HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks

Chuan Shi, *Member, IEEE*, Xiangnan Kong, Yue Huang, Philip S. Yu, *Fellow, IEEE*, and Bin Wu, *Member, IEEE*

Abstract—Similarity search is an important function in many applications, which usually focuses on measuring the similarity between objects with the same type. However, in many scenarios, we need to measure the relatedness between objects with different types. With the surge of study on heterogeneous networks, the relevance measure on objects with different types becomes increasingly important. In this paper, we study the relevance search problem in heterogeneous networks, where the task is to measure the relatedness of heterogeneous objects (including objects with the same type or different types). A novel measure HeteSim is proposed, which has the following attributes: (1) a uniform measure: it can measure the relatedness of objects with the same or different types in a uniform framework; (2) a path-constrained measure: the relatedness of object pairs are defined based on the search path that connects two objects through following a sequence of node types; (3) a semi-metric measure: HeteSim has some good properties (e.g., self-maximum and symmetric), which are crucial to many data mining tasks. Moreover, we analyze the computation characteristics of HeteSim and propose the corresponding quick computation strategies. Empirical studies show that HeteSim can effectively and efficiently evaluate the relatedness of heterogeneous objects.

Index Terms—Heterogeneous information network, similarity search, pair-wise random walk, relevance measure

1 Introduction

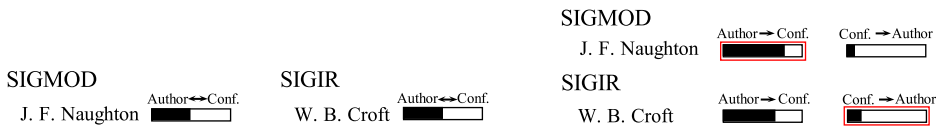
Similarity search is an important task in a wide range of applications, such as web search [1] and product recommendations [2]. The key of similarity search is similarity measure, which evaluates the similarity of object pairs. Similarity measure has been extensively studied for traditional categorical and numerical data types, such as Jaccard coefficient and cosine similarity. There are also a few studies on leveraging link information in networks to measure the node similarity, such as Personalized PageRank [3], SimRank [4], and PathSim [5]. Conventional study on similarity measure focuses on objects with the same type. That is, the objects being measured are of the same type, such as "document-to-document" and "webpage-to-webpage". There are very few studies on similarity measure on objects with different types. That is, the objects being measured are of different types, such as "author-to-conference" and "user-to-movie". It is reasonable. The similarity of objects with different types is a little against our common sense. Moreover, different from the similarity of objects with the same type, which can be measured on homogeneous situation (e.g., the same feature space or homogeneous link structure), it is even hard to define the similarity of objects with different types.

- *C. Shi, Y. Huang, and B. Wu are with Beijing University of Posts and Telecommunications, Beijing, China. E-mail: {shichuan, wubin}@bupt.edu.cn, ymoon.huang@gmail.com.*
- *X.N. Kong and P.S. Yu are with the Department of Computer Science, University of Illinois at Chicago, IL. E-mail: kongxn@gmail.com, psyu@uic.edu.*

Manuscript received 24 Jan. 2013; revised 10 Dec. 2013; accepted 23 Dec. 2013; published online xxxxx.
Recommended for acceptance by G. Karypis.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2013-01-0062.
Digital Object Identifier no. 10.1109/TKDE.2013.2297920

However, the similarity of objects with different types is not only meaningful but also useful in some scenarios. For example, the author J.F. Naughton is more relevant to SIGMOD than KDD. Moreover, the similarity measure of objects with different types are needed in many applications. For example, in a recommendation system, we need to know the relatedness between users and items to make accurate recommendations [6]. It is very important to determine the relatedness between entities (e.g., drug-disease) from medicine annotations data [7]. In an automatic profile extraction application, we need to measure the relatedness of objects with different types, such as authors and conferences, conferences and organizations, etc. Particularly, with the advent of study on heterogeneous information networks [5], [8], it is not only increasingly important but also feasible to study the relatedness among objects with different types. Heterogeneous information networks are the logical networks involving multiple-typed objects and multiple-typed links denoting different relations [9]. It is clear that heterogeneous information networks are ubiquitous and form a critical component of modern information infrastructure [9]. So it is essential to provide a relevance search function on objects with different types in such networks, which is the base of many applications. Since objects with different types coexist in the same network, their relevance measure is possible through link structure.

In this paper, we study the relevance search problem in heterogeneous information networks. The aim of relevance search is to effectively measure the relatedness of heterogeneous objects (including objects with the same type or different types). Different from the similarity search which only measures the similarity of objects with the same type, the relevance search measures the relatedness of



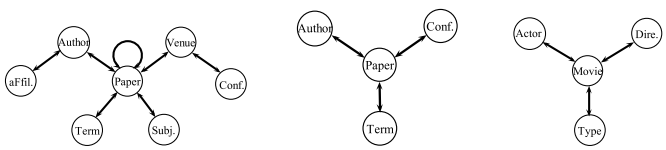
(a) Symmetric measure (b) Asymmetric measure

Fig. 1. Examples of relative importance representing by symmetric and asymmetric measures. The rectangle with partially marked black denotes the relatedness of two objects.

heterogeneous objects, not limit to objects with the same type. Distinct from relational retrieval [10], [11] in information retrieval domain, here relevance search is done on heterogeneous networks which can be constructed from metadata of objects. Moreover, we think that a desirable relevance measure should satisfy the symmetry property based on the following reasons. (1) The symmetric measure is more general and useful in many learning tasks. Although the symmetry property is not necessary in the query task, it is essential for many important tasks, such as clustering and collaborative filtering. Moreover, it is the necessary condition for a metric. (2) The symmetric measure makes more sense in many applications, especially for the relatedness of heterogeneous object pairs. For example, in some applications, we need to answer the question like who has similar importance to the SIGIR conference as J.F. Naughton to SIGMOD. Through comparing the relatedness of object pairs, we can deduce the information of their relative importance. However, it can only be done by the symmetric measure, not the asymmetric measure. It can be explained by the example shown in Fig. 1. For the symmetric measure, we can deduce that W. B. Croft[1] has the same importance to SIGIR as J. F. Naughton[2] to SIGMOD, since their relatedness scores are close. However, we cannot deduce the relative importance information from an asymmetric measure as shown in Fig. 1b. From the relatedness of author to conference and conference to author, we will draw conflicting conclusions.

Despite its value and significance, the relevance search in heterogeneous networks has seldom been studied so far. It faces the following research challenges. (1) Heterogeneous network is much more complex than traditional homogeneous network. In heterogeneous networks, different-typed objects and links coexist in a network and they carry different semantic meanings. As a bibliographic example shown in Fig. 2b (more details in Section 5.1), it includes author, paper, term, and conference type. The relation "author-paper" means author writing paper, while the relation "paper-conference" means paper published in conference. If disregarding the difference of types and semantics, it does not make sense to mix different-typed objects to measure the similarity. We can find that search paths, connecting two objects through a sequence of relations between object types, embody rich semantic information [5]. Based on different search paths, the relatedness of two objects may be totally different. As a consequence, a desirable relevance measure should be path-dependent, since such a measure can capture the semantics under paths and return meaningful values based on different paths. (2) It is difficult to design a uniform and symmetric relevance measure for heterogeneous objects. In heterogeneous networks, the paths connecting objects with the same type are usually symmetric and the path length is an even number, so it may be not difficult to design a symmetric measure based on the symmetric paths, as the PathSim [5] does. However, the paths connecting objects with different types are asymmetric and the path length may be an odd number. In this condition, it is not easy to design a symmetric relevance measure. It is more challengeable to design a uniform relevance measure for these two conditions.

Inspired by the intuition that two objects are related if they are referenced by related objects, we propose a general framework, called HeteSim, to evaluate the relatedness of heterogeneous objects in heterogeneous networks. HeteSim is a path-based relevance measure, which can effectively capture the subtle semantics of search paths. Based on pair-wise random walk (RW) model, HeteSim treats arbitrary search paths in a uniform way, which guarantees the symmetric property of HeteSim. An additional benefit is that HeteSim can evaluate the relatedness of objects with the same or different types in the same way. Moreover, HeteSim is a semi-metric measure. In other words, HeteSim satisfies the properties of non-negativity, identity of indiscernibles, and symmetry. It implies that HeteSim can be used in many learning tasks (e.g., clustering and collaborative filtering). We also consider the computation issue of HeteSim and propose four fast computation strategies. The extensive experiments validate the effectiveness of HeteSim. As a general relevance measure, HeteSim illustrates its benefits and generality in knowledge discovery of heterogeneous networks through three case studies: automatically extracting object profile, experts finding through relative importance of object pairs, and relevance search based on path semantics. HeteSim also shows its potential in the machine learning tasks (i.e., query and clustering) where HeteSim outperforms other well-established similarity measures. In addition, numerous experiments test the significance of fast computing strategies of HeteSim.



(a) ACM data (b) DBLP data (c) Movie data

Fig. 2. Examples of heterogeneous information network schema.

1. http://ciir.cs.umass.edu/personnel/croft.html.
2. http://pages.cs.wisc.edu/~naughton/.

2 Related Work

The most related work to relevance search is similarity search. Here we briefly summarize these works. Similarity search has been well studied for a long time. These studies can be roughly categorized into two types: 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 k nearest neighbor is also widely used in similarity measure [12], which aims at finding top-k nearest neighbors according to similarities defined on numerical features. Based on feature similarity, the top-k similarity pair search algorithm (i.e., top-k-join) considers similarity between tuples [13]. This type of approaches does not consider link relation among objects, so they cannot be applied to networked data.

