Mining Heterogeneous Information Networks: A Structural Analysis Approach*

Yizhou Sun
College of Computer and Information Science
Northeastern University
Boston, MA
yi.sun@neu.edu

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL
hanj@illinois.edu

ABSTRACT

Most objects and data in the real world are of multiple types, interconnected, forming complex, heterogeneous but often semi-structured information networks. However, most network science researchers are focused on homogeneous networks, without distinguishing different types of objects and links in the networks. We view interconnected, multi-typed data, including the typical relational database data, as heterogeneous information networks, study how to leverage the rich semantic meaning of structural types of objects and links in the networks, and develop a structural analysis approach on mining semi-structured, multi-typed heterogeneous information networks. In this article, we summarize a set of methodologies that can effectively and efficiently mine useful knowledge from such information networks, and point out some promising research directions.

1. INTRODUCTION

We are living in an interconnected world. Most of data or informational objects, individual agents, groups, or components are interconnected or interact with each other, forming numerous, large, interconnected, and sophisticated networks. Without loss of generality, such interconnected networks are called *information networks*. Examples of information networks include social networks, the World Wide Web, research publication networks [8], biological networks [12], highway networks [24], public health systems, electrical power grids, and so on. Clearly, information networks are ubiquitous and form a critical component of modern information infrastructure. The analysis of information networks, or their special kinds, such as social networks and the Web, has gained extremely wide attentions nowadays from researchers in computer science, social science, physics, economics, biology, and so on, with exciting discoveries and successful applications across all the disciplines.

We propose to model real-world systems from different applications as *semi-structured heterogeneous information networks*, by structuring objects and their interactions into different types, and investigate the principles and methodologies for systematically mining such networks. Different from many existing network models that view interconnected data as homogeneous graphs or networks, our semi-structured heterogeneous information network model leverages the rich semantics of typed nodes and links in a network and uncovers surprisingly rich knowledge from the network.

For example, in a bibliographic database like DBLP[1] and PubMed[2], papers are linked together via authors, venues and terms, and in Flickr[3], a social website, photos are linked together via users, groups, tags and comments. Different kinds of knowledge can be derived from such an information network view, such as discovery of clusters and hierarchies, ranking, topic analysis, classification, similarity search, and relationship prediction. These functions facilitate the generation of new knowledge in ubiquitous online databases and other online or offline systems in almost every industry. For example, different research areas and ranks for authors and conferences can be discovered by such analysis in a bibliographic database, which will be useful for the users to better understand the data and obtain valuable knowledge.

This article presents an overview of the techniques developed for information network analysis in recent years. The motivation and related concepts are briefly introduced in Section 2. The major mining tasks and techniques are presented in Section 3, and more advanced topics are in Section 4. In Section 5, we propose several research directions along the line of mining heterogeneous information networks. Finally, Section 6 concludes our study.

2. WHY HETEROGENEOUS INFORMATION NETWORKS?

In most of the current research on network science, social and information networks are usually assumed to be *homogeneous*, where nodes are objects of the same entity type (*e.g.*, person) and links are relationships from the same relation type (*e.g.*, friendship). Interesting results have been generated from such studies with numerous influential applications, such as the well-known PageRank algorithm [2] and community detection methods. However, most real world networks are *heterogeneous*, where nodes and relations are of different types. For example, in a healthcare network, nodes can be patients, doctors, medical tests, diseases, medicines,

*The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), NSF IIS-0905215, MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC, and U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

[1]http://www.informatik.uni-trier.de/∼ley/db/
[2]http://www.ncbi.nlm.nih.gov/pubmed/
[3]http://www.flickr.com/

hospitals, treatments, and so on. On one hand, treating all the nodes as of the same type (*e.g.*, homogeneous information networks) may miss important semantic information. On the other hand, treating every node as of a distinct type (*e.g.*, labeled graph) may also lose valuable schema-level information. It is important to know that patients are of the same kind, comparing with some other kinds, such as doctors or diseases. Thus, *a typed, semi-structured heterogeneous network modeling may capture essential semantics of the real world.*

Typed, semi-structured heterogeneous information networks are ubiquitous. For example, the network of Facebook consists of persons as well as objects of other types, such as photos, posts, companies, and movies; in addition to friendship between persons, there are relationships of other types, such as person-photo tagging relationships, person-movie liking relationships, person-post publishing relationships, and post-post replying relationships. A university network may consist of several types of objects like students, professors, courses, and departments, as well as their interactions, such as teaching, course registration or departmental association relationships between objects. Similar kinds of examples are everywhere, from social media to scientific, engineering or medical systems, and to online e-commerce systems. Therefore, *heterogeneous information networks are powerful and expressive representations of general real-world interactions between different kinds of network entities in diverse domains.*

2.1 What Are Heterogeneous Information Networks?

An information network represents an abstraction of the real world, focusing on the *objects* and the *interactions* between the objects. It turns out that this level of abstraction has great power in not only representing and storing the essential information about the real-world, but also providing a useful tool to mine knowledge from it, by exploring the power of links. Formally, we define an information network as follows.

