Distant Meta-Path Similarities for Text-Based Heterogeneous Information Networks

Chenguang Wang
IBM Research-Almaden
chenguang.wang@ibm.com

Yangqiu Song
Department of CSE, Hong Kong University of Science and Technology
yqsong@cse.ust.hk

Haoran Li
School of EECS, Peking University
lihaoran_2012@pku.edu.cn

Yizhou Sun
Department of Computer Science, University of California, Los Angeles
yzsun@cs.ucla.edu

Ming Zhang
School of EECS, Peking University
mzhang_cs@pku.edu.cn

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

ABSTRACT

Measuring network similarity is a fundamental data mining problem. The mainstream similarity measures mainly leverage the structural information regarding to the entities in the network without considering the network semantics. In the real world, the heterogeneous information networks (HINs) with rich semantics are ubiquitous. However, the existing network similarity doesn't generalize well in HINs because they fail to capture the HIN semantics. The meta-path has been proposed and demonstrated as a right way to represent semantics in HINs. Therefore, original meta-path based similarities (e.g., PathSim and KnowSim) have been successful in computing the entity proximity in HINs. The intuition is that the more instances of meta-path(s) between entities, the more similar the entities are. Thus the original meta-path similarity only applies to computing the proximity of two neighborhood (connected) entities. In this paper, we propose the *distant meta-path similarity* that is able to capture HIN semantics between two distant (isolated) entities to provide more meaningful entity proximity. The main idea is that even there is no shared neighborhood entities of (i.e., no meta-path instances connecting) the two entities, but if the more similar neighborhood entities of the entities are, the more similar the two entities should be. We then find out the optimum distant meta-path similarity by exploring the similarity hypothesis space based on different theoretical foundations. We show the state-of-the-art similarity performance of distant meta-path similarity on two text-based HINs and make the datasets public available.[1]

1 INTRODUCTION

Measuring network similarity (i.e., entity proximity in networks) is a fundamental problem of data mining with successful applications in information retrieval, similarity search and machine learning algorithms. Under a common assumption that the "entities are without types" (i.e., the networks don't carry semantics), most of the state-of-the-art network similarity methods only leverage the structural information of the entities to compute the proximity in homogeneous networks. However, heterogeneous information networks (HINs) [9] (e.g., social networks and biological networks) with rich semantics are ubiquitous in the real world. The traditional network similarity has been shown not generalized well in HINs.

The reason is that when measuring HIN similarity, besides considering the structural information of the entities, the entity proximity should be able to capture the HIN semantics. The meta-path represents semantics in HINs. The meta-path based similarity, e.g., PathSim [28] and KnowSim [34], naturally incorporates the HIN semantics and becomes very useful in computing the entity proximity. For example, the text-based HIN has recently been proposed to represent the texts in an HIN [33], where texts are regarded as one type of entities. By applying meta-path similarity to the text-based HIN, we observe significant improvements in text similarity computation, as well as in text clustering [33] and classification [35].

For instance, to compute the document proximity in a text-based HIN, for two documents talking about politics, a meaningful meta-path as the following defined over entity types may be very useful:
Document→Military→Government→Religion→Document.[2]
Whereas, for the two documents talking about sports, a path instance following the meta-path below may be more meaningful:
Document→Baseball→Olympics→Baseball→Document.
Given the meta-path(s), most original meta-path similarities are derived following the intuition: the more instances of meta-path(s) between entities, the more similar the entities are. We can see that original meta-path similarities can only compute the proximity of two neighborhood (connected) entities, between which path instances follow the meaningful meta-path(s) must exist.

We expect a similarity measure that can be generalized to compute the proximity of two distant (isolated) entities, between which the meaningful meta-path instances don't exist. This is of great need in most of the real world HINs. For example, given a meta-path *Document→Athlete→Document*, and a pair of documents D_0 and

[1] https://github.com/cgraywang/TextHINData

[2] Different from original meta-path similarities that assume the meta-path(s) are symmetric, we allow any meta-path(s) to be used in the similarity measures in this study.

CIKM'17 , November 6–10, 2017, Singapore,Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133029

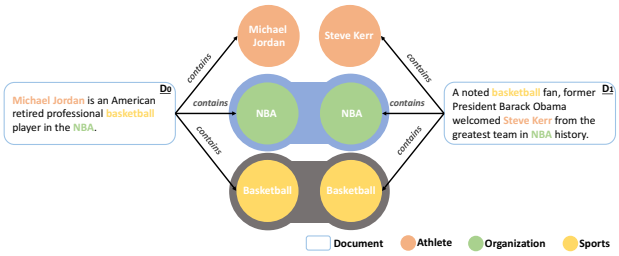


Figure 1: A text-based heterogeneous information network example.

D_1 in a text-based HIN as shown in Figure 1, we want to compute the proximity of the documents. Since there doesn't exist any path instance of the meta-path directly connecting D_0 and D_1 (*Michael Jordan* and *Steve Kerr* are not the same entity), the document proximity computed based on an original meta-path similarity will be zero. However the two documents are talking about sports and should be similar. This indicates that the original meta-path similarity cannot capture the HIN semantics between the distant entities.

In this paper, we propose the *distant meta-path similarity* to fully capture the semantics between entities in an HIN to provide more meaningful proximity of two distant entities. Intuitively, the distant meta-path similarity aims to bridge the gap between distant (relatively isolated) entities whenever there doesn't exist any meaningful path instance of a meta-path in between but originally carry similar semantics. Formally, distant meta-path similarity indicates the proximity of the two entities' neighborhood entities. The more similar neighborhood entities with the same type of the two entities are, the more similar the two entities are. The neighborhood entities of an entity refer to the entities linked via direct meta-paths to that entity. Then the two entities become distant neighbors to each other. The semantics regarding to the relationship(s) of the entity pair are thus better preserved by the distant neighbors.

Following the example in Figure 1, besides the meta-path *Document*⟶*Athlete*⟶*Document*, D_0 and D_1 have two shared neighborhood entities *NBA* and *Basketball* connected by meta-paths *Document*⟶*Organization* and *Document*⟶*Sports* respectively. Then for example, a simple similarity measure can be developed by considering the intersection in the neighborhood entity set of the two documents. The proximity of D_0 and D_1 is thus 2/3 (i.e., among six neighborhood entities, four of them are the same) and more meaningful. For the sake of distinction, we regard the original meta-path similarity as the neighborhood meta-path similarity. Compared to the neighborhood meta-path similarity, in distant meta-path similarity, even if two entities are isolated following the given meta-path(s), the more similar or the same neighborhood entities of the two entities are, the more similar the two entities should be.

To find out the optimum distant meta-path similarity, we explore similarity hypothesis space by deriving 53 different similarity measures falling into nine families. Given different assumptions, such as set assumption or probabilistic assumption, we can represent the meta-path similarities of one entity to all the other entities as different feature vectors. Then we can apply set theory or information theory to derive the corresponding similarities between two feature vectors.

To evaluate the proposed distant meta-path similarities, we develop two text-based HINs (i.e., 20NG-HIN and GCAT-HIN) based on benchmark text datasets, i.e., 20Newsgroups [18] and RCV1 [19]. We construct 20NG-HIN and GCAT-HIN by using the unsupervised semantic parsing framework [33] to ground the texts to world knowledge base, Freebase. The resultant 20NG-HIN and GCAT-HIN consist of 20 and 43 entity types, as well as 325 and 1,682 meta-paths respectively. To our best knowledge, these two datasets are the annotated HIN datasets with the largest numbers of entity types and meta-paths. Then based on the two datasets, we conduct comprehensive experiments on text clustering and classification tasks. For clustering, we employ spectral clustering algorithm [21] that can use similarities as the weights on the graph edges constructed by the data points. For classification, we use support vector machine (SVM) to incorporate the similarities into kernels [35]. Then we compare different similarity measures within intra-family and inter-families. We conclude with the best family of distant meta-path similarities as well as the correlation among families for the two datasets. We non-surprisingly find that the optimum distant meta-path similarity can be significantly better than the original neighborhood meta-path similarities (around 20% gain in clustering NMI and 20% gain in classification accuracy).

The contributions of this work can be summarized as follows.