The link based approaches measure the similarity of objects based on their link structures in a graph. The asymmetrical similarity measure, Personalized PageRank [3], evaluates the probability starting from a source object to a target object by randomly walking with restart. It is extended to the scalable calculation for online queries [14] and the top-k answers [15]. SimRank [4] is a symmetric similarity measure, which evaluates the similarity of two objects by their neighbors' similarities. SCAN [16] measures similarity of two objects by comparing their immediate neighbor sets. Recently, Jin et al. proposed RoleSim to measure the role similarity of node pair by automorphic equivalence [17]. These approaches just consider the objects with the same type, so they cannot be directly applied in heterogeneous networks. ObjectRank [18] applies authority-based ranking to keyword search in labeled graphs and PopRank [19] proposes a domain-independent object-level link analysis model. Although these two approaches noticed that heterogeneous relationships could affect the similarity, they do not consider the distinct semantics of paths that include different-typed objects, so they also cannot measure the similarity of objects in heterogeneous networks.

Recently, the relevance research in heterogeneous data emerges. Wang et al. [20] proposed a model to learn relevance from heterogeneous data, while their model more focuses on analyzing the context of heterogeneous networks, rather than network structure. Based on a Markov-chain model of random walk, Fouss et al. [21] designed a similarity metric ECTD with nice properties and interpretation. Unfortunately, absent of path constraint, ECTD cannot capture the subtle semantics in heterogeneous networks. Considering semantics in meta paths constituted by different-typed objects, Sun et al. [5] proposed PathSim to measure the similarity of same-typed objects based on symmetric paths. However, many valuable paths are asymmetric and the relatedness of different-typed objects are also meaningful. PathSim is not suitable in these conditions. In information retrieval community, Lao and Cohen [22] proposed 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. Although the PCRW model can be applied to measuring the relatedness of different-typed objects, the asymmetric property of PCRW restricts its applications. In the proposed HeteSim, users can measure the relatedness of heterogeneous objects based on an arbitrary search path. The good merits of HeteSim (e.g., symmetric and self-maximum) make it suitable for more applications.

3 Preliminary

A heterogeneous information network is a special type of information network, which either contains multiple types of objects or multiple types of links.

Definition 1 (Information Network). *Given a schema $S = (\mathcal{A}, \mathcal{R})$ which consists of a set of entities types $\mathcal{A} = \{A\}$ and a set of relations $\mathcal{R} = \{R\}$, an information network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\phi : V \to \mathcal{A}$ and a link type mapping function $\psi : E \to \mathcal{R}$. Each object $v \in V$ belongs to one particular object type $\phi(v) \in \mathcal{A}$, and each link $e \in E$ belongs to a particular relation $\psi(e) \in \mathcal{R}$. 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**.*

In information networks, we distinguish object types and relation types. As a template for a network, the network schema depicts the object types and the relations existing among object types. For a relation R existing from type A to type B, denoted as $A \xrightarrow{R} B$, A and B are the *source type* and *target type* of relation R, which are denoted as $R.S$ and $R.T$, respectively. The inverse relation R^{-1} holds naturally for $B \xrightarrow{R^{-1}} A$. Generally, R is not equal to R^{-1}, unless R is symmetric and these two types are the same.

Example 1. A bibliographic information network is a typical heterogeneous information network. The network schema of ACM data set (see Section 5.1) is shown in Fig. 2a. It contains objects from seven types of entities: papers (P), authors (A), affiliations (F), terms (T), subjects (S), venues (V), and conferences (C) (a conference includes multiple venues, e.g., KDD including KDD2010 and KDD2009). There are links connecting different-typed objects. The link types are defined by the relations between two object types. For example, links exist between authors and papers denoting the writing or written-by relations. Figs. 2b and 2c show the network schemas of DBLP data set and IMDB movie data (see Section 5.1), respectively.

Different from homogeneous networks, two objects in a heterogeneous network can be connected via different paths and these paths have different meanings. For example, in Fig. 2a, authors and conferences can be connected via "Author-Paper-Venue-Conference" (*APVC*) path, "Author-Paper-Subject-Paper-Venue-Conference" (*APSPVC*) path, and so on. The semantics underneath these two paths are different. The *APVC* path means that papers written by authors are published in conferences, while the *APSPVC* path means that papers having the same subjects with the authors' papers are published in conferences. Obviously, the distinct semantics under different paths will lead to different results. The relatedness under *APVC* path emphasizes the conferences that authors participated, while the

relatedness under *APSPVC* path emphasizes on conferences publishing the papers that have the same subjects with authors' papers. For example, most of Christos Faloutsos's papers are published in the KDD, VLDB, and SIGMOD. However, the papers having the same subjects with his papers may be published in widespread conferences, such as ICDM, SDM, and CIKM. So the relatedness of objects depends on the search path in the heterogeneous networks. Formally, we define the meta search path as the relevance path.

Definition 2 (Relevance Path). *A relevance path $\mathcal{P}$ is a path defined on a schema $S = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$ which defines a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between type A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations. The length of the path $\mathcal{P}$ is the number of relations in $\mathcal{P}$, which is l.*

For simplicity, we can also use type names denoting the relevance path if there are no multiple relations between the same pair of types: $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$. We say a concrete path $p = (a_1 a_2 \cdots a_{l+1})$ between 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 relevance path $\mathcal{P}^{-1}$ is the *reverse path* of $\mathcal{P}$, which defines an inverse relation of the one defined by $\mathcal{P}$. Similarly, we define the *reverse path instance* p^{-1} as the reverse path of p in G. Furthermore, a relevance 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 $APCPA$. Two relevance 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 $(A_1 A_2 \cdots A_l B_2 \cdots B_k)$.

4 HeteSim: A Uniform and Symmetric Relevance Measure

4.1 Basic Idea

In many domains, similar objects are more likely to be related to some other similar objects. For example, similar researchers usually publish many similar papers, and similar customers purchase similar commodities. As a consequence, two objects are similar if they are referenced by similar objects. This intuition is also fit for heterogeneous objects. For example, a researcher is more relevant to the conferences that the researcher has published papers in, and a customer is more faithful to the brands that the customer usually purchases. Although the similar idea has been applied in SimRank [4], it is limited to homogeneous networks. When we apply the idea to heterogeneous networks, it faces the following challenges. (1) The relatedness of heterogeneous objects is path-constrained. The relevance path not only captures the semantics information but also constrains the walk path. So we need to design a path-based similarity measure. (2) A uniform and symmetric measure should be designed for arbitrary paths. For a given path (symmetric or asymmetric), the measure can evaluate the relatedness of heterogeneous object pair (same or different types) with one single score. In the following section, we will illustrate these challenges and their solutions in detail.

4.2 Path-Based Relevance Measure

Different from homogeneous networks, the paths in heterogeneous networks have semantics, which makes the relatedness of object pair depend on the given relevance path. Following the basic idea that similar objects are related to similar objects, we propose a path-based relevance measure: HeteSim.

Definition 3 (HeteSim). *Given a relevance path $\mathcal{P} = R_1 \circ R_2 \circ \cdots \circ R_l$, the HeteSim score between two objects s and t ($s \in R_1.S$ and $t \in R_l.T$) is:*

$$HeteSim(s,t|R_1 \circ R_2 \circ \cdots \circ R_l) = \frac{1}{|O(s|R_1)||I(t|R_l)|} \sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(t|R_l)|} HeteSim(O_i(s|R_1), I_j(t|R_l)|R_2 \circ \cdots \circ R_{l-1}), \quad (1)$$

where $O(s|R_1)$ is the out-neighbors of s based on relation R_1, and $I(t|R_l)$ is the in-neighbors of t based on relation R_l.

When s does not have any out-neighbors (i.e., $O(s|R_1) = \emptyset$) or t does not have any in-neighbors (i.e., $I(t|R_l) = \emptyset$) following the path, we have no way to infer any relatedness between s and t in this case, so we define their relevance score to be 0. Particularly, we consider objects with the same type to have *self-relation* (denoted as I relation) and each object only has self-relation with itself. It is obvious that an object is just similar to itself for I relation. So its relevance measure can be defined as follows:

Definition 4 (HeteSim based on self-relation). *The HeteSim score between two same-typed objects s and t based on the self-relation I is:*

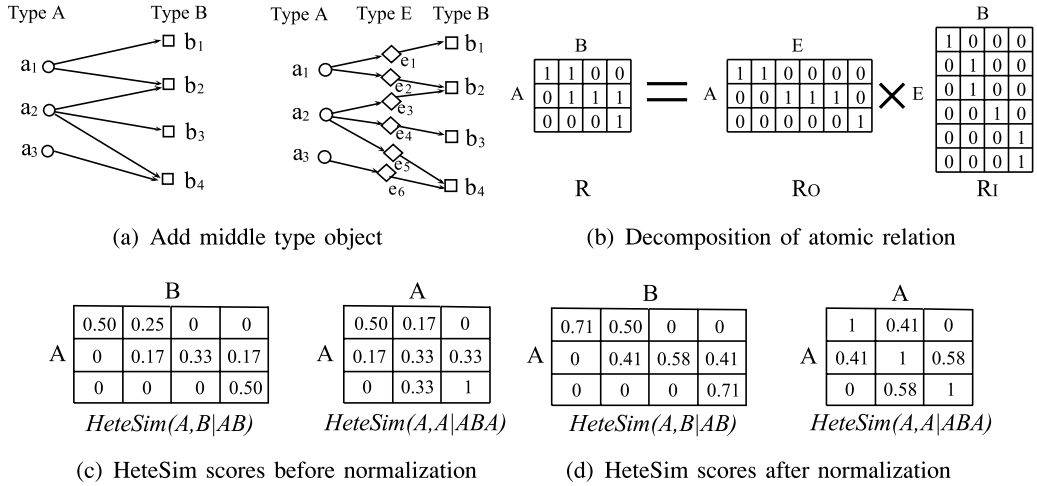
$$HeteSim(s,t|I) = \delta(s,t), \quad (2)$$

where $\delta(s,t) = 1$, if s and t are same, or else $\delta(s,t) = 0$.

