Physica A 524 (2019) 696–707

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# TSS: Temporal similarity search measure for heterogeneous information networks

Golnaz Nikmehr [a], Mostafa Salehi [a,b,*], Mahdi Jalili [c]

[a] Faculty of New Sciences and Technologies, University of Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Science (IPM), P.o.Box 19395-5746, Tehran, Iran
[c] School of Engineering, RMIT University, Australia

## HIGHLIGHTS

- We propose a temporal similarity measure for heterogeneous information networks.
- The proposed metric is based on metapath strategy and considers time of the interaction.
- Experiments represent the effectiveness of our measure, in comparison with the state-of- the-art similarity computation methods.

## ARTICLE INFO

*Article history:*
Received 10 May 2018
Received in revised form 19 October 2018
Available online 2 May 2019

*Keywords:*
Social networks
Similarity search
Heterogeneous networks
Meta-path
Time of interaction
Recommendation systems

## ABSTRACT

Many real-world phenomena can be modeled as networked systems. Some of these systems consist of heterogeneous nodes and edges. Similarity search is a fundamental operation in network systems, which is a basis for various applications such as link prediction and recommendation. This manuscript introduces a temporal similarity measure for heterogeneous networks. The proposed metric is based on metapath strategy and considers time of the interaction. We provide detailed properties of the proposed temporal similarity measures. Our experimental results on a number of real heterogeneous social networks show that incorporating time in the computation of the similarity significantly improves the performance as compared to the state-of-the-art similarity computation methods.



## 1. Introduction

Many natural and man-made phenomena can be modeled as networks ranging from social networks and the Internet to ecological systems, human brain and biological systems. Some real-world systems can be better modeled as a heterogeneous networks, where multi-typed objects are connected together with different relationships [1,2]. Heterogeneous information networks are also known as network of networks [3], interconnected networks [4], multiplex networks [3,5,6] or multilayer networks [7,8]. DBLP Network (Fig. 1(a)), is an example of heterogeneous networks, which has different kind of objects like Authors, Papers and Conference, and also different types of relationships between them such as Write and Participate.

Similarity search is a significant topic in heterogeneous information networks [8] with the goal finding objects from a dataset which are similar to a given input object (query). Nearest neighbor search and range queries [9] are two basic subjects in this context. Similarity search has significant applications in different areas including link prediction [10] and

* Corresponding author at: Faculty of New Sciences and Technologies, University of Tehran, Iran.
*E-mail addresses:* g.nikmehr@ut.ac.ir (G. Nikmehr), mostafa_salehi@ut.ac.ir (M. Salehi), mahdi.jalili@rmit.edu.au (M. Jalili).

https://doi.org/10.1016/j.physa.2019.04.207
0378-4371/

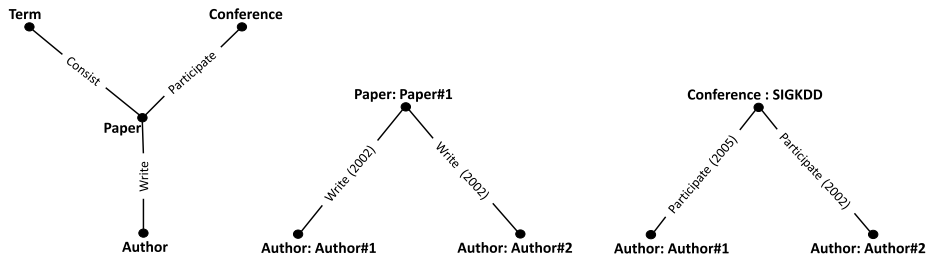


**Fig. 1.** (a) network schema of DBLP network, (b) an example of Author–Paper–Author meta-path in this networks, Author1 and Author2 have written a paper in 2002. (c) an example of Author–Conference–Author meta-path, Author1 and Author2 have participated in SIGKDD conference in 2002 and 2005, respectively.

recommendation [11]. Accuracy and computational complexity is important in such applications. Sun et al. [2] proposed a framework to represent heterogeneous networks, which has two basic elements: network schema and meta-path. The network schema is a general view of the network that shows the number of objects and relationships and the meta-path is a sequence of related objects forming a path in the network schema. Fig. 1 represent schema and example of two types of meta-paths in DBLP network.