- We define distant meta-path similarity to fully capture HIN semantics in disconnected entity proximity computation.
- We explore the similarity hypothesis space by proposing 53 newly derived distant meta-path similarities based on different theoretical foundations.
- We present the optimum meta-path similarity by conducting comprehensive experiments on two text-based HIN benchmark datasets, and show the state-of-the-art performance of the best distant meta-path similarity.
- We make the two text-based HIN datasets public available.

The rest of the paper is organized as follows. Section 2 introduces some basic concepts of the HIN. Section 3 briefly revisits neighborhood meta-path similarity and mainly presents distant meta-path similarity on the HIN. Experiments and results are discussed in Section 4. We conclude this study in Section 5.

2 PRELIMINARIES

In this section, we introduce the key related concepts of HIN. We first define the HIN and its network schema.

Definition 2.1. A **heterogeneous information network** (HIN) is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with an entity type mapping $\phi: \mathcal{V} \rightarrow \mathcal{A}$ and a relation type mapping $\psi: \mathcal{E} \rightarrow \mathcal{R}$, where $\mathcal{V}$ denotes the entity set, $\mathcal{E}$ denotes the link set, $\mathcal{A}$ denotes the entity type set, and $\mathcal{R}$ denotes the relation type set, and the number of entity types $|\mathcal{A}| > 1$ or the number of relation types $|\mathcal{R}| > 1$.

Definition 2.2. Given an HIN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the entity type mapping $\phi: \mathcal{V} \rightarrow \mathcal{A}$ and the relation type mapping $\psi: \mathcal{E} \rightarrow \mathcal{R}$, the **network schema** for network G, denoted as $\mathcal{T}_{\mathcal{G}} = (\mathcal{A}, \mathcal{R})$, is a graph with nodes as entity types from $\mathcal{A}$ and edges as relation types from $\mathcal{R}$.

The network schema provides a high-level description of a given heterogeneous information network. It defines the topology of

the entity type relationships. Another important concept, meta-path [28], is proposed to systematically define relations between entities at the schema level.

Definition 2.3. A **meta-path** $\mathcal{P}$ is a path defined on the graph of network schema $\mathcal{T}_G = (\mathcal{A}, \mathcal{R})$, 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 $R = R_1 \cdot R_2 \cdot \ldots \cdot R_L$ between types A_1 and A_{L+1}, where $\cdot$ denotes relation composition operator, and L is the length of $\mathcal{P}$.

We say a meta-path is *symmetric* if the relation R is symmetric. For simplicity, we use type names connected by "–" to denote the meta-path when there exist no multiple relations between a pair of types: $\mathcal{P} = (A_1 - A_2 - \ldots - A_{L+1})$. For example, in the Freebase network, the composite relation *two Person co-founded an Organization* can be described as *Person* $\xrightarrow{\text{found}}$ *Organization* $\xrightarrow{\text{found}^{-1}}$ *Person*, or *Person-Organization-Person* for simplicity. We say a path $p = (v_1 - v_2 - \ldots - v_{L+1})$ between v_1 and v_{L+1} in network $\mathcal{G}$ follows the meta-path $\mathcal{P}$, if $\forall l, \phi(v_l) = A_l$ and each edge $e_l = \langle v_l, v_{l+1} \rangle$ belongs to each relation type R_l in $\mathcal{P}$. We call these paths as *path instances* of $\mathcal{P}$, denoted as $p \in \mathcal{P}$. R_l^{-1} represents the reverse order of relation R_l.

The commuting matrix is defined by Y. Sun et al. [28] to compute the frequencies of all the paths related to a meta-path.

Definition 2.4. **Commuting matrix.** Given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and its network schema $\mathcal{T}_\mathcal{G}$, a commuting matrix $\mathbf{M}_\mathcal{P}$ for a meta-path $\mathcal{P} = (A_1 - A_2 - \ldots - A_{L+1})$ is defined as $\mathbf{M}_\mathcal{P} = \mathbf{W}_{A_1A_2}\mathbf{W}_{A_2A_3} \ldots \mathbf{W}_{A_LA_{L+1}}$, where $\mathbf{W}_{A_iA_j}$ is the adjacency matrix between types A_i and A_j. $\mathbf{M}_\mathcal{P}(i, j)$ represents the number of path instances between objects x_i and y_j, where $\phi(x_i) = A_1$ and $\phi(y_j) = A_{L+1}$, under meta-path $\mathcal{P}$.

We introduce two meta-path based similarities as below.

Definition 2.5. **PathSim** [28] : A meta-path based similarity measure. Given a symmetric meta-path $\mathcal{P}$, PathSim between two entities i and j of the same entity type is:

$$\begin{aligned} \text{PathSim}(i, j) &= \frac{2 \times |\{p_{i \rightsquigarrow j} \in \mathcal{P}\}|}{|\{p_{i \rightsquigarrow i} \in \mathcal{P}\}| + |\{p_{j \rightsquigarrow j} \in \mathcal{P}\}|} \\ &= \frac{2 \cdot \mathbf{M}_\mathcal{P}(i, j)}{\mathbf{M}_\mathcal{P}(i, i) + \mathbf{M}_\mathcal{P}(j, j)}, \end{aligned} \tag{1}$$

where $p_{i \rightsquigarrow j} \in \mathcal{P}$ is a path instance between i and j following meta-path $\mathcal{P}$, $p_{i \rightsquigarrow i} \in \mathcal{P}$ is that between i and i, and $p_{j \rightsquigarrow j} \in \mathcal{P}$ is that between j and j. Based on Definition 2.4, we have $|\{p_{i \rightsquigarrow j} \in \mathcal{P}\}| = \mathbf{M}_\mathcal{P}(i, j)$, $|\{p_{i \rightsquigarrow i} \in \mathcal{P}\}| = \mathbf{M}_\mathcal{P}(i, i)$, and $|\{p_{j \rightsquigarrow j} \in \mathcal{P}\}| = \mathbf{M}_\mathcal{P}(j, j)$.

Definition 2.6. **KnowSim** [35]: Given a collection of symmetric meta-paths, denoted as $\mathbf{P} = \{\mathcal{P}_m\}_{m=1}^M$, KnowSim between two entities i and j is defined as:

$$\begin{aligned} &\text{KnowSim}_{\boldsymbol{\omega}}(i, j) \\ &= \frac{2 \times \sum_m^M \omega_m |\{p_{i \rightsquigarrow j} \in \mathcal{P}_m\}|}{\sum_m^M \omega_m |\{p_{i \rightsquigarrow i} \in \mathcal{P}_m\}| + \sum_m^M \omega_m |\{p_{j \rightsquigarrow j} \in \mathcal{P}_m\}|} \\ &= \frac{2 \cdot \sum_m^M \omega_m \mathbf{M}_{\mathcal{P}_m}(i, j)}{\sum_m^M \omega_m \mathbf{M}_{\mathcal{P}_m}(i, i) + \sum_m^M \omega_m \mathbf{M}_{\mathcal{P}_m}(j, j)}. \end{aligned} \tag{2}$$

We use a vector $\boldsymbol{\omega} = [\omega_1, \cdots, \omega_m, \cdots, \omega_M]$ to denote the meta-path weights, where ω_m is the weight of meta-path $\mathcal{P}_m$.

3 HIN SIMILARITIES

In this section, we first revisit neighborhood meta-path similarities and then define distant meta-path similarities.

3.1 Neighborhood Meta-Path Similarity

We formalize the original meta-path similarity as the neighborhood meta-path similarity as defined in Def. 3.1.

Definition 3.1. **Neighborhood meta-path similarity.** The neighborhood meta-path similarity indicates the pairwise proximity between entities linked by meta-path(s) (entities are neighborhood entities to each other). Given two entities i and j connected by a meta-path $\mathcal{P}$, if $\mathbf{M}_\mathcal{P}(i, j) > 0$, the neighborhood meta-path similarity is positive. Otherwise, the neighborhood meta-path similarity between i and j is 0.

The intuition is that the more instances of meta-path(s) between entities, the more similar the entities are. Both the PathSim and KnowSim are neighborhood meta-path similarities. Let's take the following two scenarios for example to see how neighborhood meta-path similarity works. Particularly, we use the PathSim and KnowSim as examples of neighborhood meta-path similarities. First, we consider a simple document-word HIN schema consisting of two types of entities, *Document* and *Word*. Then if the meta-path is *Document*→*Word*→*Document*, the PathSim between documents i and j can be interpreted as:

$$\text{PathSim} = \frac{2 \cdot \#\text{shared words in two documents}}{\#\text{words in document } i + \#\text{words in document } j}, \tag{3}$$

when only the appearance of words instead of frequency of words in a document is considered. In the document-word HIN, this measures the semantic similarity between documents by considering the one-hop meta-path based neighborhood entities.