DEFINITION 1. ***(Information network)*** *An* information network *is defined as a directed graph $G = (\mathcal{V}, \mathcal{E})$ with an object type mapping function $\tau : \mathcal{V} \to \mathcal{A}$ and a link type mapping function $\phi : \mathcal{E} \to \mathcal{R}$, where each object $v \in \mathcal{V}$ belongs to one particular object type $\tau(v) \in \mathcal{A}$, each link $e \in \mathcal{E}$ belongs to a particular relation $\phi(e) \in \mathcal{R}$, and 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 the network. Note that, if a relation exists from type A to type B, denoted as $A\,R\,B$, the inverse relation R^{-1} holds naturally for $B\,R^{-1}\,A$. R and its inverse R^{-1} are usually not equal, unless the two types are the same and R is symmetric. When the types of objects $|\mathcal{A}| > 1$ or the types of relations $|\mathcal{R}| > 1$, the network is called **heterogeneous information network**; otherwise, it is a **homogeneous information network**.

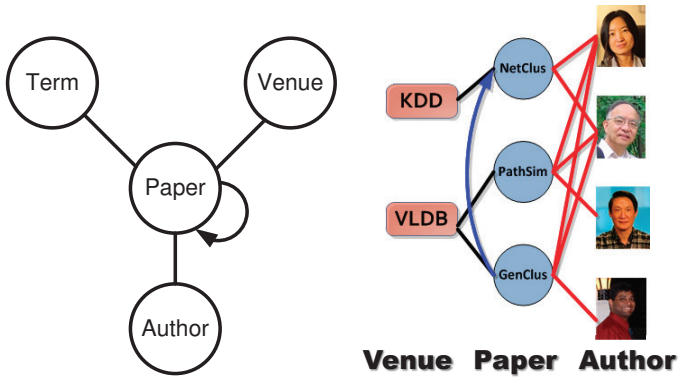
Given a complex heterogeneous information network, it is necessary to provide its meta level (*i.e.*, schema-level) description for better understanding the object types and link types in the network. Therefore, we propose the concept of network schema to describe the meta structure of a network.

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

The network schema of a heterogeneous information network specifies type constraints on the sets of objects and relationships between the objects. These constraints make a heterogeneous information network semi-structured, guiding the exploration of the semantics of the network. An information network following a network schema is then called a *network instance* of the network schema.

Heterogeneous information networks can be constructed from many interconnected, large-scale datasets, ranging from social, scientific, engineering to business applications. Here are a few examples of such networks.

1. **Bibliographic information network:** A bibliographic information network, such as the computer science bibliographic information network derived from DBLP, is a typical heterogeneous network, containing objects in four types of entities: *paper* (P), *venue* (*i.e.*, conference/journal) (V), *author* (A), and *term* (T). For each paper $p \in P$, it has links to a set of authors, a venue, and a set of terms, belonging to a set of link types. It may also contain citation information for some papers, that is, links to a set of papers cited by the paper and links from a set of papers citing the paper. The network schema for a bibliographic network and an instance of such a network are shown in Fig. 1.



(a) Schema of a bibliographic network (b) A bibliographic network

Figure 1: A bibliographic network schema and a bibliographic network instance following the schema (only papers, venues and authors are shown).

2. **Twitter information network:** Twitter as a social media can also be considered as an information network, containing objects types such as *user, tweet, hashtag* and *term*, and relation (or link) types such as *follow* between users, *post* between users and tweets, *reply* between tweets, *use* between tweets and terms, and *contain* between tweets and hashtags.
3. **Flickr information network:** The photo sharing website Flickr can be viewed as an information network, containing a set of object types: *image, user, tag, group,*

and *comment*, and a set of relation types, such as *upload* between users and images, *contain* between images and tags, *belong to* between images and groups, *post* between users and comments and *comment* between comments and images.

4. **Healthcare information network:** A healthcare system can be modeled as a healthcare information network, containing a set of object types, such as *doctor, patient, disease, treatment*, and *device*, and a set of relation types, such as *used-for* between treatments and diseases, *have* between patients and diseases, and *visit* between patients and doctors.

Heterogeneous information networks can be constructed almost in any domain, such as social networks (*e.g.*, Facebook), e-commerce (*e.g.*, Amazon and eBay), online movie databases (*e.g.*, IMDB), and numerous database applications. Heterogeneous information networks can also be constructed from text data, such as news collections, by entity and relationship extraction using natural language processing and other advanced techniques.

Diverse information can be associated with information networks. Attributes can be attached to the nodes or links in an information network. For example, location attributes, either categorical or numerical, are often associated with some users and tweets in a Twitter information network. Also, temporal information is often associated with nodes and links to reflect the dynamics of an information network. For example, in a bibliographic information network, new papers and authors emerge every year, as well as their associated links. Besides the structure information of information networks, such content information is also helpful or even critical in some tasks on mining information networks.