Equation (1) shows that the computation of $HeteSim(s,t|\mathcal{P})$ needs to iterate over all pairs $(O_i(s|R_1), I_j(t|R_l))$ of (s,t) along the path (s along the path and t against path), and sum up the relatedness of these pairs. Then, we normalize it by the total number of out-neighbors of s and in-neighbors of t. That is, the relatedness between s and t is the average relatedness between the out-neighbors of s and the in-neighbors of t. The process continues until s and t meet along the path. Similar to SimRank [4], HeteSim is also based on pair wise random walk, while it considers the path constraint. As we know, SimRank measures how soon two random surfers are expected to meet at the same node [4]. By contrast, $HeteSim(s,t|\mathcal{P})$ measures how likely s and t will meet at the same node when s follows along the path and t goes against the path.

4.3 Decomposition of Relevance Path

Unfortunately, the source object s and the target object t may not meet along a given path $\mathcal{P}$. For the similarity measure of same-typed objects, the relevance paths are usually even-length, even symmetric, so the source object and the target object will meet at the middle objects. However, for the relevance measure of different-typed objects, the relevance paths are usually odd-length. In this condition, the source and target objects will never meet at the same objects. Taking the



(a) Add middle type object (b) Decomposition of atomic relation

(c) HeteSim scores before normalization (d) HeteSim scores after normalization

Fig. 3. Decomposition of atomic relation and its HeteSim calculation.

$APVC$ path as an example, authors along the path and conferences against the path will never meet in the same objects. So the original HeteSim is not suitable for odd-length relevance paths. In order to solve this difficulty, a basic idea is to transform odd-length paths into even-length paths, and thus the source and target objects are always able to meet at the same objects. As a consequence, an arbitrary path can be decomposed as two equal-length paths.

When the length l of a relevance path $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$ is even, the source objects (along the path) and the target objects (against the path) will meet in the *middle type* object $M = A_{\frac{l}{2}+1}$ on the **middle position** $mid = \frac{l}{2} + 1$, so the relevance path $\mathcal{P}$ can be divided into two equal-length path $\mathcal{P}_L$ and $\mathcal{P}_R$. That is, $\mathcal{P} = \mathcal{P}_L \mathcal{P}_R$, where $\mathcal{P}_L = A_1 A_2 \cdots A_{mid-1} M$ and $\mathcal{P}_R = M A_{mid+1} \cdots A_{l+1}$.

When the path length l is odd, the source objects and the target objects will meet at the relation $A_{\frac{l+1}{2}} A_{\frac{l+1}{2}+1}$. In order to let the source and target objects meet at same-typed objects, we can add a middle type object E between the atomic relation $A_{\frac{l+1}{2}} A_{\frac{l+1}{2}+1}$ and maintain the relation between $A_{\frac{l+1}{2}}$ and $A_{\frac{l+1}{2}+1}$ at the same time. Then the new path becomes $\mathcal{P}' = (A_1 \cdots E \cdots A_{l+1})$ whose length is $l+1$, an even number. The source objects and the target objects will meet in the *middle type* object $M = E$ on the **middle position** $mid = \frac{l+1}{2} + 1$. As a consequence, the new relevance path $\mathcal{P}'$ can also be decomposed into two equal-length paths $\mathcal{P}_L$ and $\mathcal{P}_R$.

Definition 5 (Decomposition of relevance path). *An arbitrary relevance path $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$ can be decomposed into two equal-length path $\mathcal{P}_L$ and $\mathcal{P}_R$ (i.e., $\mathcal{P} = \mathcal{P}_L \mathcal{P}_R$), where $\mathcal{P}_L = A_1 A_2 \cdots A_{mid-1} M$ and $\mathcal{P}_R = M A_{mid+1} \cdots A_{l+1}$. M and mid are defined as above.*

Obviously, for a symmetric path $\mathcal{P} = \mathcal{P}_L \mathcal{P}_R$, $\mathcal{P}_R^{-1}$ is equal to $\mathcal{P}_L$. For example, the relevance path $\mathcal{P} = APCPA$ can be decomposed as $\mathcal{P}_L = APC$ and $\mathcal{P}_R = CPA$. For the relevance path $APSPVC$, we can add a middle type object E in SP and thus the path becomes $APSEPVC$, so $\mathcal{P}_L = APSE$ and $\mathcal{P}_R = EPVC$.

The next question is how we can add the middle type object E in an atomic relation R between $A_{\frac{l+1}{2}}$ and $A_{\frac{l+1}{2}+1}$. In order to contain original atomic relation, we need to make the R relation be the composition of two new relations. To do so, for each instance of relation R, we can add an instance of E to connect the source and target objects of the relation instance. An example is shown in Fig. 3a, where the middle type object E is added in between the atomic relation AB along each path instance.

Definition 6 (Decomposition of atomic relation). *For an atomic relation R, we can add an object type E (called edge object) between the $R.S$ and $R.T$. And thus the atomic relation R is decomposed as R_O and R_I where R_O represents the relation between $R.S$ and E and R_I represents that between E and $R.T$. For each relation instance $r \in R$, an instance $e \in E$ connects $r.S$ and $r.T$. The paths $r.S \to e$ and $e \to r.T$ are the instances of R_O and R_I, respectively.*

It is clear that the relation decomposition has the following property, whose proof can be found in the Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.2297920.

Property 1. *An atomic relation R can be decomposed as R_O and R_I, $R = R_O \circ R_I$, and this decomposition is unique.*

Based on this decomposition, the relatedness of two objects with an atomic relation R can be calculated as follows:

Definition 7 (HeteSim based on atomic relation). *The HeteSim score between two different-typed objects s and t based on an atomic relation R ($s \in R.S$ and $t \in R.T$) is:*

$$\begin{aligned} HeteSim(s,t|R) &= HeteSim(s,t|R_O \circ R_I) \\ &= \frac{1}{|O(s|R_O)||I(t|R_I)|} \sum_{i=1}^{|O(s|R_O)|} \sum_{j=1}^{|I(t|R_I)|} \\ &\quad \delta(O_i(s|R_O), I_j(t|R_I)). \end{aligned} \tag{3}$$

It is easy to find that $HeteSim(s,t|I)$ is a special case of $HeteSim(s,t|R)$, since, for the self-relation I, $I = I_O \circ I_I$ and $|O(s|I_O)| = |I(t|I_I)| = 1$. Definition 7 means that HeteSim can measure the relatedness of two different-typed objects

with an atomic relation R directly through calculating the average of their mutual influence.

Example 2. Fig. 3a shows an example of decomposition of atomic relation. The relation AB is decomposed into the relations AE and EB. Moreover, the relation AB is the composition of AE and EB as shown in Fig. 3b. Two HeteSim examples are illustrated in Fig. 3c. We can find that HeteSim justly reflects relatedness of objects. Taking a_2 as example, although a_2 equally connects with b_2, b_3, and b_4, it is more close to b_3, because b_3 only connects with a_2. This information is correctly reflected in the HeteSim score of a_2 based on AB path.

We also find that the similarity of an object and itself is not 1 in HeteSim. Taking the right figure of Fig. 3c as example, the relatedness of a_2 and itself is 0.33. It is obviously unreasonable. In the following section, we will normalize the HeteSim and make the relevance measure more reasonable.

4.4 Normalization of HeteSim

Firstly, we introduce the calculation of HeteSim between any two objects given an arbitrary relevance path.

Definition 8 (Transition probability matrix). *For relation $A \xrightarrow{R} B$, W_{AB} is an adjacent matrix between type A and B. U_{AB} is a normalized matrix of W_{AB} along the row vector, which is the transition probability matrix of $A \longrightarrow B$ based on relation R. V_{AB} is a normalized matrix of W_{AB} along the column vector, which is the transition probability matrix of $B \longrightarrow A$ based on relation R^{-1}.*

It is easy to prove that the transition probability matrix has the following property. The proof can be found in the Appendix A, available in the online supplemental material.