Second, if we consider all the possible meta-paths in a network, then the unweighted version of KnowSim degenerates to the following formulation:

$$\text{KnowSim}^\infty = \frac{2 \cdot \#\text{paths between two entities}}{\#\text{cirles with entity } i + \#\text{circles with entity } j}, \tag{4}$$

which computes the document proximity based on the two-hop meta-path based neighborhood entities.

Both above similarities are able to capture the HIN semantics but do not consider the meta-path connection of the two entities with other entities, i.e., only consider the direct multi-hop meta-path based neighborhood entities. Thus, we call both PathSim and KnowSim as neighborhood meta-path similarities, meaning that they consider the neighborhood entities connected by a meta-path.

As we can see, the neighborhood meta-path similarity works pretty well when calculating the proximity of two neighborhood entities, between which the path instances of the meaningful meta-path(s) exist. However, in the real world HINs, such as the example in Figure 1, when the two entities are distant entities (between which no path instances of a meta-path exist), the neighborhood meta-path similarities cannot be applied to compute the proximity of the entities (because the similarity will always equal to zero). This indicates the neighborhood meta-path similarity fails to capture HIN semantics between distant entities.

Table 1: Representative meta-path based similarity measures from ten families. The first two similarities are neighborhood meta-path similarities, the others are distant meta-path similarities.

Family	Similarity	Formulation (all are the similarities for entities i and j of the same type)
Neighborhood similarity	KnowSim	$S_{KnowSim} = \frac{2 \cdot \sum_m^M \mathbf{M}_{\mathcal{P}_m}(i,j)}{\sum_m^M \mathbf{M}_{\mathcal{P}_m}(i,i) + \sum_m^M \mathbf{M}_{\mathcal{P}_m}(j,j)}$ (5)
Neighborhood similarity	Avg(PathSim)	$S_{Avg(PathSim)} = \frac{1}{M}\sum_m^M \frac{2 \cdot \mathbf{M}_{\mathcal{P}_m}(i,j)}{\mathbf{M}_{\mathcal{P}_m}(i,i) + \mathbf{M}_{\mathcal{P}_m}(j,j)}$ (6)
Intersection	Hamming	$S_{Ham} = \frac{MN}{\sum_{m=1}^M \sum_{k=1}^N \lvert\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)\rvert}$ (7)
Inner product	Cosine	$S_{Cos} = \frac{\sum_{m=1}^M \sum_{k=1}^N \mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(j,k)}{\sqrt{\sum_{m=1}^M \sum_{k=1}^N \mathbf{M}_{\mathcal{P}_m}(i,k)^2}\sqrt{\sum_{m=1}^M \sum_{k=1}^N \mathbf{M}_{\mathcal{P}_m}(j,k)^2}}$ (8)
L_p Minkowski	Euclidean L_2	$S_{Euc} = \frac{1}{\sqrt{\sum_{m=1}^M \sum_{k=1}^N \lvert\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)\rvert^2}}$ (9)
L_1	SØrensen	$S_{S\varnothing r} = 1 - \frac{\sum_{m=1}^M \sum_{k=1}^N \lvert\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)\rvert}{\sum_{m=1}^M \sum_{k=1}^N (\mathbf{M}_{\mathcal{P}_m}(i,k) + \mathbf{M}_{\mathcal{P}_m}(j,k))}$ (10)
Squared L_2	Clark	$S_{Cla} = \frac{1}{\sqrt{\sum_{m=1}^M \sum_{k=1}^N (\frac{\lvert\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)\rvert}{\mathbf{M}_{\mathcal{P}_m}(i,k) + \mathbf{M}_{\mathcal{P}_m}(j,k)})^2}}$ (11)
Binary	Russell-Rao	$S_{Rus} = 1 - \frac{MN - \sum_{m=1}^M \sum_{k=1}^N \mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(j,k)}{MN}$ (12)
Fidelity	Hellinger	$S_{Hel} = 2 \times (1 - \sqrt{1 - \sum_{m=1}^M \sum_{k=1}^N \sqrt{\mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(j,k)}})$ (13)
Shannon's entropy	Kullback-Leibler	$S_{KL} = \frac{1}{\sum_{m=1}^M \sum_{k=1}^N \mathbf{M}_{\mathcal{P}_m}(i,k) \ln \frac{\mathbf{M}_{\mathcal{P}_m}(i,k)}{\mathbf{M}_{\mathcal{P}_m}(j,k)}}$ (14)
Hybrids	Avg(L_1, L_∞)	$S_{Avg} = \frac{2}{\sum_{m=1}^M \sum_{k=1}^N \lvert\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)\rvert + \max_{j,k} \lvert\mathbf{M}_{\mathcal{P}_m}(i,k) - \mathbf{M}_{\mathcal{P}_m}(j,k)\rvert}$ (15)

3.2 Distant Meta-Path Similarity

We then propose distant meta-path similarity aiming to fully capture the HIN semantics to provide more meaningful proximity of distant (relatively isolated) entities in HINs.

For instance, in Figure 1, given two documents (D_0 and D_1) and the meta-path *Document→Athlete→Document*, the proximity of the two documents is zero based on the neighborhood meta-path similarity. However both documents are talking about sports thus should be relatively similar (similarity score greater than zero). We expect to use distant meta-path similarity to bridge the gap between such distant entities and provide entity proximity with more accurate semantics. Formally, the distant meta-path similarity is defined in Def. 3.2.

Definition 3.2. **Distant meta-path similarity.** The distant meta-path similarity between an entity pair describes the proximity of the pair's neighborhood entities. Neighborhood entities are defined as the entities linked via meta-path(s) to the pair. Let $\{\mathbf{M}_{\mathcal{P}}(i,k)\}_{k=1}^N$ denotes the meta-path instances between entity i and its neighborhood entities. The distant meta-path similarity between i and j is then decided by the proximity of $\{\mathbf{M}_{\mathcal{P}}(i,k)\}_{k=1}^N$ and $\{\mathbf{M}_{\mathcal{P}}(j,k)\}_{k=1}^N$. Entities i and j are called as distant neighbors to each other.

The intuition is that the more similar neighborhood entities with the same type of two entities are, the more similar the two entities are. For example, to consider all the neighborhood entities of the same type, one simplest way is to use the intersection between the two sets of neighborhood entities to compute the similarity:

$$S_{It}^{\mathcal{P}}(i,j) = \sum_{k=1}^N \min(\mathbf{M}_{\mathcal{P}}(i,k), \mathbf{M}_{\mathcal{P}}(j,k)), \tag{16}$$

which considers the intersection of all meta-paths between either entities i or j with neighborhood entities. Note that here to implement intersection, we need to have $\mathbf{M}_{\mathcal{P}}(i,k) \leftarrow I[\mathbf{M}_{\mathcal{P}}(i,k) > 0]$ where $I[true] = 1$ and $I[false] = 0$ are indicator functions. If we consider only all meta-paths, then the similarity is:

$$S_{It}(i,j) = \sum_{m=1}^M \sum_{k=1}^N \min(\mathbf{M}_{\mathcal{P}_m}(i,k), \mathbf{M}_{\mathcal{P}_m}(j,k)). \tag{17}$$

Let's revisit the two scenarios in Sec. 3.1 to see how distant meta-path similarity uses more HIN semantics. We first use *Document→Word →Document* meta-path. Then in Eq. (16), we have

$$\begin{aligned} S_{It}^{\mathcal{P}}(i,j) = \quad & \sum \min(I[\#\text{shared words in document } i \text{ and } k], \\ & I[\#\text{shared words in document } j \text{ and } k]). \end{aligned} \tag{18}$$

This means that, when there are more documents that are "similar" (or sharing words) to both documents i and j, the two documents are more similar. Interestingly, because there is no meta-path connecting two documents, the original network structure does not

support the neighborhood document proximity. We now can compute the distant document proximity which preserves right HIN semantics between the two documents.