2.2 Why Is Mining Heterogeneous Networks a New Game?

Numerous methods have been developed for the analysis of homogeneous information networks, especially on social networks [1], such as ranking, community detection, link prediction, and influence analysis. However, most of these methods cannot be directly applied to mining heterogeneous information networks. This is not only because heterogeneous links across entities of different types may carry rather different semantic meanings but also because a heterogeneous information network in general captures much richer information than its homogeneous network counterpart. A homogeneous information network is usually obtained by projection from a heterogeneous information network, but with significant information loss. For example, a co-author network can be obtained by projection on co-author information from a more complete heterogeneous bibliographic network. However, such projection will lose valuable information on what subjects and which papers the authors were collaborating on. Moreover, with rich heterogeneous information preserved in an original heterogeneous information network, many powerful and novel data mining functions need to be developed to explore the rich information hidden in the heterogeneous links across entities.

Why is mining heterogeneous networks a new game? Clearly, information propagation across heterogeneous nodes and links can be very different from that across homogeneous nodes and links. Based on our research into mining heterogeneous information networks, especially our studies on ranking-based clustering [19; 22], ranking-based classification [10; 9], meta-path-based similarity search [18], relationship prediction [15; 16], and relation strength learning [14; 20], we believe there are a set of new principles that may guide systematic analysis of heterogeneous information networks. We summarize these principles as follows.

1. **Information propagation across heterogeneous types of nodes and links.** Similar to most of the network analytic studies, links should be used for information propagation in mining tasks. However, the new game is *how to propagate information across heterogeneous types of nodes and links*, in particular, how to compute ranking scores, similarity scores, and clusters, and how to make good use of class labels, across heterogeneous nodes and links. No matter how we work out new, delicate measures, definitions, and methodologies, a golden principle is that *objects in the networks are interdependent, and knowledge can only be mined using the holistic information in a network*.
2. **Search and mining by exploring network meta structures.** Different from homogeneous information networks where objects and links are being treated either as of the same type or as of un-typed nodes or links, heterogeneous information networks in our model are semi-structured and typed, that is, nodes and links are structured by a set of types, forming a network schema. The network schema provides a meta structure of the information network. It provides guidance of search and mining of the network and help analyze and understand the semantic meaning of the objects and relations in the network. Meta-path-based similarity search and mining has demonstrated the usefulness and the power of exploring network meta structures.
3. **User-guided exploration of information networks.** In a heterogeneous information network, there often exist numerous semantic relationships across multiple types of objects, carrying subtly different semantic meanings. A certain weighted combination of relations or meta-paths may best fit a specific application for a particular user. Therefore, it is often desirable to automatically select the right relation (or meta-path) combinations with appropriate weights for a particular search or mining task based on user's guidance or feedback. User-guided or feedback-based network exploration is a useful strategy.

3. MAJOR TASKS AND TECHNIQUES

In this section, we first introduce some fundamental mining tasks in heterogeneous information networks, which include clustering, ranking, classification, similarity search, relationship prediction, and relation strength-aware learning, and the methodologies we have developed to solve these tasks. We partition these tasks into three parts, mainly following the three principles introduced in the last section.

3.1 Clustering and Classification in Heterogeneous Information Networks

Clustering, classification and ranking are basic mining functions for information networks. We introduce several studies that address these tasks in heterogeneous information networks by distinguishing different types of links.

Ranking-based clustering in heterogeneous information networks. For link-based clustering of heterogeneous

Table 1: Rank scores for venues, authors and terms for the net-cluster of the database research area.

Venue	Rank score
SIGMOD	0.315
VLDB	0.306
ICDE	0.194
PODS	0.109
EDBT	0.046
CIKM	0.019
...	...

Author	Rank score
Michael Stonebraker	0.0063
Surajit Chaudhuri	0.0057
C. Mohan	0.0053
Michael J. Carey	0.0052
David J. DeWitt	0.0051
H. V. Jagadish	0.0043
...	...

Term	Rank score
database	0.0529
system	0.0322
query	0.0313
data	0.0251
object	0.0138
management	0.0113
...	...

information networks, we need to explore links across heterogeneous types of data. Recent studies develop a ranking-based clustering approach (*e.g.*, RankClus [19] and NetClus [22]) that generates both clustering and ranking results efficiently. This approach is based on the observation that ranking and clustering can mutually enhance each other because objects highly ranked in each cluster may contribute more towards unambiguous clustering, and objects more dedicated to a cluster will be more likely to be ranked high in the same cluster. It turns out that the accuracy of clustering results can be significantly enhanced compared with that either using projected homogeneous information networks or using only partial link information. Moreover, by integrating ranking and clustering, a cluster can be understood easily by reading the top-ranked objects in that cluster. Table 1 is a net-cluster (*i.e.*, a network cluster following the schema of the original network) generated by NetClus [22] on the DBLP network, representing the database research area.

Classification of heterogeneous information networks. Classification can also take advantage of links in heterogeneous information networks. Knowledge can be effectively propagated across a heterogeneous network because the nodes that are linked together are likely to be similar, and different types of links have different level of strengths in determining this similarity. Moreover, following the idea of ranking-based clustering, one can explore ranking-based classification since objects highly ranked in a class are likely to play a more important role in classification. These ideas lead to effective algorithms, such as GNetMine [10] and RankClass [9]. It turns out that by distinguishing different types of links in a heterogeneous information network, classification accuracy can be significantly enhanced.