A number of measures have been proposed to find similarity between two objects in heterogeneous information networks [2,12–15]. PathSim is a meta-path based similarity measure introduced for to find similar pairs of objects in heterogeneous networks [2]. This measure has become a basic measure in this field. Some other measures were introduced to improve similarity search considering special conditions such as considering both symmetric and non-symmetric meta-paths [12] or using path weighting with new attributes [16]. In heterogeneous networks interaction, time is a valuable attribute, as properties of real social networks may change over time. There is only one work that have considered the effect of interaction time on the similarity between objects [17] and path time weighting. In this paper we proposed a novel similarity measure that incorporates the interaction time in the calculations and show its effectiveness over existing state-of-the-art methods.

In Fig. 1(b), for APA (Author–Paper–Author) meta-path, we assume that Author1 and Author2 are co-authors of a paper written in a particular time. We also know that meta-paths are formed with relationships. In this particular example, we have two relationships of type Author1–Paper1 and Paper1–Author2. The Author1–Paper1 relationship means that Author1 has written the paper in year 2002. The same is true for Paper1–Author2. In this case, when two authors collaborate in the same time, they are interested in common research areas and they will likely collaborate in the future. In Fig. 1(c) however in ACA (Author–Conference–Author) meta-path, having Author1 and Author2 participated in *SIGKDD* conference, unlike the previous example, this participation may occur in different times. In this example, Author1 has participated in 2005 and Author2 in 2002. Thus, in this case relationships of the meta-path have different times. So in this simple example, we showed a general view of our main idea about considering time of interactions between objects in heterogeneous networks.

The main contributions of this study are as follows:

- It proposes a new similarity measure, TSS—Temporal, that captures the time of interactions between objects in heterogeneous networks.
- Our experiments represent the effectiveness of our measure, in comparison with PathSim and other methods.

## 2. Related works

Similarity measures have been widely studied in networks. For homogeneous networks, there are two popular measures in similarity measure: Personalized PageRank and SimRank. Personalized PageRank [18] is an asymmetric similarity measure that evaluates the probability of visiting the objects which is based on the concept of random walks. SimRank [19] is a symmetric similarity measure, which works by propagating the pairwise similarity score to neighboring pairs.

In heterogeneous networks, some similarity measures are based on meta-path, while some others do not use the meta-path information. PathSim [2] is the most widely-used meta-path-based measure. It compute the similarity between any two objects based on a given meta-path using path counting. Number of paths between the objects and the balance of their visibility, where the visibility is defined as the number of path instances between them. PathSimExt [16] is a measure that revisit the definition of PathSim by introducing external support to enrich the result of PathSim. External support is the supportive information such as the number of citations in bibliographic data. AvgSim [15] is based on random walk processes along the given meta-path and the reverse meta-path. AsymSim [20] uses asymmetric relations of meta-paths in heterogeneous networks to compute the similarity between objects. HeteSim [12] uses pairwise random walks to measure the similarity of objects in a given meta-path. NetSim [13] is a structural-based similarity measure which computes the similarity between centers in an x-star network. JoinSim's goal is to improve PathSim to work

with large-scale networks [14], it uses pruning expensive similarity computation by introducing bucket pruning based locality sensitive hashing indexing. There are only few measures that use the time information to compute the similarity values [17]. As discussed on Fig. 1, this paper did not consider interaction times and it ignored some meta-paths in while involving time information.

## 3. Prelliminaries

Before detailing the proposed similarity measure based on meta-path ... in heterogeneous information networks, let us introduce definitions.

**Definition 1** (*Information Network*)**.** An information network is defined as a digraph $G = (V, E)$, with an object type mapping function $F : V \rightarrow A$ and a link type mapping function $H : E \rightarrow R$. an object $v \in V$ belongs to one particular object type, i.e., $F(v) \in A$, and a link $e \in E$ belongs to particular relation i.e., $H(e) \in R$. Information networks are divided into two types: (1) heterogeneous information network if $|A| > 1$ or $|R| > 1$; and (2) homogeneous information network otherwise [2].

**Definition 2** (*Network Schema*)**.** The network schema, $T_G = (A, R)$, is a meta-level description for a heterogeneous network $G = (V, E)$ with an object type mapping function $F : V \rightarrow A$ and a link type mapping function $H : E \rightarrow R$ [2].