For the second case, the distant similarity is approximately:

$$S_{It}^{\infty} = \#\text{paths between two entities bridged by other entities.} \tag{19}$$

Again this similarity provides more accurate semantic proximity that neighborhood meta-path similarity cannot provide.

From the above examples we can see that, distant meta-path similarity captures the HIN semantics (i.e., distant semantics) between disconnected entities that neighborhood meta-path similarities cannot capture. This leads to distant meta-path similarity that provides more meaningful entity proximity in HINs. Then the remaining problem is that *which is the best way to define a distant meta-path similarity?* In the rest of this section, we explore the similarity hypothesis space based on different theoretical foundations, and derive 53 different similarities categorized into nine families. In the interests of space, we only mention the original names of different similarities/distances, cite them, and show one example that is customized to the meta-path based similarity in each family. We will explain the meaning of that similarity as a representative of the family. Note that even in the same family, the semantic meaning of the similarity can be different. A summary of the families and example similarities is shown in Table 1.

3.2.1 Intersection Family. The first family is intersection family which involves the intersection operator inside the similarity. We list the similarities we have implemented as following: **1.** Intersection [6]. **2.** Wave Hedges [10]. **3.** Czekanowski Coefficient [4]. **4.** Motyka similarity [4]: half of Czekanowski Coefficient. **5.** Ruzicka similarity [4]. **6.** $1 - S_{Ruzicka}$ is known as Tanimoto distance [6], a.k.a., Jaccard distance. **7.** Hamming distance [4] based similarity. We have shown that using S_{It} can achieve new semantic meaning of the similarity. For all the similarities in this family, we set $\mathbf{M}_{\mathcal{P}}(i,k) \leftarrow I[\mathbf{M}_{\mathcal{P}}(i,k) > 0]$. The Hamming distance based similarity is shown as Eq. (7) in Table 1. The distance is defined as the number of entities with different meta-paths corresponding to the two entities i and j. Then the similarity is referred to as the inverse number of the distance. For each meta-path, this similarity is related to intersection since larger intersection of entities means lower number of Hamming distance.

3.2.2 Inner Product Family. The inner product family involves the inner product value for each meta-path $\mathcal{P}_m$: $\sum_k \mathbf{M}_{\mathcal{P}_m}(i,k)\mathbf{M}_{\mathcal{P}_m}(i,k)$. We have the following variants: **8.** Simple inner product [6]. **9.** Harmonic mean [4]. **10.** Cosine coefficient (a.k.a., Ochiai [4, 23] and Carbo [23]). **11.** Kumar and Hassebrook based similarity measuring the Peak-to-correlation energy [16]. **12.** Jaccard coefficient (a.k.a. Tanimoto) [30]. **13.** Dice Coefficient or SØrensen, Czekannowski, Hodgkin-Richards [23] or Morisita [24]. **14.** Correlation (Pearson) [4]. The example of cosine meta-path similarity is shown as Eq. (8) in Table 1. Inner product is similar to intersection but also considers the weights of each meta-path value. Cosine similarity normalizes the weights by each of the entity i and j's values.

3.2.3 L_p Minkowski Family. The L_p Minkowski family is a general formulation of p-norm based distance. We derive the following similarities: **15.** L_2 Euclidean distance based similarity. **16.** L_1 City block distance [13] based similarity (rectilinear distance, taxicab norm, and Manhattan distance, proposed by Hermann Minkowski). **17.** L_p Minkowski distance based similarity [3]. **18.** L_∞ Chebyshev distance (chessboard distance and the minimax approximation) [32] based similarity where p goes to infinite. We show the L_2 Euclidean distance based similarity as Eq. (9) in Table 1. It simply computes the Euclidean distance between two vectors comprised by meta-path values from entities i and j to all the other entities, and then uses the inverse value as the similarity. This distance is similar to Hamming distance, but treats the values of each meta-path independently. Moreover, for arbitrary L_p norm, it computes the distances by making different geometric assumptions of the vectors in the high-dimensional space.

3.2.4 L_1 Family. Besides city block distance based similarity, we show more L_1 distance based similarities here: **19.** SØrensen distance [26] (a.k.a., Bray-Curtis [2, 4, 23] based similarity). **20.** Gower distance [8] based similarity. **21.** Soergel [23] distance based similarity. **22.** Kulczynski [4] distance based similarity. **23.** Canberra similarity [4]. **24.** Lorentzian similarity [4]. The differences among these L_1 distance based similarities and the city block distance based similarity introduced previously are the way they weight the distance and the way they convert distance to similarity. For example, for the SØrensen distance based similarity, which is shown in Eq. (10) in Table 1, it uses the sum of all the related meta-path values as denominator to normalize the L_1 distance in the range of $[0, 1]$ and regards "1 - the distance" as the similarity.

3.2.5 Squared L_2 Family. Here we explore more similarities related to the squared value of L_2 norm: **25.** Squared Euclidean distance [4]. **26.** Pearson χ^2 divergence [25]. **27.** Neyman χ^2 [4]. **28.** Squared χ^2 [7] (a.k.a. triangular discrimination [5, 31]). **29.** Probabilistic symmetric χ^2 [4], which is identical to Sangvi χ^2 between populations [4]. **30.** Divergence [14]. **31.** Clark [4]: squared root of half of divergence as defined in the Eq. (11). **32.** Additive Symmetric χ^2 [4, 37]. Squared L_2 family incorporates the squared L_2 norms in the similarity function. For example, squared Euclidean distance is the squared value of Euclidean distance. The difference among the above similarities is how to weight the squared L_2 norm. For example, we show the Clark similarity as Eq. (11) in Table 1. The way to normalize the squared L_2 norm is similar to the way SØrensen distance normalizes the L_1 distance except for the squared value and the way to sum all the values.

3.2.6 Binary Family. We introduce a set of distant meta-path similarities based on binary values instead of scale values. In this case, we set binary values as what we did in intersection. Then the similarities are listed as follows: **33.** Yule similarity [4]. **34.** Matching distance [4]. **35.** Kulsinski is defined as a variation of Yule similarity. **36.** Roger-Tanimoto similarity [4]. **37.** Russel-Rao similarity [4] is formally defined in Eq. (12). **38.** Sokal-Michener's simple matching [4] (a.k.a. Rand similarity). The corresponding metric $1 - S_{Sokal-Michener}$ is called the variance or Manhattan similarity (a.k.a. Penrose size distance). **39.** Sokal-Sneath similarity [4]. Binary similarity is more complicated than intersection since it can introduce a lot of logical operators over the binary values. We choose the simplest one of Russel-Rao similarity shown

Table 2: Statistics of entities in two text-based HINs: #(Document) is the number of all documents; similar for #(Word) (# of distinct words), #(FBEntity) (# of distinct Freebase entities), #(Total) (the total # of distinct entities), and #Types (the total # of entity types).

	20NG-HIN	GCAT-HIN
#(Document)	19,997	60,608
#(Word)	60,691	95,001
#(FBEntity)	28,034	110,344
#(Total)	108,722	265,953
#(Types)	1,904	1,937

as Eq. (12) in Table 1. In Russel-Rao similarity, we use an "AND" operation to generate the similarity.

3.2.7 Fidelity Family. Fidelity family incorporates geometric mean of both meta-path values of entities i and j, and further sum or average the mean values. We summarize the similarities we use here: **40.** Fidelity similarity [4], a.k.a. Bhattacharyya coefficient or Hellinger affinity [4]. **41.** Bhattacharyya distance based similarity [1]. **42.** Hellinger [4]. **43.** Matusita [22]. **44.** Squared-chord distance based similarity [3] is the Matusita but without the square root. A typical fidelity family similarity Hellinger is shown in Eq. (13) in Table 1. Hellinger distance is originally defined with measure theory based on two probability distributions. Therefore, we normalize the frequencies of path instances to probabilities as $\mathbf{M}_{\mathcal{P}}(i,k) \leftarrow \mathbf{M}_{\mathcal{P}}(i,k)/\sum_{k'} \mathbf{M}_{\mathcal{P}}(i,k')$. It can be proven that Hellinger distance is in the range of $[0,1]$ based on the Cauchy-Schwarz inequality. Thus, in our case, we simply use "1 - Hellinger distance" as the similarity.