3.2 Meta-Path-Based Similarity Search and Mining

We then introduce a systematic approach for dealing with general heterogeneous information networks with a specified network schema, by using meta-path-based methodologies. Under this framework, similarity search and interesting mining tasks, such as relationship prediction, can be addressed. Different from homogeneous information networks, two objects can be connected via different types of paths in a heterogeneous information network. For example, two authors can be connected via "author-paper-author" path, "author-paper-venue-paper-author" path, and so on. Formally, these paths are called *meta-paths*, defined as follows.

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

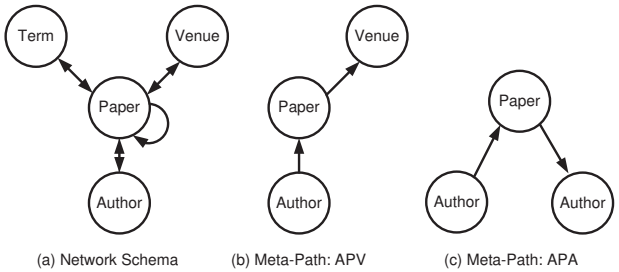


Figure 2: Bibliographic network schema and meta-paths.

For the bibliographic network schema shown in Figure 2 (a), we list two examples of meta-paths in Figure 2 (b) and (c), where an arrow explicitly shows the direction of a relation. We say a path $p = (a_1 a_2 \ldots a_{l+1})$ between a_1 and a_{l+1} in network G follows the meta-path $\mathcal{P}$, if $\forall i$, $a_i \in A_i$ and each link $e_i = \langle a_i a_{i+1} \rangle$ belongs to each relation R_i in $\mathcal{P}$. We call these paths as *path instances* of $\mathcal{P}$, denoted as $p \in \mathcal{P}$. Some path instance examples are shown in Table 2.

Table 2: Path instances and their corresponding meta-paths in heterogeneous information networks.

	Connection Type I	Connection Type II
Path instance	Jim-P_1-Ann Mike-P_2-Ann Mike-P_3-Bob	Jim-P_1-SIGMOD-P_2-Ann Mike-P_3-SIGMOD-P_2-Ann Mike-P_4-KDD-P_5-Bob
Meta-path	A(uthor)-P(aper)-A	A-P-V(enue)-P-A

Via meta-paths, one can systematically specify how object types are connected in a network. Different meta-paths lead to different kinds of features. Multiple mining tasks can be explored under this framework.

Meta-path-based similarity search in heterogeneous information networks. Similarity search plays an important role in the analysis of networks. By considering different linkage paths (*i.e.*, meta-path) in a network, one can derive various semantics on similarity in a heterogeneous information network. For example, Table 3 shows that using different meta-paths, one can find different author lists that are most similar to Christos Faloutsos. A meta-path based similarity measure, PathSim, is introduced in [18], for finding peer objects in the network, which generates better results, compared with random-walk based similarity mea-

sures. Another measure, HeteSim, introduced in [13], computes relevance score between objects of different types.

Table 3: Top-10 similar authors to "Christos Faloutsos" under different meta-paths on the *full-DBLP* dataset.

(a) Path: APA

Rank	Author
1	Christos Faloutsos
2	Spiros Papadimitriou
3	Jimeng Sun
4	Jia-Yu Pan
5	Agma J. M. Traina
6	Jure Leskovec
7	Caetano Traina Jr.
8	Hanghang Tong
9	Deepayan Chakrabarti
10	Flip Korn

(b) Path: $APVPA$

Rank	Author
1	Christos Faloutsos
2	Jiawei Han
3	Rakesh Agrawal
4	Jian Pei
5	Charu C. Aggarwal
6	H. V. Jagadish
7	Raghu Ramakrishnan
8	Nick Koudas
9	Surajit Chaudhuri
10	Divesh Srivastava

Meta-path-based relationship prediction in heterogeneous information networks. Heterogeneous information network brings interactions among multiple types of objects and hence the possibility of predicting relationships across heterogeneous typed objects. By systematically designing meta-path-based topological features and measures in the network, supervised models can be used to learn the best weights associated with different topological features for effective relationship prediction [15; 16].

As a case study, the co-authorship prediction problem is examined in [15], which outputs the most significant meta-paths for predicting co-authorships, as shown in Table 4, and also provides better understanding why co-author relationships are built. Note that, predicting co-authors for a given author is an extremely difficult task, as there are too many candidate target authors (3-hop candidates are used in analysis), but the number of real new relationships are usually very small. Table 5 shows the top-5 predicted co-authors in time interval T_2 (2003-2009) using the $T_0 - T_1$ (topological features are collected in 1989-1995 and co-authorship building ground truths are collected in 1996-2002) training framework, for both the proposed hybrid topological features and the shared co-author feature. We can see that the results generated by heterogeneous features has a higher accuracy compared with the homogeneous one.

