous social networks," *Data Mining and Knowledge Discovery*, vol. 25, no. 3, pp. 511–544, 2012.
- [159] A. Zhang, X. Xie, K. C. C. Chang, C. A. Gunter, J. Han, and X. Wang, "Privacy risk in anonymized heterogeneous information networks," in *EDBT*, 2014, pp. 595–606.
- [160] J. Tang, M. Qu, and Q. Mei, "Pte: Predictive text embedding through large-scale heterogeneous text networks," in *KDD*, 2015, pp. 1165–1174.
- [161] X. Yu, Y. Sun, P. Zhao, and J. Han, "Query-driven discovery of semantically similar substructures in heterogeneous networks," in *KDD*, 2012, pp. 1500–1503.
- [162] M. Danilevsky, C. Wang, F. Tao, S. Nguyen, G. Chen, N. Desai, L. Wang, and J. Han, "Amethyst: A system for mining and exploring topical hierarchies of heterogeneous data," in *KDD*, 2013, pp. 1458–1461.
- [163] X. Jin, C. Wang, J. Luo, X. Yu, and J. Han, "Likeminer: a system for mining the power of 'like' in social media networks," in *KDD*, 2011, pp. 753–756.
- [164] X. Jin, C. X. Lin, J. Luo, and J. Han, "Socialspamguard: A data mining-based spam detection system for social media networks," in *PVLDB*, vol. 4, no. 12, 2011, pp. 1458–1461.
- [165] F. Tao, X. Yu, K. H. Lei, G. Brova, X. Cheng, J. Han, R. Kanade, Y. Sun, C. Wang, L. Wang *et al.*, "Research-insight: Providing insight on research by publication network analysis," in *SIGMOD*, 2013, pp. 1093–1096.
- [166] F. Tao, G. Brova, J. Han, H. Ji, C. Wang, B. Norick, A. El-Kishky, J. Liu, X. Ren, and Y. Sun, "Newsnetexplorer: automatic construction and exploration of news information networks," in *SIGMOD*, 2014, pp. 1091–1094.
- [167] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *KDD*, 2008, pp. 990–998.
- [168] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li *et al.*, "Patentminer: topic-driven patent analysis and mining," in *KDD*, 2012, pp. 1366–1374.
- [169] X. Yin, J. Han, and P. Yu, "Object distinction: Distinguishing objects with identical names," in *ICDE*, 2007, pp. 1242–1246.
- [170] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 5, 2007.
- [171] W. Shen, J. Han, and J. Wang, "A probabilistic model for linking named entities in web text with heterogeneous information networks," in *SIGMOD*, 2014, pp. 1199–1210.
- [172] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, and J. Han, "Clustype: Effective entity recognition and typing by relation phrase-based clustering," in *KDD*, 2015, pp. 995–1004.
- [173] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo, "Mining advisor-advisee relationships from research publication networks," in *KDD*, 2010, pp. 203–212.
- [174] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han, "A bayesian approach to discovering truth from conflicting sources for data integration," *PVLDB*, vol. 5, no. 6, pp. 550–561, 2012.

- [175] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.
- [176] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Identify online store review spammers via social review graph," *TIST*, vol. 3, no. 4, p. 61, 2012.
- [177] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable topical phrase mining from text corpora," *PVLDB*, vol. 8, no. 3, pp. 305–316, 2014.
- [178] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han, "Mining quality phrases from massive text corpora," in *SIGMOD*, 2015, pp. 1729–1744.
- [179] C. Wang, J. Han, Q. Li, X. Li, W.-P. Lin, and H. Ji, "Learning hierarchical relationships among partially ordered objects with heterogeneous attributes and links," in *SDM*, 2012, pp. 516–527.
- [180] Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao, "Community evolution detection in dynamic heterogeneous information networks," in *MLG*, 2010, pp. 137–146.
- [181] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, "Discovering meta-paths in large heterogeneous information networks," in *WWW*, 2015, pp. 754–764.
- [182] J. Cohen, "Graph twiddling in a mapreduce world," *Computing in Science & Engineering*, vol. 11, no. 4, pp. 29–41, 2009.
- [183] U. Kang, C. E. Tsourakakis, and C. Faloutsos, "Pegasus: A peta-scale graph mining system implementation and observations," in *ICDM*, 2009, pp. 229–238.
- [184] G. Colliat, "Olap, relational, and multidimensional database systems," *ACM Sigmod Record*, vol. 25, no. 3, pp. 64–69, 1996.
- [185] P. Zhao, X. Li, D. Xin, and J. Han, "Graph cube: on warehousing and olap multidimensional networks," in *SIGMOD*, 2011, pp. 853–864.
- [186] C. Li, P. S. Yu, L. Zhao, Y. Xie, and W. Lin, "Infonetolaper: Integrating infonetwarehouse and infonetcube with infonetolap," *PVLDB*, vol. 4, no. 12, pp. 1422–1425, 2011.
- [187] M. Yin, B. Wu, and Z. Zeng, "Hmgraph olap: a novel framework for multi-dimensional heterogeneous network analysis," in *DOLAP*, 2012, pp. 137–144.
- [188] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *WWW*, 2004, pp. 491–501.
- [189] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *CIKM*, 2010, pp. 199–208.