3.2.8 Shannon's Entropy Family. The Shannon's entropy family is listed as follows: **45.** Kullback and Leibler (KL) [15] divergence (relative entropy or information deviation). **46.** Jeffreys or J divergence [12, 15, 29]. **47.** K divergence based similarity [4]. **48.** K divergence's symmetric form TopsØe distance [4] (a.k.a. information statistics [7]). **49.** Jensen-Shannon divergence [4, 20]. **50.** Jensen difference [29]. Since the entropy is also defined on probabilities, we normalize the frequencies to be probabilities as we did for Hellinger distance, e.g., the KL divergence is shown as Eq. (14) in Table 1. KL divergence is originally used to evaluate the difference between two distributions. We regard the inverse value as the similarity.

3.2.9 Hybrid Family. We include some combinations of the above similarities. **51.** Taneja [11]: arithmetic and geometric means that come up with the arithmetic and geometric mean divergence. **52.** Symmetric χ^2: arithmetic and geometric mean divergence is presented according to [17]. **53.** Avg(L_1, L_∞): average of city block and Chebyshev distances [13] is shown as Eq. (15) in Table 1.

4 EXPERIMENTS

In this section, we report experimental results that demonstrate the effectiveness of distant meta-path similarities compared with neighborhood meta-path similarities. We also analyze the relationships between different similarity families.

Table 3: Results of clustering and classification of different meta-path based similarities on 20NG-HIN and GCAT-HIN datasets. Clust. means clustering and Class. means classification. We use underline to emphasize each best similarity in every family. We use boldface to emphasize the overall best similarity and the best mean value among all the families.

	20NG-HIN		GCAT-HIN	
	Clust.	Class.	Clust.	Class.
KnowSim	0.223	52.4%	0.299	81.6%
Avg(PathSim)	0.218	13.0%	0.329	69.4%
Mean(neighborhood)	**0.221**	32.7%	0.314	75.5%
1. Intersection	0.218	65.1%	0.328	92.1%
2. Wave Hedges	0.057	39.0%	0.159	57.2%
3. Czekanowski	0.119	66.5%	0.229	90.7%
4. Motyka	0.219	66.9%	0.286	85.1%
5. Ruzicka	0.059	42.9%	0.043	46.9%
6. Tanimoto	0.044	28.6%	0.038	41.2%
7. Hamming	0.168	56.9%	0.188	83.6%
Mean(Intersection)	0.126	52.3%	0.182	71.0%
8. Inner Product	0.154	63.4%	0.237	92.7%
9. Harmonic Mean	0.191	62.1%	0.205	89.2%
10. Cosine	0.248	**67.4%**	0.242	**93.1%**
11. Kumar-Hassebrook	0.104	50.5%	0.233	82.4%
12. Jaccard	0.104	50.5%	0.225	82.4%
13. Dice	0.04	25.2%	0.037	56.5%
14. Correlation	0.243	**67.4%**	0.251	**93.1%**
Mean(Inner product)	0.155	55.2%	0.204	84.2%
15. Euclidean L_2	0.254	65.2%	**0.376**	89.7%
16. City block L_1	0.055	57.8%	0.309	77.6%
17. Minkowski L_p	**0.278**	62.8%	0.366	90.9%
18. Chebyshev L_∞	0.192	50.1%	0.288	88.1%
Mean(L_p Minkowski)	0.195	**59.0%**	**0.335**	**86.6%**
19. SØrensen	0.227	66.4%	0.333	91.8%
20. Gower	0.23	49.9%	0.311	89.3%
21. Soergel	0.044	28.6%	0.038	41.2%
22. Kulczynski	0.06	42.9%	0.042	46.9%
23. Canberra	0.197	26.0%	0.209	57.2%
24. Lorentzian	0.063	57.7%	0.31	77.6%
Mean(L_1)	0.137	45.3%	0.207	67.3%
25. Squared Euclidean	0.099	61.2%	0.347	90.8%
26. Pearson χ^2	0.036	41.9%	0.105	73.5%
27. Neyman χ^2	0.041	10.1%	0.099	72.4%
28. Squared χ^2	0.187	30.5%	0.193	57.2%
29. Prob. Symmetric χ^2	0.186	15.7%	0.18	57.2%
30. Divergence	0.19	19.6%	0.273	57.2%
31. Clark	0.201	25.8%	0.238	78.6%
32. Add. Symmetric χ^2	0.236	41.3%	0.248	68.0%
Mean(Squared L_2)	0.147	30.8%	0.21	69.4%
33. Yule	0.036	14.3%	0.039	47.6%
34. Matching	0.257	47.6%	0.344	75.6%
35. Kulsinski	0.171	47.6%	0.189	72.5%
36. Rogers-Tanimoto	0.257	47.6%	0.344	79.7%
37. Russell-Rao	0.156	47.6%	0.238	72.5%
38. Sokal-Michener	0.255	47.6%	0.345	79.7%
39. Sokal-Sneath	0.052	25.2%	0.086	57.1%
Mean(Binary)	0.169	39.6%	0.226	69.2%
40. Fidelity	0.243	65.9%	0.311	92.1%
41. Bhattacharyya	0.161	30.5%	0.285	76.9%
42. Hellinger	0.186	32.4%	0.29	47.1%
43. Matusita	0.191	31.4%	0.292	46.6%
44. Squared-chord	0.257	65.3%	0.321	92.1%
Mean(Fidelity)	0.208	45.1%	0.3	71.0%
45. Kullback-Leibler	0.032	43.8%	0.078	55.6%
46. Jeffreys	0.037	10.2%	0.016	47.4%
47. K divergence	0.043	41.9%	0.03	52.5%
48. TopsØe	0.112	62.8%	0.299	82.6%
49. Jensen-Shannon	0.144	64.8%	0.311	85.7%
50. Jensen difference	0.144	64.8%	0.312	85.8%
Mean(Shannon)	0.085	48.1%	0.174	68.3%
51. Taneja	0.147	62.6%	0.314	81.4%
52. Kumar-Johnson	0.247	47.6%	0.256	58.2%
53. Avg(L_1, L_∞)	0.11	59.5%	0.278	81.2%
Mean(Hybrids)	0.168	56.6%	0.283	73.6%

4.1 Datasets

To evaluate the similarities, we use the framework that converts texts as HINs [33]. In this case we have a lot of annotated documents for evaluation. Moreover, the meta-schema of the network is much richer than the traditional HINs such as DBLP academic network [28]. We use two benchmark text datasets to perform clustering and classification:

20Newsgroups dataset. The 20newsgroups dataset [18] contains about 20,000 newsgroups documents across 20 newsgroups. After converting them to an HIN, we call the data **20NG-HIN**.

GCAT in RCV1 dataset. The RCV1 dataset contains manually labeled newswire stories from Reuter Ltd [19]. The news documents are categorized with respect to three controlled vocabularies: industries, topics and regions. There are 103 categories including all nodes except for root in the topic hierarchy. We select the 60,608 documents under top category GCAT-HIN (Government/Social) to convert them to another HIN: **GCAT-HIN**. There are 16 leaf categories under GCAT.

After grounding the texts to the knowledge base, Freebase, the numbers of entities in different datasets are summarized in Table 2. We can see that there are more entity types than the entity types of the data used before for HIN studies. The entity types are the types of named entities mentioned in texts, such as *Politician*, *Musician* and *President*. The Freebase also has relations between entity types. The numbers of relation instances (logical forms parsed out by semantic parsing and filtering [33]) in 20NG-HIN and GCAT-HIN are $9,655,466$ and $18,008,612$. In practice we find that a lot of entity types related to a small number of path instances, thus resulting in meta-paths with few path instances. Therefore, we prune these entities using a threshold. Moreover, we limited the length of meta-paths to be less than seven. Finally we got 325 meta-paths and $1,682$ meta-paths for 20NG-HIN and GCAT-HIN, respectively [34].

4.2 Evaluation Tasks

Now we introduce the two tasks: document clustering and classification, to evaluate the similarities.