3.3 User-Guided Relation Strength-Aware Mining

The heterogeneity of relations between object types leads to different mining results that can be chosen by users. With user guidance, the strength of each relation should be automatically learned and used for better mining. We introduce

Table 4: Significance of meta-paths with *Normalized Path Count measure* for $HP3hop$ dataset.

Meta-path	p-value	Significance level[1]
$A-P \rightarrow P-A$	0.0378	**
$A-P \leftarrow P-A$	0.0077	***
$A-P-V-P-A$	1.2974e-174	****
$A-P-A-P-A$	1.1484e-126	****
$A-P-T-P-A$	3.4867e-51	****
$A-P \rightarrow P \rightarrow P-A$	0.7459	
$A-P \leftarrow P \leftarrow P-A$	0.0647	*
$A-P \rightarrow P \leftarrow P-A$	9.7641e-11	****
$A-P \leftarrow P \rightarrow P-A$	0.0966	*

[1] *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$

Table 5: Top-5 predicted co-authors for Jian Pei in 2003-2009.

Rank	Hybrid heterogeneous features	# of shared authors as features
1	**Philip S. Yu**	**Philip S. Yu**
2	**Raymond T. Ng**	Ming-Syan Chen
3	Osmar R. Zaïane	Divesh Srivastava
4	**Ling Feng**	Kotagiri Ramamohanarao
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

[*] Bold font indicates true new co-authors of Jian Pei in the period of 2003-2009.

two different kinds of relation strength-aware mining tasks.

Table 6: Case studies of cluster membership results.

Object	DB	DM	IR	ML
SIGMOD	0.8577	0.0492	0.0482	0.0449
KDD	0.0786	0.6976	0.1212	0.1026
CIKM	0.2831	0.1370	0.4827	0.0971
Jennifer Widom	0.7396	0.0830	0.1061	0.0713
Jim Gray	0.8359	0.0656	0.0536	0.0449
Christos Faloutsos	0.4268	0.3055	0.1380	0.1296

(a) A-V network (b) A-V-P network

Figure 3: Strengths for link types in two *DBLP four-area networks*.

Relation strength-aware clustering via attribute selection. Links in networks are frequently used to regularize the attribute-based clustering tasks, that is, linked objects should have similar cluster labels. However, shall we trust links from different types equally? We propose GenClus [14] to address this problem. By specifying a set of attributes, the strengths of different relations in heterogeneous information networks can be automatically learned to help the clustering of objects. Table 6 shows a clustering case study for objects from different types in a DBLP network, where the network schema is in Fig. 3(a). Fig. 3 demonstrates the

learned strengths for each relation for the clustering task for two different network schemas.

Integrating user-guided clustering with meta-path selection. Different meta-paths in a heterogeneous information network represent different relations with different semantic meanings. User guidance in the form of a small set of training examples for some object types can indicate their preference on the results of clustering. The preferred meta-path or weighted meta-path combinations can be learned to reach better consistency between mining results and the training examples [20].

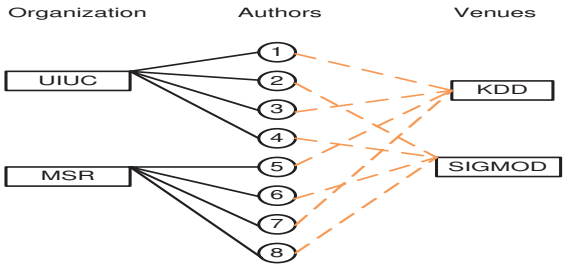


Figure 4: A toy heterogeneous information network containing organizations, authors and venues.

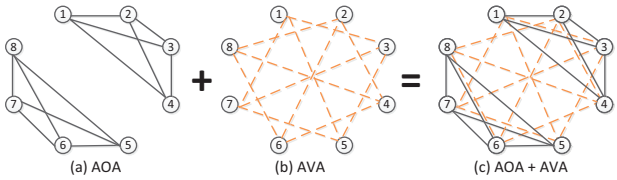


Figure 5: Author connection graphs under different meta-paths.

EXAMPLE 1. ***(Meta-path-based clustering)*** *A toy heterogeneous information network is shown in Figure 4, which contains three types of objects: organization (O), author (A) and venue (V), and two types of links: the solid line represents the affiliation relation between author and organization, whereas the dashed one the publication relation between author and venue. Authors are then connected (indirectly) via different meta-paths. For example,* $A-O-A$ *is a meta-path denoting a relation between authors via organizations (i.e., colleagues), whereas* $A-V-A$ *denotes a relation between authors via venues (i.e., publishing in the same venues). A question then arises:* which type of connections should we use to cluster the authors?
Obviously, there is no unique answer to this question: different meta-paths lead to different author connection graphs, which may lead to different clustering results.
In Figure 5(a), authors are connected via organizations and form two clusters: $\{1,2,3,4\}$ *and* $\{5,6,7,8\}$*; in Figure 5(b), authors are connected via venues and form two different clusters:* $\{1,3,5,7\}$ *and* $\{2,4,6,8\}$*; whereas in Figure 5(c), a connection graph combining both meta-paths generates 4 clusters:* $\{1,3\},\{2,4\},\{5,7\}$ *and* $\{6,8\}$.