Property 2. *$U_{AB} = V'_{BA}$ and $V_{AB} = U'_{BA}$, where V'_{BA} is the transpose of V_{BA}.*

Definition 9 (Reachable probability matrix). *Given a network $G = (V, E)$ following a network schema $S = (\mathcal{A}, \mathcal{R})$, a reachable probability matrix PM for a path $\mathcal{P} = (A_1 A_2 \cdots A_{l+1})$ is defined as $PM_{\mathcal{P}} = U_{A_1A_2} U_{A_2A_3} \cdots U_{A_lA_{l+1}}$ (PM for simplicity). $PM(i, j)$ represents the probability of object $i \in A_1$ reaching object $j \in A_{l+1}$ under the path $\mathcal{P}$.*

According to the definition and Property 2 of HeteSim, the relevance between objects in A_1 and A_{l+1} based on the relevance path $\mathcal{P} = A_1 A_2 \cdots A_{l+1}$ is

$$\begin{aligned} HeteSim(A_1, A_{l+1}|\mathcal{P}) &= HeteSim(A_1, A_{l+1}|\mathcal{P}_L\mathcal{P}_R) \\ &= U_{A_1A_2} \cdots U_{A_{mid-1}M} V_{MA_{mid+1}} \cdots V_{A_lA_{l+1}} \\ &= U_{A_1A_2} \cdots U_{A_{mid-1}M} U'_{A_{mid+1}M} \cdots U'_{A_{l+1}A_l} \\ &= U_{A_1A_2} \cdots U_{A_{mid-1}M} (U_{A_{l+1}A_l} \cdots U_{A_{mid+1}M})' \\ &= PM_{\mathcal{P}_L} PM'_{\mathcal{P}_R{}^{-1}}. \end{aligned} \tag{4}$$

The above equation shows that the relevance of A_1 and A_{l+1} based on the path $\mathcal{P}$ is the inner product of two probability distributions that A_1 reaches the middle type object M along the path and A_{l+1} reaches M against the path. For two instances a and b in A_1 and A_{l+1}, respectively, their relevance based on path $\mathcal{P}$ is

$$HeteSim(a, b|\mathcal{P}) = PM_{\mathcal{P}_L}(a, :) PM'_{\mathcal{P}_R{}^{-1}}(b, :), \tag{5}$$

where $PM_{\mathcal{P}}(a, :)$ means the ath row in $PM_{\mathcal{P}}$.

We have stated that HeteSim needs to be normalized. It is reasonable that the relatedness of the same objects is 1, so the HeteSim can be normalized as follows:

Definition 10 (Normalization of HeteSim). *The normalized HeteSim score between two objects a and b based on the relevance path $\mathcal{P}$ is:*

$$HeteSim(a, b|\mathcal{P}) = \frac{PM_{\mathcal{P}_L}(a, :) PM'_{\mathcal{P}_R{}^{-1}}(b, :)}{\sqrt{\|PM_{\mathcal{P}_L}(a, :)\| \|PM'_{\mathcal{P}_R{}^{-1}}(b, :)\|}}. \tag{6}$$

In fact, the normalized HeteSim is the cosine of the probability distributions of the source object a and target object b reaching the middle type object M. It ranges from 0 to 1. Fig. 3d shows the normalized HeteSim scores. It is clear that the normalized HeteSim is more reasonable. The normalization is an important step for HeteSim with the following advantages. (1) The normalized HeteSim has nice properties. The following Property 4 shows that HeteSim satisfies the identity of indiscernibles. (2) It has nice interpretation. The normalized HeteSim is the cosine of two vectors representing reachable probability. As Fouss et al. pointed out [21], the angle between the node vectors is a much more predictive measure than the distance between the nodes. In the following section, the HeteSim means the normalized HeteSim.

4.5 Properties of HeteSim

HeteSim has good properties, which makes it useful in many applications. The proof of these properties can be found in the Appendix A, available in the online supplemental material.

Property 3 (Symmetric). $HeteSim(a, b|\mathcal{P}) = HeteSim(b, a|\mathcal{P}^{-1})$.

Property 3 shows the symmetric property of HeteSim. Although PathSim [5] also has the similar symmetric property, it holds only when the path is symmetric and a and b are with the same type. The HeteSim has the more general symmetric property not only for symmetric paths (note that $\mathcal{P}$ is equal to $\mathcal{P}^{-1}$ for symmetric paths) but also for asymmetric paths.

Property 4 (Self-maximum). $HeteSim(a, b|\mathcal{P}) \in [0, 1]$. *$HeteSim(a, b|\mathcal{P})$ is equal to 1 if and only if $PM_{\mathcal{P}_L}(a, :)$ is equal to $PM_{\mathcal{P}_R{}^{-1}}(b, :)$.*

Property 4 shows HeteSim is well constrained. For a symmetric path $\mathcal{P}$ (i.e., $\mathcal{P}_L = \mathcal{P}_R{}^{-1}$), $PM_{\mathcal{P}_L}(a, :)$ is equal to $PM_{\mathcal{P}_R{}^{-1}}(a, :)$, and thus $HeteSim(a, a|\mathcal{P})$ is equal to 1. If we define the distance between two objects (i.e., $dis(s, t)$) as $dis(s, t) = 1 - HeteSim(s, t)$, the distance of the same object is zero (i.e., $dis(s, s) = 0$). As a consequence, HeteSim satisfies the identity of indiscernibles. Note that it is a general identity of indiscernibles. For two objects with different types, their HeteSim score is also 1 if they have the same probability distribution on the middle type object. It is reasonable, since they have the similar structure based on the given path.

Since HeteSim obeys the properties of non-negativity, identity of indiscernibles, and symmetry, we can say that HeteSim is a semi-metric measure [23]. Because of a path-based measure, HeteSim does not obey the triangle inequality. A semi-metric measure has many good merits and can be widely used in many applications [23].

TABLE 1
Comparison of Different Similarity Measures

	Symmetry	Triangle Inequation	Path based	Model	Features
HeteSim	√	×	√	PRW	evaluate relevance of heterogeneous objects based on arbitrary path
PathSim[5]	√	×	√	Path Count	evaluate similarity of same-typed objects based on symmetric path
PCWR[11]	×	×	√	RW	measure proximity to the query nodes based on given path
SimRank[4]	√	×	×	PRW	measure similarity of node pairs based on the similarity of their neighbors
RoleSim[17]	√	√	×	PRW	measure real-valued role similarity based on automorphic equivalence
P-PageRank[3]	×	×	×	RW	measure personalized views of importance based on linkage strcutre

Property 5 (Connection to SimRank). *For a bipartite graph $G = (V, E)$ based on the schema $S = (\{A, B\}, \{R\})$, suppose the constant* C *in SimRank is 1,*

$$SimRank(a_1, a_2) = \lim_{n \rightsquigarrow \infty} \sum_{k=1}^{n} HeteSim(a_1, a_2 \,|\, (RR^{-1})^k),$$

$$SimRank(b_1, b_2) = \lim_{n \rightsquigarrow \infty} \sum_{k=1}^{n} HeteSim(b_1, b_2 \,|\, (R^{-1}R)^k),$$

where $a_1, a_2 \in A$, $b_1, b_2 \in B$ and $A \xrightarrow{R} B$. Here HeteSim is the non-normalized version.

This property reveals the connection of SimRank and HeteSim. SimRank sums up the meeting probability of two objects after all possible steps. HeteSim just calculates the meeting probability along the given relevance path. If the relevance paths explore all possible meta paths among the two types of objects, the sum of HeteSim based on these paths is the SimRank. So we can say that HeteSim is a path-constrained version of SimRank. Through relevance paths, HeteSim can subtly evaluate the similarity of heterogeneous objects with fine granularity. This property also implies that HeteSim is more efficient than SimRank, since HeteSim only needs to calculate the meeting probability along the given relevance path, not all possible meta paths.

4.6 Discussion

It is a big issue for heterogeneous networks to choose relevance path. There are several ways to do it. (1) Users can select proper paths according to their domain knowledge and experiences. (2) Supervised learning can be used to automatically determine the importance of relevance paths. In information retrieval field, Lao and Cohen [22] proposed a learnable proximity measure where proximity is defined by a weighted combination of simple "path experts". Through labeled training data, a learning algorithm can infer the weights of paths. The similar strategy can also be used for path selection. (3) Recently, Sun et al. [24] combined meta path selection and user-guided information for clustering in heterogeneous networks. The similar user-guided information can also been applied in the selection of relevance paths in HeteSim.

There are numbers of similarity measures, most of which are based on three basic strategies [5]: (1) Path count strategy measures the number of path instances connecting source and target objects; (2) Random walk strategy measures the probability of the random walk from source to target objects; and (3) Pairwise random walk (PRW) strategy measures the pairwise random walk probability starting from source and target objects and reaching the same middle objects. Due to symmetry and arbitrary path constraints, we employ the PRW model in this work. Although the RW model can also satisfy the symmetric property through the combination of the reachable probability based on the paths $\mathcal{P}$ and $\mathcal{P}^{-1}$, it is redundant for symmetric path, as well as lacks of nice interpretation. For the PRW model, it has to face the problem that the source and target object will not meet when the length of relevance path is odd. In order to solve it, some strategies need to be devised, such as assigning the meeting object type. This paper adopts the path decomposition strategy based on the following advantages. (1) It has a uniform framework to evaluate the relevance of same or different-typed objects for arbitrary paths. (2) It provides a simple but effective method to evaluate the relevance of two different-typed objects based on an atomic relation (see Definition 7).

Furthermore, we compare six well-established similarity measures in Table 1. There are three similarity measures for heterogeneous networks (i.e., HeteSim, PathSim, and PCWR) and three measures for homogeneous networks (i.e., P-PageRank, SimRank, and RoleSim), respectively. Although these similarity measures all evaluate the similarity of nodes by utilizing network structure, they have different properties and features. Three measures for heterogeneous networks all are path-based, since meta paths in heterogeneous networks embody semantics and simplify network structure. Two RW model based measures (i.e., P-PageRank and PCRW) do not satisfy the symmetric property. Because of satisfying the triangle inequation, RoleSim is a metric, while HeteSim, PathSim, and SimRank are semi-metric. Different from PathSim, which can only measure the similarity of objects with the same type under symmetric paths, the proposed HeteSim can measure the relevance of heterogeneous (same or different-typed) objects under arbitrary (symmetric or asymmetric) paths. Although HeteSim can be considered as a path-constrained extension of SimRank, HeteSim is a general similarity measure in heterogeneous networks with arbitrary schema, not limited to bipartite or N-partite networks.