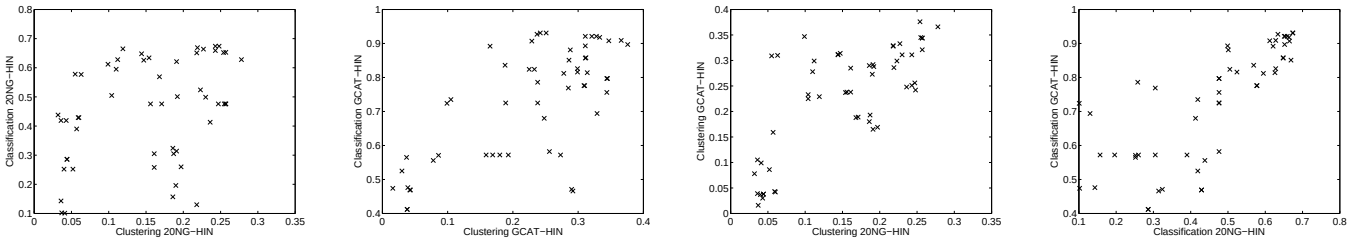
Spectral Clustering Using Similarities. To check the quality of different similarity measures in the real application scenario, we use different similarity measures as the weight matrices in the spectral clustering [36] for document clustering task. The spectral clustering algorithm is self-tuning algorithm, which means the parameters in the radial basis functions can automatically scale with the input. We compare the clustering results of 55 different similarity measures with each other. Two of them are neighborhood meta-path similarities, as shown in Table 1. The Avg(PathSim) is the average PathSim similarity over every single meta-path. The other 53 HIN similarities are distant meta-path similarities introduced in Section 3.2. We set the numbers of clusters as 20 and 16 for 20NG-HIN and GCAT-HIN according to their ground-truth labels, respectively. We employ the widely-used normalized mutual information (NMI) [27] as the evaluation measure. The NMI score is 1 if the clustering results match the category labels perfectly and 0 if the clusters are obtained from a random partition. In general, the larger the scores, the better the clustering results. Note that we have demonstrated that HIN based similarity (i.e., KnowSim) performs significantly better than BOW feature based similarities (e.g., cosine and Jaccard) used in the spectral clustering on document datasets [34], so now we focus on comparing the clustering performance between HIN based similarities.

In Table 3, we show the performance of the clustering results with different similarity measures on both 20NG-HIN and GCAT-HIN datasets. The NMI is the average of five random trials per experimental setting. The best NMI scores of both datasets are achieved by Minkowski L_p and Euclidean L_2 similarities respectively for two datasets. On average, the neighborhood meta-path similarity performs best for 20NG-HIN dataset, and L_p Minkowski family performs best for GCAT-HIN dataset. Shannon family does not perform as good as the other similarities for clustering. Especially for the similarities that are not symmetric, i.e., Kullback-Leibler, Jeffreys, and K divergence, the performances are the worst. This is reasonable since our task of clustering prefers to have a symmetric measure to evaluate pairwise document similarities. By comparing the distant meta-path similarities with neighborhood meta-path similarities, we can see that the best distant meta-path similarity is better than the best neighborhood meta-path similarity.

SVM Classification Using Similarities. We also evaluate the effectiveness of the 55 similarity measures by using the similarity measures as kernels in the document classification task with support vector machine (SVM). For the similarity measures that are kernels, we use the similarity matrix as the kernel in SVM. For the similarities that are not kernels, we adopt indefinite SVM to perform kernel based classification [35]. We perform 20-class classification and 16-class classification for 20NG-HIN and GCAT-HIN according to the number of the corresponding ground-truth categories in the dataset. Each dataset is randomly divided into 80% training and 20% testing data. We apply 5-fold cross validation on the training set to determine the optimal hyperparameter C for SVM. Then all the classification models are trained based on the full training set, and tested on the test set. We use classification accuracy as the evaluation measure. Note that we have demonstrated that HIN based similarity (i.e., KnowSim) performs significantly better than BOW feature based similarities used in the SVM classification on document datasets [35], so we just focus on comparing the classification performance between HIN based similarities.

We also show the results in Table 3. Each number is an average based on five random trials. From the table we can see that, L_p Minkowski family performs consistently the best for both datasets. Cosine and correlation similarities perform almost the same and are the best among all the similarities. The difference between these two similarities is whether we centralize the vectors. For correlation, we need to centralize the data while for cosine we do not. However, it seems the classification results are not affected by centralization. For the neighborhood meta-path similarities, we find KnowSim performs relatively better than Avg(PathSim) but still worse than the best distant meta-path similarity. Since classification is more deterministic and clustering may contain more randomness in the results, we would suggest using KnowSim when considering neighborhood meta-path similarities.

Moreover, by considering both classification and clustering results, we can see that L_p Minkowski family is in general good for both tasks and datasets. Cosine similarity, which is widely used for text data, is also good, and for classification, it is the best among all



(a) **Correlation** $\rho = 0.3766$**, significance** $p = 0.0046$**.**

(b) **Correlation** $\rho = 0.7057$**, significance** $p = 0.0000$**.**

(c) **Correlation** $\rho = 0.7404$**, significance** $p = 0.0000$**.**

(d) **Correlation** $\rho = 0.7709$**, significance** $p = 0.0000$**.**

Figure 2: Correlation results of 55 similarities based on two tasks, clustering and classification, on two datasets, 20NG-HIN and GCAT-HIN.

Table 4: Correlations of Top-10 best similarity measures. Each is with top-3 intra-family and inter-family similarity measures on both 20NG-HIN and GCAT-HIN.

Intra-family							
Datesets	Top-10 Best	Top 1	Corr.	Top 2	Corr.	Top 3	Corr.
20NG-HIN	10. Cosine	14. Correlation	1.0	8. Inner Product	0.605	9. Harmonic Mean	0.242
	14. Correlation	10. Cosine	1.0	8. Inner Product	0.605	9. Harmonic Mean	0.242
	4. Motyka	3. Czekanowski	0.736	1. Intersection	0.561	5. Ruzicka	0.551
	3. Czekanowski	4. Motyka	0.736	7. Hamming	0.498	1. Intersection	0.299
	19. SØrensen	23. Canberra	0.75	24. Lorentzian	0.721	21. Soergel	0.506
	40. Fidelity	44. Squared-chord	1.0	43. Matusita	0.985	42. Hellinger	0.983
	44. Squared-chord	40. Fidelity	1.0	43. Matusita	0.985	42. Hellinger	0.983
	15. Euclidean L_2	17. Minkowski L_p	0.982	18. Chebyshev L_∞	0.952	16. City block L_1	0.671
	1. Intersection	4. Motyka	0.561	5. Ruzicka	0.472	3. Czekanowski	0.299
	49. Jensen-Shannon	50. Jensen difference	1.0	48. TopsØe	0.992	45. Kullback-Leibler	0.164
GCAT-HIN	10. Cosine	14. Correlation	0.999	8. Inner Product	0.873	13. Dice	0.662
	14. Correlation	10. Cosine	0.999	8. Inner Product	0.875	13. Dice	0.662
	8. Inner Product	14. Correlation	0.875	10. Cosine	0.873	13. Dice	0.612
	1. Intersection	4. Motyka	0.617	2. Wave Hedges	0.365	3. Czekanowski	0.342
	40. Fidelity	44. Squared-chord	1.0	43. Matusita	0.92	42. Hellinger	0.917
	44. Squared-chord	40. Fidelity	1.0	43. Matusita	0.92	42. Hellinger	0.917
	19. SØrensen	21. Soergel	0.699	22. Kulczynski	0.674	23. Canberra	0.669
	17. Minkowski L_p	15. Euclidean L_2	0.943	18. Chebyshev L_∞	0.449	16. City block L_1	0.293
	25. Squared Euclidean	26. Pearson χ^2	0.476	32. Add. Symmetric χ^2	0.47	31. Clark	0.338
	3. Czekanowski	4. Motyka	0.73	6. Tanimoto	0.676	5. Ruzicka	0.662
Inter-family							
Datesets	Top-10 Best	Top 1	Corr.	Top 2	Corr.	Top 3	Corr.
20NG-HIN	10. Cosine	3. Czekanowski	0.674	4. Motyka	0.609	35. Kulsinski	0.605
	14. Correlation	3. Czekanowski	0.674	4. Motyka	0.609	35. Kulsinski	0.605
	4. Motyka	19. SØrensen	0.685	34. Matching	0.678	36. Rogers-Tanimoto	0.678
	3. Czekanowski	19. SØrensen	0.952	10. Cosine	0.674	14. Correlation	0.674
	19. SØrensen	3. Czekanowski	0.952	31. Clark	0.762	7. Hamming	0.734
	40. Fidelity	1. Intersection	0.815	20. Gower	0.815	53. Avg(L_1, L_∞)	0.807
	44. Squared-chord	1. Intersection	0.815	20. Gower	0.815	53. Avg(L_1, L_∞)	0.807
	15. Euclidean L_2	25. Squared Euclidean	0.984	1. Intersection	0.976	20. Gower	0.976
	1. Intersection	20. Gower	1.0	15. Euclidean L_2	0.976	25. Squared Euclidean	0.948
	49. Jensen-Shannon	51. Taneja	0.982	17. Minkowski L_p	0.799	15. Euclidean L_2	0.797
GCAT-HIN	10. Cosine	3. Czekanowski	0.928	19. SØrensen	0.925	35. Kulsinski	0.875
	14. Correlation	3. Czekanowski	0.926	19. SØrensen	0.924	35. Kulsinski	0.878
	8. Inner Product	35. Kulsinski	1.0	37. Russell-Rao	1.0	19. SØrensen	0.82
	1. Intersection	20. Gower	1.0	15. Euclidean L_2	0.798	17. Minkowski L_p	0.77
	40. Fidelity	36. Rogers-Tanimoto	0.883	38. Sokal-Michener	0.883	34. Matching	0.882
	44. Squared-chord	36. Rogers-Tanimoto	0.883	38. Sokal-Michener	0.883	34. Matching	0.882
	19. SØrensen	3. Czekanowski	0.989	10. Cosine	0.925	14. Correlation	0.924
	17. Minkowski L_p	25. Squared Euclidean	0.883	20. Gower	0.773	1. Intersection	0.77
	25. Squared Euclidean	15. Euclidean L_2	0.978	17. Minkowski L_p	0.883	50. Jensen difference	0.82
	3. Czekanowski	19. SØrensen	0.989	10. Cosine	0.928	14. Correlation	0.926