In [20], the PathSelClus algorithm is proposed to learn the importance of each meta-path as well as output the clustering results that are consistent with the user guidance. For example, to cluster authors into clusters in Example 1, a user may seed $\{1\}$ and $\{5\}$ for two clusters, which implies a selection of meta-path $A-O-A$; or seed $\{1\},\{2\},\{5\}$, and $\{6\}$ for four clusters, which implies a combination of both meta-paths $A-O-A$ and $A-V-A$ with about equal weight.

4. ADVANCED TOPICS

Beyond the basic mining tasks discussed above, in this section, we introduce several advanced topics for mining information networks, which include role discovery, trustworthiness analysis, co-evolution analysis, text mining in information networks, and OLAP in information networks. Many of these tasks can help better refine the quality of information networks, and others will help better understand the content rich information networks. More advanced operators such as OLAP is also necessary for better exploring the networks.

4.1 Role Discovery in Information Networks

An information network contains abundant knowledge about relationships among objects. Unfortunately, such knowledge, such as advisor-advisee relationships among researchers in a bibliographic network, is often hidden. Role discovery is to uncover such hidden relationships by information network analysis. For example, a time-constrained probabilistic factor graph model, which takes a research publication network as input and models the advisor-advisee relationship mining problem using a jointly likelihood objective function has been developed [25]. It successfully mines advisor-advisee hidden roles in the DBLP database with high accuracy. Such mechanism can be further developed to discover hierarchical relationships [26] and ontology among objects under different kinds of user-provided constraints or rules.

4.2 Trustworthiness Analysis in Information Networks

A major challenge for data integration is to derive the most complete and accurate integrated records from diverse and sometimes conflicting sources. The *truth finding* problem is to decide which piece of information being merged is most likely to be true. By constructing an information network that links multiple information providers with multiple versions of the stated facts for each entity to be resolved, novel network analysis methods, such as TruthFinder [28] and LTM [30], can be developed to resolve the conflicting source problem effectively. In [7], the authors propose to detect copying relationships among sources, which turns out to be critical in resolving conflicts among sources. Trustworthy analysis will help data cleaning and data integration, hence improve the quality of information networks.

4.3 Evolution Analysis in Heterogeneous Information Networks

Many current studies on network evolution are on homogeneous networks. However, in the real cases, different relationships exist in the heterogeneous network, and multi-typed relationships will co-evolve together. Modeling co-evolution of multi-typed objects will capture richer semantics than modeling on single-typed objects alone. For example, studying co-evolution of authors, venues and terms in a bibliographic network can tell better the evolution of

research areas than just examining co-author network or term network alone. Thus an important direction is how to model the co-evolution of multi-typed objects in the form of multi-typed cluster evolution in heterogeneous networks, such as EvoNetClus which builds a hierarchical Dirichlet process mixture model-based online model to study the real heterogeneous networks formed by DBLP and twitter [21].

4.4 Integrating Text Mining and Information Networks

Objects in information networks are usually associated with text information, and it is interesting to integrate the traditional text mining problem with information networks. In [17], we propose a topic model, iTopicModel, which can generate topics not only based on the text information of documents but also based on the link information among documents. It turns out that the topic quality can be significantly enhanced especially when the text is sparse and link quality is high. In [6], the authors further distinguish the types of links between documents, and a biased propagation among different documents is considered. A novel topic model with biased propagation (TMBP) algorithm is proposed, which directly incorporates heterogeneous information network with topic modeling in a unified way. The underlying intuition is that multi-typed objects should be treated differently along with their inherent textual information and the rich semantics of the heterogeneous information network. Besides topic models, [5] proposes a joint regularization framework to enhance expertise retrieval by modeling heterogeneous networks as regularization constraints on top of document-centric model, which can find high quality experts for a given query.

4.5 Online Analytical Processing of Heterogeneous Information Networks

The power of online analytical processing (OLAP) has been shown in multidimensional analysis of structured, relational data. Similarly, users may like to view a heterogeneous information network from different angles, in different dimension combinations, and at different levels of granularity. For example, in a bibliographic network, by specifying the object type as paper and link type as citation relation, and rolling up papers into research topics, we can immediately see the citation relationships between different research topics and figure out which research topic could be the driving force for others. However, the extension of the concept of online analysis processing (OLAP) to multi-dimensional data analysis of heterogeneous information networks is nontrivial. Not only different applications may need different ontological structures and concept hierarchies to summarize information networks but also because multiple pieces of semantic information in heterogeneous networks are intertwined, determined by multiple nodes and links. There are some preliminary studies on this issue, such as [23; 3; 31], but the large territories of online analytical processing of information networks are still waiting to be explored.

5. RESEARCH FRONTIERS

Viewing interconnected data as an information network and studying systematically the methods for mining heterogeneous information networks is a promising frontier in data mining research. There are still many challenging research issues. Here we illustrate only a few.