5 Experiments

In the experiments, we validate the effectiveness of the HeteSim on three data sets with three case studies and two learning tasks.

TABLE 2
Automatic Object Profiling Task on Author "Christos Faloutsos" on ACM Data Set

Path	*APVC*		*APT*		*APS*		*APA*	
Rank	Conf.	Score	Terms	Score	Subjects	Score	Authors	Score
1	KDD	0.1198	mining	0.0930	H.2 (database management)	0.1023	Christos Faloutsos	1
2	SIGMOD	0.0284	patterns	0.0926	E.2 (data storage representations)	0.0232	Hanghang Tong	0.4152
3	VLDB	0.0262	scalable	0.0869	G.3 (probability and statistics)	0.0175	Agma Juci M. Traina	0.3250
4	CIKM	0.0083	graphs	0.0816	H.3 (information storage and retrieval)	0.0136	Spiros Papadimitriou	0.2785
5	WWW	0.0060	social	0.0672	H.1 (models and principles)	0.0135	Caetano Traina, Jr.	0.2680

5.1 Data Sets

Three heterogeneous information networks are employed in our experiments.

ACM data set: The ACM data set was downloaded from ACM digital library[3] in June 2010. The ACM data set comes from 14 representative computer science conferences: KDD, SIGMOD, WWW, SIGIR, CIKM, SODA, STOC, SOSP, SPAA, SIGCOMM, MobiCOMM, ICML, COLT, and VLDB. These conferences include 196 corresponding venue proceedings. The data set has 12K papers, 17K authors, and 1.8K author affiliations. After removing stop words in the paper titles and abstracts, we get 1.5K terms that appear in more than 1 percent of the papers. The network also includes 73 subjects of these papers in ACM category. The network schema of ACM data set is shown in Fig. 2a. Furthermore, we label the data with the ACM category (i.e., subjects) information. That is, with three major subjects (i.e., H.3, H.2, and C.2), we label seven conferences, 6,772 authors, and 4,526 papers.

DBLP data set [25]: The DBLP data set is a sub-network collected from DBLP website[4] involving major conferences in four research areas: database, data mining, information retrieval and artificial intelligence, which naturally form four classes. The data set contains 14K papers, 20 conferences, 14K authors and 8.9K terms, with a total number of 17K links. In the data set, 4,057 authors, all 20 conferences and 100 papers are labeled with one of the four research areas. The network schema is shown in Fig. 2b.

Movie data set [26]: The IMDB movie data comes from the Internet Movie Database,[5] which includes movies, actors, directors and types. A movie heterogeneous network is constructed from the movie data and its schema is shown in Fig. 2c. The movie data contains 1.5K movies, 5K actors, 551 directors, and 112 types.

5.2 Case Study

In this section, we demonstrate the traits of HeteSim through case study in three tasks: automatic object profiling, expert finding, relevance search.

5.2.1 Task 1: Automatic Object Profiling

We first study the effectiveness of HeteSim on different-typed relevance measurement in the automatic object profiling task. If we want to know the profile of an object, we can measure the relevance of the object to objects that we are interested in. For example, the academic profile of Christos Faloutsos[6] can be constructed through measuring the relatedness of Christos Faloutsos with related objects, e.g., conferences, affiliations, other authors, etc. Table 2 shows the lists of top relevant objects with various types on ACM data set. $APVC$ path shows the conferences he actively participates. Note that KDD and SIGMOD are the two major conferences Christos Faloutsos participates, which are mentioned in his homepage.[7] From the path APT, we can obtain his research interests: data mining, pattern discovery, scalable graph mining and social network. Using APS path, we can discover his research areas represented as ACM subjects: database management (H.2) and data storage (E.2). Based on APA path, HeteSim finds the most important co-authors, most of which are his PhD students. Another interesting case about the KDD conference profile can be seen in Appendix B, available in the online supplemental material.

5.2.2 Task 2: Expert Finding

In this case, we want to validate the effectiveness of HeteSim to reflect the relative importance of object pairs through an expert finding task. As we know, the relative importance of object pairs can be revealed through comparing their relatedness. Suppose we know the experts in one domain, the expert finding task here is to find experts in other domains through their relative importance. Table 3 shows the relevance scores returned by HeteSim and PCRW on six "conference-author" pairs on ACM data set. The relatedness of conferences and authors are defined based on the $APVC$ and $CVPA$ paths which have the same semantics: authors publishing papers in conferences. Due to the symmetric property, HeteSim returns the same value for both paths, while PCRW returns different values for these two paths. Suppose that we are familar with data mining area, and already know that C. Faloutsos is an influential researcher in KDD. Comparing these HeteSim scores, we can find influential researchers in other research areas even if we are not quite familiar with these areas. J.F. Naughton, W.B. Croft and A. Gupta should be influential researchers in SIGMOD, SIGIR and SODA, respectively, since they have very similar HeteSim scores to C. Faloutsos. Moreover, we can also deduce that Luo Si and Yan Chen may be active researchers in SIGIR and SIGCOMM, respectively, since

3. http://dl.acm.org/.
4. http://www.informatik.uni-trier.de/~ley/db/.
5. www.imdb.com/.
6. http://www.cs.cmu.edu/~christos/.
7. http://www.cs.cmu.edu/~christos/misc.html.

TABLE 3
Relatedness Scores of Authors and Conferences Measured by HeteSim and PCRW on ACM Data Set

HeteSim		PCRW					
APVC&CVPA		*APVC*			*CVPA*		
Pair	Score	Pair	Score		Pair		Score
C. Faloutsos, KDD	**0.1198**	C. Faloutsos, KDD	0.5517		KDD, C. Faloutsos		**0.0087**
W. B. Croft, SIGIR	**0.1201**	W. B. Croft, SIGIR	0.6481		SIGIR, W. B. Croft		**0.0098**
J. F. Naughton, SIGMOD	**0.1185**	J. F. Naughton, SIGMOD	**0.7647**		SIGMOD, J. F. Naughton		0.0062
A. Gupta, SODA	**0.1225**	A. Gupta, SODA	**0.7647**		SODA, A. Gupta		**0.0090**
Luo Si, SIGIR	0.0734	Luo Si, SIGIR	**0.7059**		SIGIR, Luo Si		0.0030
Yan Chen, SIGCOMM	0.0786	Yan Chen, SIGCOMM	**1**		SIGCOMM, Yan Chen		0.0013

they have moderate HeteSim scores. In fact, C. Faloutsos, J.F. Naughton, W.B. Croft and A. Gupta are top ranked authors in their research communities. Luo Si and Yan Chen are the young professors and they have done good work in their research areas. However, if the relevance measure is not symmetric (e.g., PCRW), it is very hard to tell which authors are more influential when comparing these relevance scores. For example, the PCRW score of Yan Chen and SIGCOMM is the largest one in the *APVC* path. However, the value is the smallest one for the reversed path (i.e., *CVPA* path). A quantitative experiment in the Appendix C, available in the online supplemental material, illustrates that, compared to PCRW, HeteSim can reveal the relative importance of author-conference pairs more accurately.

5.2.3 Task 3: Relevance Search Based on Path Semantics

As we have stated, the path-based relevance measure can capture the semantics of paths. In this relevance search task, we will observe the importance of paths and the effectiveness of semantics capture through the comparison of three path-based measures (i.e., HeteSim, PCRW, and PathSim) and SimRank. This task is to find the top 10 related authors to Christos Faloutsos based on the *APVCVPA* path which means authors publishing papers in same conferences. By ignoring the heterogeneity of objects, we directly run SimRank on whole network and select top ten authors from the rank results which mix different-typed objects together. The comparison results are shown in Table 4. At first sight, we can find that three path-based measures all return researchers having the similar reputation with C. Faloutsos in slightly different orders. However, the results of SimRank are totally against our common sense. We think the reason of bad performances is that SimRank only considers link structure but ignores the link semantics.