the similarities. Besides Cosine, another similarity measure, Correlation, from inner product family performs competitive with Cosine for classification. The reason is that the inner product operation just fits the formulation of SVM kernel.

4.3 Correlation of Clustering/Classification

We further use the Pearson correlation coefficient to test the consistency of clustering and classification. Figure 2(a) shows the correlation of clustering NMI and classification accuracy results on 20NG-HIN dataset. Figure 2(b) shows the correlation of clustering NMI and classification accuracy results on GCAT-HIN dataset. Both

Table 5: Correlations of Top-10 worst similarity measures. Each is with top-3 intra-family and inter-family similarity measures on both 20NG-HIN and GCAT-HIN.

Intra-family							
Datesets	Top-10 Worst	Top 1	Corr.	Top 2	Corr.	Top 3	Corr.
20NG-HIN	27. Neyman χ^2	28. Squared χ^2	0.745	29. Prob. Symmetric χ^2	0.745	25. Squared Euclidean	0.679
	46. Jeffreys	48. Tops∅e	0.021	49. Jensen-Shannon	0.02	50. Jensen difference	0.02
	Avg(PathSim)	KnowSim	0.593	-	-	-	-
	33. Yule	34. Matching	0.021	36. Rogers-Tanimoto	0.021	38. Sokal-Michener	0.021
	29. Prob. Symmetric χ^2	28. Squared χ^2	1.0	27. Neyman χ^2	0.745	25. Squared Euclidean	0.642
	30. Divergence	32. Add. Symmetric χ^2	0.414	28. Squared χ^2	0.403	29. Prob. Symmetric χ^2	0.403
	13. Dice	10. Cosine	0.006	14. Correlation	0.006	9. Harmonic Mean	0.004
	39. Sokal-Sneath	35. Kulsinski	0.03	37. Russell-Rao	0.03	34. Matching	0.001
	31. Clark	32. Add. Symmetric χ^2	0.138	26. Pearson χ^2	0.128	27. Neyman χ^2	0.111
	23. Canberra	21. Soergel	0.788	19. SØrensen	0.75	24. Lorentzian	0.681
GCAT-HIN	6. Tanimoto	5. Ruzicka	0.984	3. Czekanowski	0.676	4. Motyka	0.261
	21. Soergel	22. Kulczynski	0.984	19. SØrensen	0.699	23. Canberra	0.442
	43. Matusita	42. Hellinger	1.0	41. Bhattacharyya	0.998	40. Fidelity	0.92
	5. Ruzicka	6. Tanimoto	0.984	3. Czekanowski	0.662	4. Motyka	0.268
	22. Kulczynski	21. Soergel	0.984	19. SØrensen	0.674	23. Canberra	0.423
	42. Hellinger	43. Matusita	1.0	41. Bhattacharyya	0.999	40. Fidelity	0.917
	46. Jeffreys	48. Tops∅e	0.047	49. Jensen-Shannon	0.039	50. Jensen difference	0.039
	33. Yule	34. Matching	0.006	36. Rogers-Tanimoto	0.006	38. Sokal-Michener	0.006
	47. K divergence	50. Jensen difference	0.158	49. Jensen-Shannon	0.157	48. Tops∅e	0.144
	45. Kullback-Leibler	48. Tops∅e	0.369	49. Jensen-Shannon	0.323	50. Jensen difference	0.318
Inter-family							
Datesets	Top-10 Worst	Top 1	Corr.	Top 2	Corr.	Top 3	Corr.
20NG-HIN	27. Neyman χ^2	9. Harmonic Mean	0.757	1. Intersection	0.703	20. Gower	0.703
	46. Jeffreys	51. Taneja	0.02	27. Neyman χ^2	0.017	53. Avg(L_1, L_∞)	0.017
	Avg(PathSim)	3. Czekanowski	0.416	19. SØrensen	0.387	4. Motyka	0.207
	33. Yule	5. Ruzicka	0.026	22. Kulczynski	0.026	4. Motyka	0.021
	29. Prob. Symmetric χ^2	9. Harmonic Mean	0.981	1. Intersection	0.725	20. Gower	0.725
	30. Divergence	41. Bhattacharyya	0.78	42. Hellinger	0.78	43. Matusita	0.78
	13. Dice	KnowSim	0.023	27. Neyman χ^2	0.02	40. Fidelity	0.009
	39. Sokal-Sneath	52. Kumar-Johnson	0.043	8. Inner Product	0.03	32. Add. Symmetric χ^2	0.03
	31. Clark	23. Canberra	0.995	7. Hamming	0.981	11. Kumar-Hassebrook	0.981
	23. Canberra	7. Hamming	0.995	11. Kumar-Hassebrook	0.995	12. Jaccard	0.995
GCAT-HIN	6. Tanimoto	21. Soergel	1.0	22. Kulczynski	0.984	40. Fidelity	0.811
	21. Soergel	6. Tanimoto	1.0	5. Ruzicka	0.984	40. Fidelity	0.811
	43. Matusita	6. Tanimoto	0.763	21. Soergel	0.763	36. Rogers-Tanimoto	0.751
	5. Ruzicka	22. Kulczynski	1.0	21. Soergel	0.984	36. Rogers-Tanimoto	0.788
	22. Kulczynski	5. Ruzicka	1.0	6. Tanimoto	0.984	36. Rogers-Tanimoto	0.788
	42. Hellinger	6. Tanimoto	0.759	21. Soergel	0.759	36. Rogers-Tanimoto	0.747
	46. Jeffreys	51. Taneja	0.057	11. Kumar-Hassebrook	0.056	12. Jaccard	0.056
	33. Yule	52. Kumar-Johnson	0.015	2. Wave Hedges	0.008	KnowSim	0.007
	47. K divergence	25. Squared Euclidean	0.141	15. Euclidean L_2	0.135	4. Motyka	0.132
	45. Kullback-Leibler	53. Avg(L_1, L_∞)	0.384	16. City block L_1	0.377	24. Lorentzian	0.377

the Pearson correlation coefficient and its significant test value are shown in each caption of the sub-figure. The correlation on 20NG-HIN dataset is not as high as GCAT-HIN dataset, but it is still significantly correlated at 0.01 level. Both results mean that the clustering and classification results are consistent. There are some differences between spectral clustering and SVM classification. Spectral clustering assumes data points are on a manifold and assumes local linearities. SVM using kernel assumes the high dimensional Hilbert space is linearly separable given a kernel. The similarities preserve more locality may be better for spectral clustering, while the similarities that can map the data onto a linearly separable space may work better for classification.