5.1 Constructing and Refining Heterogeneous Information Networks

Many studies on mining heterogeneous information networks assume that a heterogeneous information network to be investigated contains a well-defined network schema and a large set of relatively clean and unambiguous objects and links. However, in the real world, things are more complicated.

A network extracted from a relational database may contain a well-defined schema which can be used to define the schema of its corresponding heterogeneous information network. Nevertheless, objects and links even in such a database-formed information network can still be noisy. For example, in the DBLP network, different authors may share the same name [27], that is, one node in a network may refer to multiple real-world entities; whereas in some other cases, different nodes in a network may refer to the same entity. Entity resolution will need to be integrated with network mining in order to merge and split objects or links and derive high quality results. Moreover, links in a network, roles of a node with respect to some other nodes may not be explicitly given. For example, the advisor-advisee relationship in the DBLP network [25] is not given, but such kind of relationships can be critical for understanding the growth of a research community or for some other data mining tasks. Furthermore, sometimes the connections between different nodes may not be reliable or trustable. For example, the author information for a book provided by an online book store could be erroneous or inaccurate. Multiple Web-sites may provide conflicting or compensating information for the properties of certain objects. Trustworthiness modeling [30] could be critically important for data cleaning, data integration, and quality network construction.

Construction of high-quality heterogeneous information networks becomes increasingly more challenging when we move away from relational databases towards increasingly more complicated, unstructured data, from text documents, to online web-based systems, multimedia data, and multilingual data. Information extraction, natural language understanding, and many other information processing techniques should be integrated with network construction and analysis techniques to ensure high-quality information networks can be constructed and progressively refined so that quality mining can be performed on better-quality heterogeneous information networks.

Notice that entity extraction, data cleaning, detection of hidden semantic relationships, and trustworthiness analysis should be integrated with the network construction and mining processes to progressively and mutually enhance the quality of construction and mining of information networks.

5.2 Diffusion Analysis in Heterogeneous Information Networks

Diffusion analysis has been studied on homogeneous networks extensively, from the innovation diffusion analysis in social science [11] to obesity diffusion in health science [4]. However, in the real world, pieces of information or diseases are propagated in more complex ways, where different types of links may play different roles. For example, diseases could propagate among people, different kinds of animals and food, via different channels. Comments on a product may propagate among people, companies, and news agen-

cies, via traditional news feeds, social media, reviews, and so on. It is highly desirable to study the issues on information diffusion in heterogeneous information networks in order to capture the spreading models that better represent the real world patterns.

5.3 Discovery and Mining of Hidden Information Networks

Although a network can be huge, a user at a time could be only interested in a tiny portion of nodes, links, or subnetworks. Instead of directly mining the entire network, it is more fruitful to mine hidden networks "extracted" dynamically from some existing networks, based on user-specified constraints or expected node/link behaviors. For example, instead of mining an existing social network, it could be more fruitful to mine networks containing suspects and their associated links; or mine subgraphs with nontrivial nodes and high connectivity. How to discover such hidden networks and how to mine knowledge (*e.g.*, clusters, behaviors, and anomalies) from such hidden but non-isolated networks (*i.e.*, still intertwined with the gigantic network in both network linkages and semantics) could be an interesting but challenging problem.

5.4 Discovery of Application-Oriented Ontological Structures in Heterogeneous Information Networks

As shown in the studies on ranking-based clustering and ranking-based classification, interconnected, multiple typed objects in a heterogeneous information network often provide critical information for generating high quality, fine-level concept hierarchies. For example, it is often difficult to identify researchers just based on their research collaboration networks. However, putting them in a heterogeneous network that links researchers with their publication, conferences, terms and research papers, their roles in the network becomes evidently clear. Moreover, people may have different preferences over ontological structures at handling different kinds of tasks. For example, some people may be interested in the research area hierarchy in the DBLP network, whereas others may be interested in finding the author lineage hierarchy. How to incorporate user's guidance, and generate adaptable ontological structures to meet users's requirement and expectation could be an interesting and useful topic to study.

5.5 Intelligent Querying and Semantic Search in Heterogeneous Information Networks

Given real-world data are interconnected, forming gigantic and complex heterogeneous information networks, it poses new challenges to query and search in such networks intelligently and efficiently. Given the enormous size and complexity of a large network, a user is often only interested in a small portion of the objects and links most relevant to the query. However, objects are connected and inter-dependent on each other, how to search effectively in a large network for a given user's query could be a challenge. Similarity search that returns the most similar objects to a queried object, as studied in this thesis [18] and its follow-up [13], will serve as a basic function for semantic search in heterogeneous networks. Such kind of similarity search may lead to useful applications, such as product search in e-commerce networks and patent search in patent networks. Search functions should be further enhanced and integrated with many other functions. For example, structural search [29], which tries to find semantically similar structures given a structural query, may be useful for finding pattern in an e-commerce network involving buyers, sellers, products, and their interactions. Also, a recommendation system may take advantage of heterogeneous information networks that link among products, customers and their properties to make improved recommendations. Querying and semantic search in heterogeneous information networks opens another interesting frontier on research related to mining heterogeneous information networks.