In addition, let's analyze the subtle differences of results returned by three path-based measures. The PathSim finds the similar peer authors, such as P. Yu and J. Han. They have the same reputation in data mining field. It is strange for PCRW that the most similar author to C. Faloutsos is not himself, but C. Aggarwal and J. Han. It is obviously not reasonable. Our conjecture is that C. Aggarwal and J. Han published many papers in the conferences that C. Faloutsos participated in, so C. Faloutsos has more reachable probability on C. Aggarwal and J. Han than himself along the *APVCVPA* path. HeteSim's results are a little different. The most similar authors are S. Parthasarathy and X. Yan, instead of P. Yu and J. Han. Let's revisit the semantics of the path *APVCVPA*: authors publishing papers in the same conferences. Fig. 4 shows the reachable probability distribution from authors to

TABLE 4
Top 10 Related Authors to "Christos Faloutsos" Based on *APVCVPA* Path on ACM Data Set

	HeteSim		PathSim		PCRW		SimRank	
Rank	Author	Score	Author	Score	Author	Score	Author	Score
1	Christos Faloutsos	1	Christos Faloutsos	1	Charu C. Aggarwal	0.0063	Christos Faloutsos	1
2	Srinivasan Parthasarathy	0.9937	Philip Yu	0.9376	Jiawei Han	0.0061	Edoardo Airoldi	0.0789
3	Xifeng Yan	0.9877	Jiawei Han	0.9346	Christos Faloutsos	0.0058	Leejay Wu	0.0767
4	Jian Pei	0.9857	Jian Pei	0.8956	Philip Yu	0.0056	Kensuke Onuma	0.0758
5	Jiong Yang	0.9810	Charu C. Aggarwal	0.7102	Alia I. Abdelmoty	0.0053	Christopher R. Palmer	0.0699
6	Ruoming Jin	0.9758	Jieping Ye	0.6930	Chris B. Jones	0.0053	Anthony Brockwell	0.0668
7	Wei Fan	0.9743	Heikki Mannila	0.6928	Jian Pei	0.0034	Hanghang Tong	0.0658
8	Evimaria Terzi	0.9695	Eamonn Keogh	0.6704	Heikki Mannila	0.0032	Evan Hoke	0.0651
9	Charu C. Aggarwal	0.9668	Ravi Kumar	0.6378	Eamonn Keogh	0.0031	Jia-Yu Pan	0.0650
10	Mohammed J. Zaki	0.9645	Vipin Kumar	0.6362	Mohammed J. Zaki	0.0027	Roberto Santos Filho	0.0648

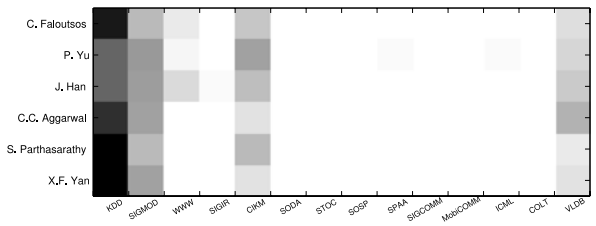


Fig. 4. Probability distribution of authors' papers on 14 conferences of ACM data set.

conferences along the path $APVC$. It is clear that the probability distribution of papers of S. Parthasarathy and X. Yan on conferences are more close to that of C. Faloutsos, so they should be more similar to C. Faloutsos based on the same conference publication. Although P. Yu and J. Han have the same reputation with C. Faloutsos, their papers are more broadly published in different conferences. So they are not the most similar authors to C. Faloutsos based on the $APVCVPA$ path. As a consequence, the HeteSim more accurately captures the semantics of the path.

Since relevance path can embody semantics, we can apply HeteSim to do semantic recommendation based on paths given by users. Due to space limitation, we illustrate such a case study on IMDB movie data set in the Appendix D, available in the online supplemental material. Following this idea, a semantic-based recommendation system HeteRecom [26] has been designed.

5.3 Performance on Query Task

The query task will validate the effectiveness of HeteSim on query search of heterogeneous objects. Since PathSim cannot measure the relatedness of different-typed objects, we only compare HeteSim with PCRW in this experiment. On DBLP data set, we measure the proximity of conferences and authors based on the CPA and $CPAPA$ paths. For each conference, we rank its related authors according to their measure scores. Then we draw the ROC curve of top 100 authors according to the labels of authors (when the labels of author and conference are the same, it is true, else it is false). After that, we calculate the AUC (Area Under ROC Curve) score to evaluate the performances of the ranked results. Note that all conferences and some authors on the DBLP data set are labeled with one of the four research areas (see Section 5.1). The larger score means the better performance. We evaluate the performances on nine representative conferences and their AUC scores are shown in Table 5. We can find that HeteSim consistently outperforms PCRW in most conferences under these two paths. It shows that the proposed HeteSim method can work better than the asymmetric similarity measure PCRW on proximity query task.

5.4 Performance on Clustering Task

Due to the symmetric property, HeteSim can be applied to clustering tasks directly. In order to evaluate its performance, we compare HeteSim with five well-established similarity measures, including two path-based measures (i.e., PathSim and PCRW) and three homogeneous measures (i.e., SimRank, RoleSim, and P-PageRank). These measures use the same information to determine the pairwise similarity between objects. We evaluate the clustering performances on DBLP and ACM data sets. There are three tasks: conference clustering based on $CPAPC$ path, author clustering based on $APCPA$ path, and paper clustering based on $PAPCPAP$ path. For asymmetric measures (i.e., PCRW and P-PageRank), the symmetric similarity matrix can be obtained through the average of similarity matrices based on paths $\mathcal{P}$ and $\mathcal{P}^{-1}$. For RoleSim, it is applied in the network constructed by path $\mathcal{P}$. For SimRank and P-PageRank, they are applied in the subnetwork constructed by path $\mathcal{P}_L$ (note that the three paths in the experiments are symmetric). Then we apply Normalized Cut [27] to perform clustering based on the similarity matrices obtained by different measures. The number of clusters are set as 4 and 3 for DBLP and ACM data sets, respectively. The NMI criterion (Normalized Mutual Information) [28] is used to evaluate the clustering performances on conferences, authors, and papers. NMI is between 0 and 1 and the higher the better. In experiments, the damping factors for P-PageRank, SimRank, and RoleSim are set as 0.9, 0.8, and 0.1, respectively.

The average clustering accuracy results of 100 runs are summarized in Table 6. We can find that, on all six tasks, HeteSim achieves best performances on four of them as well as good performances on other two tasks. The mediocre results of PCWR and P-PageRank illustrate that, although symmetric similarity measures can be constructed by the combination of two random walk processes, the simple combination cannot generate good similarity measures. RoleSim aims to detect role similarity, a little bit different from structure similarity, so it has bad performances in these clustering tasks. The experiments show that HeteSim not only does well on similarity measure of same-typed objects but also has the potential as the similarity measure in clustering.

6 Quick Computation Strategies and Experiments

HeteSim has a high computation demand for time and space. It is not affordable for online query in large-scale

TABLE 5
AUC Values for the Relevance Search of Conferences and Authors Based on Different Paths on DBLP Data Set

Paths	Methods	KDD	ICDM	SDM	SIGMOD	VLDB	ICDE	AAAI	IJCAI	SIGIR
CPA	HeteSim	0.811	0.675	0.950	0.766	0.826	0.732	0.811	0.875	0.613
	PCRW	0.803	0.673	0.939	0.758	0.820	0.726	0.806	0.871	0.606
CPAPA	HeteSim	0.845	0.767	0.715	0.831	0.872	0.791	0.817	0.895	0.952
	PCRW	0.844	0.762	0.710	0.822	0.886	0.789	0.807	0.900	0.949

TABLE 6
Comparison of Clustering Performances for Similarity Measures on DBLP and ACM Data Sets

Methods	DBLP dataset						ACM dataset					
	Venue NMI		Author NMI		Paper NMI		Venue NMI		Author NMI		Paper NMI	
	Mean	Dev.	Mean	Dev.	Mean	Dev.	Mean	Dev.	Mean	Dev.	Mean	Dev.
HeteSim	0.768	0.071	**0.728**	0.083	**0.498**	0.067	**0.843**	0.140	0.405	0.1	**0.439**	0.063
PathSim	0.816	0.078	0.672	0.085	0.383	0.058	0.785	0.164	0.378	0.091	0.432	0.087
PCRW	0.709	0.072	0.710	0.080	0.488	0.039	0.840	0.141	**0.414**	0.092	0.429	0.074
SimRank	**0.888**	0.092	0.685	0.066	0.469	0.031	0.835	0.139	0.375	0.115	0.410	0.073
RoleSim	0.278	0.034	0.501	0.040	0.388	0.049	0.389	0.095	0.293	0.016	0.304	0.017
P-PageRank	0.731	0.086	0.441	0.001	0.421	0.063	0.840	0.164	0.363	0.104	0.407	0.093

information networks. So a primary strategy is to compute relevance matrix off-line and do online queries with these matrices. For frequently-used relevance paths, the relatedness matrix $HeteSim(A, B|\mathcal{P})$ can be materialized ahead of time. The online query on $HeteSim(a, B|\mathcal{P})$ will be very fast, since it only needs to locate the row and column in the matrix. However, it also costs much time and space to materialize all frequently-used paths. As a consequence, we propose four strategies to fast compute the relevance matrix. Moreover, experiments validate the effectiveness of these strategies.

6.1 Quick Computation Strategies

The computation of HeteSim includes two phases: matrix multiplication (denoted as MUL, i.e., the computation of $PM_{\mathcal{P}_L}$ and $PM_{\mathcal{P}_R^{-1}}$), relevance computation (denoted as REL, i.e., the computation of $PM_{\mathcal{P}_L} * PM_{\mathcal{P}_R^{-1}}$ and normalization). Through analyzing the running time of HeteSim on different phases and paths (the details can be seen in Appendix E, available in the online supplemental material), we find two characteristics of HeteSim computation. (1) The relevance computation is the main time-consuming phase. It implies that the speedup of matrix multiplication may not significantly reduce HeteSim's running time, although this kind of strategies is widely used in accelerating SimRank [4] and PCWR [22]. (2) The dimension and sparsity of matrix greatly affect the efficiency of HeteSim. Although we cannot reduce the running time of relevance computation phase directly, we can accelerate the computation of HeteSim through adjusting matrix dimension and keeping matrix sparse. Based on above idea, we design the following four quick computation strategies.