Moreover, Figure 2(c) shows the correlation of clustering results between two datasets, and Figure 2(d) shows the correlation of classification results between two datasets. It seems the correlation scores are higher than the scores between clustering and classification. This is reasonable as we have analyzed that spectral clustering and SVM may have different preferences. This also indicates that the similarities are robust and scalable.

4.4 Correlation Between Similarities

We finally analyze the correlation between each pair of similarity measures. For both datasets, we have the similarities' scores of pairwise documents. Then for each pair of similarity measures, we use the lists of similarity scores to compute the correlation between the pair of similarity measures.

We sort Table 3 based on the classification results of both datasets, and obtain the top ten best similarity measures and top ten worst ones. For each similarity measure, we use the correlation to retrieve top three similarity measures. For the ten best similarity measures, we show the results in Table 4. For the ten worst similarity measures, we show the results in Table 5.

From Table 4 we can see that, for the best similarity measures, such as cosine, the top intra-family similar measures are *14. Correlation*, *8. Inner Product*, and *9. Harmonic Mean* for 20NG-HIN dataset, and *14. Correlation*, *8. Inner Product*, and *13. Dice* for GCAT-HIN dataset. If we refer back to Table 3, we can see that *13. Dice* similarity performs on 20NG-HIN not as good as it performs on GCAT-HIN

dataset. For the inter-family similarities, we can also see some interesting results. For example, for cosine, the most correlated similarity measures are *3. Czekanowski* (Intersection family), *4. Motyka* (Intersection family), and *35. Kulsinski* (Binary family) on 20NG-HIN dataset, and *3. Czekanowski* (Intersection family), *19. SØrensen* (L_1 family), and *35. Kulsinski* (Binary family) on GCAT-HIN dataset. Inner product is similar to intersection in the sense that the only difference is whether considering the weights. Intersection is further similar to binary if the logic of binary operation is "AND."

From Table 5 we can see that, for the worst similarity measures, there are also interesting findings. Some of the bad similarity measures are highly correlated. For example, for GCAT-HIN dataset, *6. Tanimoto* is highly correlated with *5. Ruzicka* inside family, and *21. Soergel* outside family. The classification results are *6. Tanimoto*: 41.2%, *5. Ruzicka*: 46.9%, and *21. Soergel*: 41.2%. Moreover, the good similarity measures in Shannon family such as *49. Jensen-Shannon* is relatively highly correlated with *45. Kullback-Leibler* on GCAT-HIN dataset. However the correlation score is not as high as the other top similar scores. This is because *45. Kullback-Leibler* is not a symmetric similarity when *49. Jensen-Shannon* is.

5 CONCLUSION

In this paper, we study the problem of entity proximity in HINs, and propose distant meta-path similarity to fully capture HIN semantics between entities when measuring the proximity. We then derive 53 distant meta-path similarity measures and experimentally compare them in two text-based HIN datasets. Experimental results show that cosine similarity is consistently good for general use, and the L_p Minkowski family is outstanding on both datasets. Although our similarities are tested on text-based HINs, they can be simply applied to other HIN datasets such as academic networks (e.g., DBLP and PubMed) or social networks (e.g., Facebook and Twitter).

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for the helpful comments. The co-author Yangqiu Song is supported by China 973 Fundamental R&D Program (No.2014CB340304), HKUST Initiation Grant IGN16EG01, and Hong Kong CERG Project 26206717; Haoran Li and Ming Zhang are supported by the National Natural Science Foundation of China (NSFC Grant Nos. 61472006, 61772039 and 91646202); Yizhou Sun is supported by NSF Career Award #1741634 and NSF IIS #1705169; and Jiawei Han is supported by the U.S. Army Research Lab No. W911NF-09-2-0053 (NSCTA), U.S. NSF IIS-1320617, IIS 16-18481 and IIS 17-04532, and NIH BD2K 1U54GM114838 from NIGMS.

REFERENCES

[1] A Bhattycharrya. 1943. On a Measure of Divergence Between Two Statistical Populations Defined by Probability Distributions. *Bulletin of the Calcutta Mathematical Society* 35 (1943), 99–110.
[2] J Roger Bray and John T Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs* 27, 4 (1957), 325–349.
[3] Sung-Hyuk Cha. 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences* 1, 4 (2007), 300–307.
[4] Michel Marie Deza and Elena Deza. 2009. *Encyclopedia of distances*. Springer.
[5] SS Dragomir, J Sunde, and C Buse. 2000. New inequalities for Jeffreys divergence measure. *Tamsui Oxford Journal of Mathematical Sciences* 16, 2 (2000), 295–309.
[6] Richard O Duda, Peter E Hart, and David G Stork. 2012. *Pattern classification*. John Wiley & Sons.
[7] Daniel G Gavin, W Wyatt Oswald, Eugene R Wahl, and John W Williams. 2003. A statistical approach to evaluating distance metrics and analog assignments for pollen records. *Quaternary Research* 60, 3 (2003), 356–367.
[8] John C Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics* (1971), 857–871.
[9] Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S. Yu. 2010. Mining Knowledge from Databases: An Information Network Analysis Approach. In *SIGMOD*. 1251–1252.
[10] Ts Hedges. 1976. An empirical modification to linear wave theory. *ICE* 61, 3 (1976), 575–579.
[11] JETA Inder. 1995. New developments in generalized information measures. *Advances in Imaging and Electron Physics* 91 (1995), 37–135.
[12] Harold Jeffreys. 1946. An invariant form for the prior probability in estimation problems. In *RCLA*, Vol. 186. 453–461.
[13] Eugene F Krause. 2012. *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation.
[14] Joseph B Kruskal and Myron Wish. 1978. *Multidimensional scaling*. Vol. 11. Sage.
[15] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.
[16] BVK Vijaya Kumar and Laurence Hassebrook. 1990. Performance measures for correlation filters. *Applied optics* 29, 20 (1990), 2997–3006.
[17] Pranesh Kumar and Andrew Johnson. 2005. On a symmetric divergence measure and information inequalities. *Journal of Inequalities in pure and applied Mathematics* 6, 3 (2005).
[18] Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *ICML*. 331–339.
[19] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *JMLR* 5 (2004), 361–397.
[20] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *TIT* 37, 1 (1991), 145–151.
[21] Ulrike Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 395–416.
[22] Kameo Matusita. 1955. Decision rules, based on the distance, for problems of fit, two samples, and estimation. *The Annals of Mathematical Statistics* (1955), 631–640.
[23] Valentin Monev. 2004. Introduction to similarity searching in chemistry. *MCMCC* 51 (2004), 7–38.
[24] Masaaki Morisita. 1959. Measuring of interspecific association and similarity between communities. *MFSKU* 3 (1959), 65–80.
[25] Karl Pearson. 1900. On the Criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling. *Philos. Mag.* 50, 302 (1900), 157–175.
[26] Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter* 5 (1948), 1–34.
[27] Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *JMLR* 3 (2003), 583–617.
[28] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. *PVLDB* (2011), 992–1003.
[29] IJ Taneja. 2001. Generalized Information Measuresand Their Applications. *on-line book* (2001).
[30] Taffee T Tanimoto. 1957. IBM internal report. *Nov* 17 (1957), 1957.
[31] Flemming Topsøe. 2000. Some inequalities for information divergence and related measures of discrimination. *TIT* 46, 4 (2000), 1602–1609.
[32] Ferdinand Van Der Heijden, Robert Duin, Dick De Ridder, and David MJ Tax. 2005. *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. John Wiley & Sons.
[33] Chenguang Wang, Yangqiu Song, Ahmed El-Kishky, Dan Roth, Ming Zhang, and Jiawei Han. 2015. Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. In *KDD*. 1215–1224.
[34] Chenguang Wang, Yangqiu Song, Haoran Li, Ming Zhang, and Jiawei Han. 2015. KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. In *ICDM*. 506–513.
[35] Chenguang Wang, Yangqiu Song, Haoran Li, Ming Zhang, and Jiawei Han. 2016. Text Classification with Heterogeneous Information Network Kernels. In *AAAI*. 2130–2136.
[36] Lihi Zelnik-manor and Pietro Perona. 2005. Self-Tuning Spectral Clustering. In *NIPS*, L.K. Saul, Y. Weiss, and L. Bottou (Eds.). 1601–1608.
[37] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. 2006. *Similarity search: the metric space approach*. Vol. 32. Springer Science & Business Media.