6. CONCLUSIONS

Most objects and data in the real world are interconnected, forming complex, heterogeneous but often semi-structured information networks. However, many database researchers consider a database merely as a data repository that supports storage and retrieval rather than an information-rich, inter-related and multi-typed information network that supports comprehensive data analysis; whereas many network researchers focus on homogeneous networks. Departing from both, we view interconnected, semi-structured datasets as heterogeneous, information-rich networks and study how to uncover hidden knowledge in such networks. In this article, we present an organized picture on mining heterogeneous information networks and introduce a set of interesting, effective and scalable network mining methods. In addition, we also present several promising research topics in this exciting direction.

7. REFERENCES

[1] C. C. Aggarwal, editor. *Social Network Data Analytics*. Springer, 2011.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th Int. World Wide Web Conf. (WWW'98)*, pages 107–117, Brisbane, Australia, April 1998.

[3] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu. Graph OLAP: Towards online analytical processing on graphs. In *Proc. 2008 Int. Conf. Data Mining (ICDM'08)*, Pisa, Italy, Dec. 2008.

[4] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357(4):370–379, 2007.

[5] H. Deng, J. Han, M. R. Lyu, and I. King. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries (JCDL'12)*, pages 71–80, 2012.

[6] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proc. 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, CA, Aug. 2011.

[7] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *Proc. VLDB Endow.*, 3(1-2):1358–1369, Sept. 2010.

[8] C. L. Giles. The future of citeseer: *citeseer^x*. In *Proc. 10th European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD'06)*, Berlin, Germany, September 2006.

[9] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proc. 2011 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, San Diego, CA, Aug. 2011.

[10] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. In *Proc. 2010 European Conf. Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD'10)*, Barcelona, Spain, Sept. 2010.

[11] E. M. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 2003.

[12] T. L. S. Roy and M. Werner-Washburne. Integrative construction and analysis of condition-specific biological networks. In *Proc. 2007 AAAI Conf. on Artificial Intelligence (AAAI'07)*, Vancouver, BC, July 2007.

[13] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *Proc. 2012 Int. Conf. on Extending Database Technology (EDBT'12)*, pages 180–191, Berlin, Germany, March 2012.

[14] Y. Sun, C. C. Aggarwal, and J. Han. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *PVLDB*, 5:394–405, 2012.

[15] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proc. 2011 Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'11)*, Kaohsiung, Taiwan, July 2011.

[16] Y. Sun, J. Han, C. C. Aggarwal, and N. Chawla. When will it happen? relationship prediction in heterogeneous information networks. In *Proc. 2012 ACM Int. Conf. on Web Search and Data Mining (WSDM'12)*, Seattle, WA, Feb. 2012.

[17] Y. Sun, J. Han, J. Gao, and Y. Yu. iTopicModel: Information network-integrated topic modeling. In *Proc. 2009 Int. Conf. Data Mining (ICDM'09)*, Miami, FL, Dec. 2009.

[18] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proc. 2011 Int. Conf. Very Large Data Bases (VLDB'11)*, Seattle, WA, Aug. 2011.

[19] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In *Proc. 2009 Int. Conf. Extending Data Base Technology (EDBT'09)*, Saint-Petersburg, Russia, Mar. 2009.

[20] 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 *Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12)*, Beijing, China, Aug. 2012.

[21] Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao. Community evolution detection in dynamic heterogeneous information networks. In *Proc. 2010 KDD Workshop on Mining and Learning with Graphs (MLG'10)*, Washington D.C., July 2010.

[22] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proc. 2009 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, June 2009.

[23] Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *Proc. 2008 ACM SIGMOD Int. Conf. Management of Data (SIGMOD'08)*, pages 567–580, Vancouver, BC, Canada, June 2008.

[24] Z. B. C. C. W. Jiang, J. Vaidya and B. Banich. Knowledge discovery from transportation network data. In *Proc. 2005 Int. Conf. Data Mining (ICDE'05)*, Tokyo, Japan, April 2005.

[25] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *Proc. 2010 ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10)*, Washington D.C., July 2010.

[26] 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 *Proc. 2012 SIAM Int. Conf. on Data Mining (SDM'12)*, Anaheim, CA, April 2012.

[27] X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[28] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the Web. *IEEE Trans. Knowledge and Data Engineering*, 20:796–808, 2008.

[29] X. Yu, Y. Sun, P. Zhao, and J. Han. Query-driven discovery of semantically similar substructures in heterogeneous networks. In *Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12)*, Beijing, China, Aug. 2012.

[30] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A Bayesian approach to discovering truth from conflicting sources for data integration. In *Proc. 2012 Int. Conf. Very Large Data Bases (VLDB'12)*, Istanbul, Turkey, Aug. 2012.

[31] P. Zhao, X. Li, D. Xin, and J. Han. Graph cube: On warehousing and OLAP multidimensional networks. In *Proc. 2011 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'11)*, Athens, Greece, June 2011.