6.1.1 Dynamic Programming Strategy (DP)

The matrix multiplication obeys the associative property. Moreover, different computation sequences have different time complexities. The Dynamic Programming strategy changes the sequence of matrix multiplication with the associative property. The basic idea of DP is to assign low-dimensioned matrix with the high computation priority. For a path $\mathcal{P} = R_1 \circ R_2 \circ \cdots \circ R_l$, the expected minimal computation complexity of HeteSim can be calculated by the following equation and the computation sequence is recorded by i

$$Com(R_1 \cdots R_l) = \begin{cases} 0 & l = 1, \\ |R_1.S| \times |R_1.T| \times |R_2.T| & l = 2, \\ \arg\min_i \{Com(R_1 \cdots R_i) + Com(R_{i+1} \cdots R_l) & \\ \quad + |R_1.S| \times |R_i.T| \times |R_l.T|\} & l > 2. \end{cases} \tag{7}$$

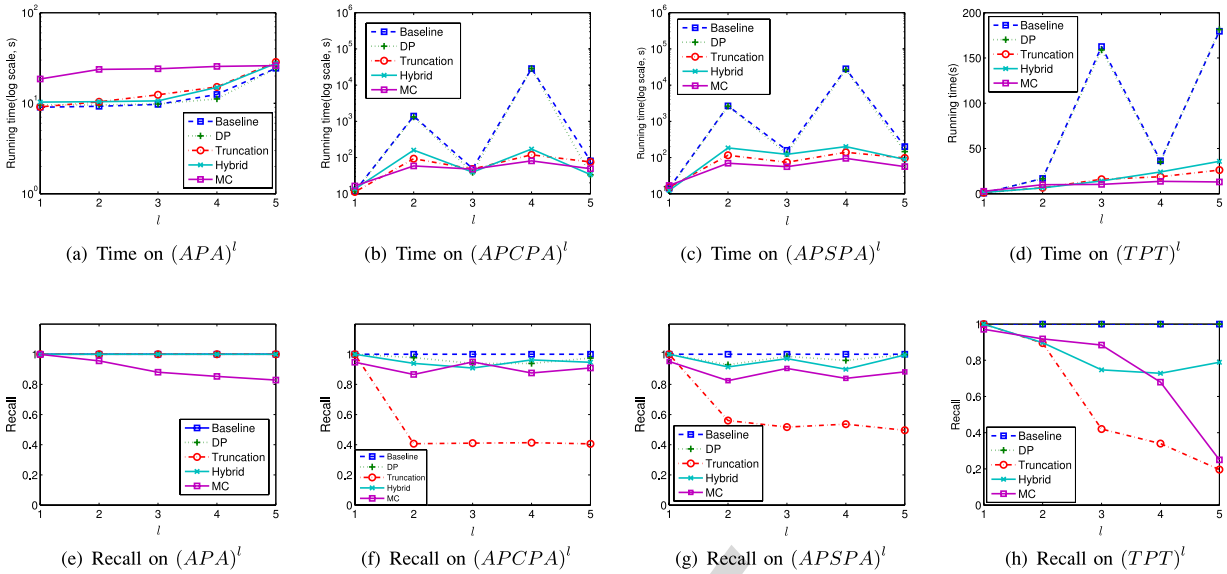
The above equation can be easily solved by dynamic programming method with the $O(l^2)$ complexity. The running time can be omitted, since l is much smaller than the matrix dimension. Note that the DP strategy only accelerates the MUL phase (i.e., matrix multiplication) and it does not change relevance result, so the DP is an information-lossless strategy.

6.1.2 Truncation Strategy

The truncation strategy is based on the hypothesis that removing the probability on those less important nodes would not significantly degrade the performance, which has been proved by many researches [22]. One advantage of this strategy is to keep matrix sparse. The sparse matrix greatly reduces the amount of space and time consumption. The basic idea of truncation strategy is to add a truncation step at each step of random walk. In the truncation step, the relevance value is set with 0 for those nodes when their relevance values are smaller than a threshold ε. A static threshold is usually used in many methods (e.g., ref. [22]). However, it has the following disadvantage: it may truncate nothing for matrix whose elements all have high probability and it may truncate most nodes for matrix whose elements all have low probability. Since we usually pay close attention to the top k objects in query task, the threshold ε can be set as the top k relevance value for each search object. For a similarity matrix with size $M \times L$, the k can be dynamically adjusted as follows:

$$k = \begin{cases} L & \text{if } L \leq W, \\ \lfloor (L - W)^{\beta} \rfloor + W(\beta \in [0, 1]) & \text{others}, \end{cases}$$

where W is the number of top objects, decided by users. The basic idea of dynamic adjustment is that the k slowly increases for super object type (i.e., L is large). The W and β determine the truncation level. The larger W or β will cause the larger k, which means a denser matrix. It is expensive to determine the top k relevance value for each object, so we can estimate



(a) Time on $(APA)^l$ (b) Time on $(APCPA)^l$ (c) Time on $(APSPA)^l$ (d) Time on $(TPT)^l$

(e) Recall on $(APA)^l$ (f) Recall on $(APCPA)^l$ (g) Recall on $(APSPA)^l$ (h) Recall on $(TPT)^l$

Fig. 5. Running time and accuracy of computing HeteSim based on different strategies and paths.

the value by the top kM value for the whole matrix. Furtherly, the top kM value can be approximated by the sample data with ratio γ from the raw matrix. The larger γ leads to more accurate approximation with longer running time. In summary, the truncation strategy is an information-loss strategy, which keeps matrix sparse with small sacrifice on accuracy. In addition, it needs additional time to estimate the threshold ε.

6.1.3 Hybrid Strategy

As discussed above, the DP strategy can accelerate the MUL phase and the truncation strategy can indirectly speed up the REL phase by keeping sparse matrix. So a hybrid strategy can be designed to combine these two strategies. For the MUL phase, the DP strategy is applied. After obtaining the $PM_{\mathcal{P}_L}$ and $PM_{\mathcal{P}_R{}^{-1}}$, the truncation strategy is added. Different from the above truncation strategy, the hybrid strategy only truncates the $PM_{\mathcal{P}_L}$ and $PM_{\mathcal{P}_R{}^{-1}}$. The hybrid strategy utilizes the benefits of DP and truncation strategies. It is also an information-loss strategy, since the truncation strategy is employed.

6.1.4 Monte Carlo (MC) Strategy

Monte Carlo method is a class of computational algorithms that estimate results through repeating random sampling. It has been applied to compute approximate values of matrix multiplication [22], [29]. In this study, we applied the MC strategy to estimate the value of $PM_{\mathcal{P}_L}$ and $PM_{\mathcal{P}_R{}^{-1}}$. The value of $PM_{\mathcal{P}}(a,b)$ can be approximated by the normalized count of the number of times that the walkers visit the node b from a along the path $\mathcal{P}$

$$PM_{\mathcal{P}}(a,b) = \frac{\#times\ the\ walkers\ visit\ b\ along\ \mathcal{P}}{\#walkers\ from\ a}.$$

The number of walkers from a (i.e., K) controls the accuracy and amount of computation. The larger K will achieve more accurate estimation with more time cost. An advantage of the MC strategy is that its running time is not affected by the dimension and sparsity of matrix. However, the high-dimension matrix needs larger K for high accuracy. As a sampling method, the MC is also an information-loss strategy.

6.2 Quick Computation Experiments

We validate the efficiency and effectiveness of quick computation strategies on the ACM data set. The four paths are used: $(APA)^l$, $(APCPA)^l$, $(APSPA)^l$, and $(TPT)^l$. l means times of path repetition and ranges from 1 to 5. Four quick computation strategies and the original method (i.e., baseline) are employed. The parameters in truncation process are set as follows: the number of top objects W is 200, β is 0.5, and γ is 0.005. The number of walkers (i.e., K) in MC strategy is 500. The running time and accuracy of all strategies are recorded. In the accuracy evaluation, the relevance matrices obtained by the original method are regarded as the baseline. The accuracy is the *recall* criterion on the top 100 objects obtained by each strategy. All experiments are conducted on machines with Intel Xeon 8-Core CPUs of 2.13 GHz and 64 GB RAM.

Fig. 5 shows the running time and accuracy of four strategies on different paths. The running time of these strategies are illustrated in Figs. 5a, 5b, 5c, and 5d. We can observe that the DP strategy almost has the same running time with the baseline. It only speeds up the HeteSim computation when the MUL phase dominates the whole running time (e.g., $(APCPA)^5$ and $(APSPA)^5$). It is not the case for the truncation and hybrid strategies, which significantly accelerate the HeteSim computation and have a close speedup ratio on most conditions. Except the APA path, the MC strategy has the highest speedup ratio among all four strategies on most conditions. Then, let's observe their accuracy from Figs. 5e, 5f, 5g, and 5h. The accuracy of the DP strategy is always close to 1. The hybrid strategy achieves the second performances for most paths. The accuracy of the MC strategy is also high for most paths, while it fluctuates on different paths. Obviously, the truncation strategy has the lowest accuracy on most conditions.

As we have noted, the DP, as an information-lossless strategy, only speeds up the MUL phase which is not the main time-consuming part for most paths. So the DP strategy trivially accelerates HeteSim with the accuracy close to 1. The truncation strategy is an information-loss strategy to keep matrix sparse, so it can effectively accelerate HeteSim. That is the reason why the truncation strategy has the high speedup ratio but low accuracy. Because the hybrid strategy combines the benefits of DP and truncation strategy, it has a close speedup ratio to the truncation strategy with higher accuracy. In order to achieve high accuracy, more walkers in the MC strategy are needed for high-dimension or sparse matrix, while the fixed walkers in experiments (i.e., K is 500) makes the MC strategy the poor accuracy on some conditions.

According to the analysis above, these strategies are suitable for different paths and scenarios. For very sparse matrix (e.g., $(APA)^l$) and low-dimension matrix (e.g., $(APCPA)^3$), all strategies cannot significantly improve efficiency. However, in these conditions, the HeteSim can be quickly computed without any strategies. For those dense (e.g., $(APCPA)^4$) and high-dimension matrix (e.g., $(APSPA)^4$) which have huge computation overhead, the truncation, hybrid, and MC strategies can effectively improve the HeteSim's efficiency. Particularly, the speedup of the hybrid and MC strategies are up to 100 with little loss in accuracy. If the MUL phase is the main time-consuming part for a path, the DP strategy can also speed up HeteSim greatly without loss in accuracy. The MC strategy has very high efficiency, but its accuracy may degrade for high-dimension matrix. So the appropriate K needs to be set through balancing the efficiency and effectiveness.

7 Conclusion

In this paper, we study the relevance search problem which measures the relatedness of heterogeneous objects in heterogeneous networks. We propose a general relevance measure, called HeteSim. As a path-constraint measure, HeteSim can measure the relatedness of same-typed or different-typed objects in a uniform framework. In addition, HeteSim is a semi-metric measure, which can be used in many applications. Extensive experiments validate the effectiveness and efficiency of HeteSim on evaluating the relatedness of heterogeneous objects.

There are some interesting directions for future work. First, more methods can be explored to measure the relatedness of heterogeneous objects, such as path count and RW strategies. Secondly, since the proposed quick computation strategies are all in-memory methods, we can design the parallel computation methods of HeteSim. Last but not least, the problem on how to choose and weight different meta paths is also important issues for heterogeneous networks.

Acknowledgments

This work was supported by the National Basic Research Program of China (2013CB329603), the National Natural Science Foundation of China (No. 61375058, 61074128, 71231002, 60905025) and the Fundamental Research Funds for the Central Universities. This work was also supported in part by US National Science Foundation (NSF) through grants CNS-1115234, DBI-0960443, and OISE-1129076, and US Department of Army through grant W911NF-12-1-0066.

References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web," technical report, Stanford Univ. Database Group, 1998.

[2] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl, "Grouplens: Applying Collaborative Filtering to Usenet News," *Comm. ACM*, vol. 40, no. 3, pp. 77-87, 1997.

[3] G. Jeh and J. Widom, "Scaling Personalized Web Search," *Proc. 12th Int'l Conf. World Wide Web (WWW)*, pp. 271-279, 2003.

[4] G. Jeh and J. Widom, "Simrank: A Measure of Structural-Context Similarity," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 538-543, 2002.

[5] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu, "Pathsim: Meta Path-Based Top-k Similarity Search in Heterogeneous Information Networks," *Proc. Int'l Conf. Very Large Databases (VLDB)*, pp. 992-1003, 2011.

[6] M. Jamali and L. Lakshmanan, "Heteromf: Recommendation in Heterogeneous Information Networks Using Context Dependent Factor Models," *Proc. 22nd Int'l Conf. World Wide Web (WWW)*, pp. 643-654, 2013.

[7] G. Palma, M.E. Vidal, E. Haag, L. Raschid, and A. Thor, "Measuring Relatedness between Scientific Entities in Annotation Datasets," *Proc. Int'l Conf. Bioinformatics, Computational Biology and Biomedical Informatics (BCB)*, pp. 367-376, 2013.

[8] Y. Sun, Y. Yu, and J. Han, "Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 797-806, 2009.

[9] J. Han, "Mining Heterogeneous Information Networks by Exploring the Power of Links," *Proc. 12th Int'l Conf. Discovery Science*, pp. 13-30, 2009.

[10] J. Zhu, A.P. de Vries, G. Demartini, and T. Iofciu, "Evaluating Relation Retrieval for Entities and Experts," *Proc. ACM SIGIR Workshop*, 2008.

[11] 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.

[12] M. Kolahdouzan and C. Shahabi, "Voronoi-Based K Nearest Neighbor Search for Spatial Network Databases," *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB)*, pp. 840-851, 2004.

[13] C. Xiao, W. Wang, X. Lin, and H. Shang, "Top-k Set Similarity Joins," *Proc. IEE 25th Int'l Conf. Data Eng. (ICDE)*, pp. 916-927, 2009.

[14] H. Tong, C. Faloutsos, and J. Pan, "Fast Random Walk with Restart and Its Applications," *Proc. Sixth Int'l Conf. Data Mining (ICDM)*, pp. 613-622, 2006.

[15] M. Gupta, A. Pathak, and S. Chakrabarti, "Fast Algorithms for Top-k Personalized Pagerank Queries," *Proc. 17th Int'l Conf. World Wide Web (WWW)*, pp. 1225-1226, 2008.

[16] X. Xu, N. Yuruk, Z. Feng, and T.A.J. Schweiger, "Scan: An Structural Clustering Algorithm for Networks," *Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 824-833, 2007.

[17] R. Jin, V.E. Lee, and H. Hong, "Axiomatic Ranking of Network Role Similarity," *Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 922-930, 2011.

[18] A. Balmin, V. Hristidis, and Y. Papakonstantinou, "Objectrank: Authority-Based Keyword Search in Databases," *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB)*, pp. 564-575, 2004.

[19] Z. Nie, Y. Zhang, J. Wen, and W. Ma, "Object-Level Ranking: Bringing Order to Web Objects," *Proc. 14th Int'l Conf. World Wide Web (WWW)*, pp. 422-433, 2005.

[20] 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," *Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, pp. 655-664, 2011.

[21] F. Fouss, A. Pirotte, J.M. Renders, and M. Saerens, "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," *IEEE Trans. Knowledge and Data Eng.*, vol. 19, no. 3, pp. 355-369, Mar. 2007.

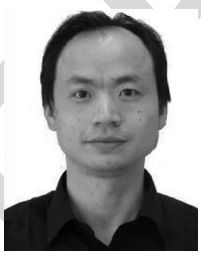
[22] N. Lao and W. Cohen, "Fast Query Execution for Retrieval Models Based on Path Constrained Random Walks," *Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 881-888, 2010.
[23] Q. Xia, "The Geodesic Problem in Quasimetric Spaces," *J. Geometric Analysis*, vol. 19, no. 2, pp. 452-479, 2009.
[24] 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," *Proc. 18th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 1348-1356, 2012.
[25] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao, "Graph Regularized Transductive Classification on Heterogeneous Information Networks," *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pp. 570-586, 2010.
[26] C. Shi, C. Zhou, X. Kong, P.S. Yu, G. Liu, and B. Wang, "HeteRecom: A Semantic-Based Recommendation System in Heterogeneous Networks," *Proc. 18th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 1552-1555, 2012.
[27] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, Aug. 2000.
[28] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis," *Proc. 12th Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT)*, pp. 565-576, 2009.
[29] D. Fogaras, B. Rácz, K. Csalogány, and T. Sarlós, "Towards Scaling Fully Personalized Pagerank: Algorithms, Lower Bounds, and Experiments," *Internet Math.*, vol. 2, no. 3, pp. 333-358, 2005.

Chuan Shi received the BS degree from the Jilin University in 2001, the MS degree from the Wuhan University in 2004, and the PhD degree from the ICT of Chinese Academic of Sciences in 2007. He joined the Beijing University of Posts and Telecommunications as a lecturer in 2007, and is currently an associate professor. His research interests include machine learning, data mining, evolutionary computing. He has published more than 30 papers in refereed journals and conferences. He is a member of the IEEE.

Xiangnan Kong received the bachelor's and master's degrees in computer science from Nanjing University, China, in 2006 and 2009, respectively. He is currently working toward the PhD degree in the Department of Computer Science, University of Illinois at Chicago. He has been working in data mining and machine learning, particularly in graph mining, multi-label learning, and semi-supervised learning.

Yue Huang received the BS degree from the Beijing University of Posts and Telecommunications in 2013. He is currently working toward the master's degree at Columbia University. His research interests include machine learning and data mining.

Philip S. Yu received the BS degree in EE from National Taiwan University, the MS and PhD degrees in EE from Stanford University, and the MBA degree from New York University. He is a distinguished professor in computer science at the University of Illinois at Chicago and also holds the Wexler chair in Information Technology. He spent most of his career at IBM, where he was a manager of the Software Tools and Techniques group at the Watson Research Center. His research interest is on big data, including data mining, data stream, database and privacy. He has published more than 790 papers in refereed journals and conferences. He holds or has applied for more than 250 US patents. He is the editor-in-chief of *ACM Transactions on Knowledge Discovery from Data*. He is on the steering committee of the IEEE Conference on Data Mining and ACM Conference on Information and Knowledge Management and was a member of the IEEE Data Engineering steering committee. He was the editor-in-chief of *IEEE Transactions on Knowledge and Data Engineering* (2001-2004). He is a fellow of the ACM and the IEEE.



Bin Wu received the BS degree from the Beijing University of Posts and Telecommunications in 1991, and the MS and PhD degrees from the ICT of Chinese Academic of Sciences in 1998 and 2002, respectively. He joined the Beijing University of Posts and Telecommunications as a lecturer in 2002, and is a professor at present. His research interests include data mining, complex network, and cloud computing. He has published more than 100 papers in refereed journals and conferences. He is a member of the IEEE.