**Definition 3** (*Meta-path*)**.** A meta-path is a sequence of typed relations between objects. In other words, it is a path defined on network schema $T_G = (A, R)$. meta-path $P$ is shown as $P = A_1A_2 \ldots A_l$, where each $A_1$ to $A_l$ shows the typed object. The formal definition of a meta-path is $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_l$, which defines a composite relation $R_1 \circ R_2 \circ \cdots \circ R_l$ between types $A_1$ to $A_l$ where $\circ$ denotes the composition operator on relations [2].

**Definition 4** (*Commuting Matrix*)**.** A commuting matrix of a meta-path $P = A_1A_2 \ldots A_l$ is multiplication of adjacency matrix of typed objects. This concept is shown as $M = W_{A_1A_2}W_{A_2A_3} \ldots W_{A_{l-1}A_l}$, where $W_{A_iA_j}$ is the adjacency matrix between type $A_i$ and type $A_j$. $M(i, j)$ represents the number of path instances between object $x_i \in A_i$ and object $y_i \in A_l$ under meta-path $P$.

**Definition 5** (*TimeInfo Matrix*)**.** TimeInfo matrix is a matrix that stores time information. Given a meta-path $P = A_1A_2 \ldots A_m \ldots A_{l-1}A_l$ and time period $t_1 \leq Time \leq t_2$, we define matrix $Y_{\{A_{m-1}\} \times \{A_m Time\}}$ as TimeInfo matrix. As we see, one of the dimensions of this matrix is depended on the time. The size of the first dimension of this matrix is the number of the instance of type $A_{m-1}$ and the size of the second dimension is a combination of all instances of type $A_m$ with all years in the period of *Time*. In general, each cell of this matrix shows the presence or absence of an instance of type $A_{m-1}$ in years of *Time* based on the instance of type $A_m$.

**Definition 6** (*Time Weight Matrix*)**.** Time weight matrix is a matrix that stores time weight gained from a time function. Given a meta-path $P = A_1A_2 \ldots A_m \ldots A_{l-1}A_l$ and time period $t_1 \leq Time \leq t_2$, we define matrix $F_{\{A_m Time\} \times \{A_m Time\}}$ as time weight matrix. The dimensions of this square matrix are based on the time and the size of each dimension is the number of combination of all instances of type $A_m$ with all years in the period of *Time*. Each cell of the matrix shows relation between times $t_i$ and $t_j$ based on a time function $f(t_i, t_j)$. In this matrix, if instances of type $A_m$ are equal during a time period, that cell is filled with the time function, otherwise it is filled by zero.

## 4. TSS - temporal similarity search

In this section we propose our new temporal measure. Given a meta-path $P = A_1A_2 \ldots A_m \ldots A_{l-1}A_l$, we define our base matrix that is the Time commuting matrix, which is used to find similarity of two objects in network:

$$MT_\rho = W_{A_1A_2}W_{A_2A_3} \ldots Y_{\{A_{m-1}\}\{A_m Time\}} F_{\{A_m Time\}\{A_m Time\}} Y^T \ldots W_{A_2A_1} \tag{1}$$

As we see in Eq. (1), the Time Commuting Matrix is a multiplication of the adjacency matrix of typed objects, TimeInfo matrix, time weight matrix and its transpose $Y^T$. Now we can compute similarity of two object $a$ and $b$ as below:

$$S(a, b) = \frac{2 \times MT_\rho[a, b]}{MT_\rho[a, a] + MT_\rho[b, b]} \tag{2}$$

As we mentioned earlier to consider interaction time in our measure, we need both $\triangle d$ as distance of interaction times to current time and $\triangle t$ as distance between interaction times (denote in Eq. (3)) [21]. We affect these in time functions exponential, linear or any other kinds. Here we use two exponential functions: $f_1 = e^{(-\alpha \times \triangle d) + (-\alpha \times \triangle t)}$ [22], $f_2 = \alpha^{(\triangle d) + (\triangle t)}$ [17] and one linear function: $f_3(t_i, t_j) = \frac{1}{(\beta)\triangle t + (1-\beta)\triangle d}$ [21], where $\alpha$ and $\beta$ are parameters between zero and one. We use exponential functions in bibliographic fields based on the assumption that older collaboration are generally

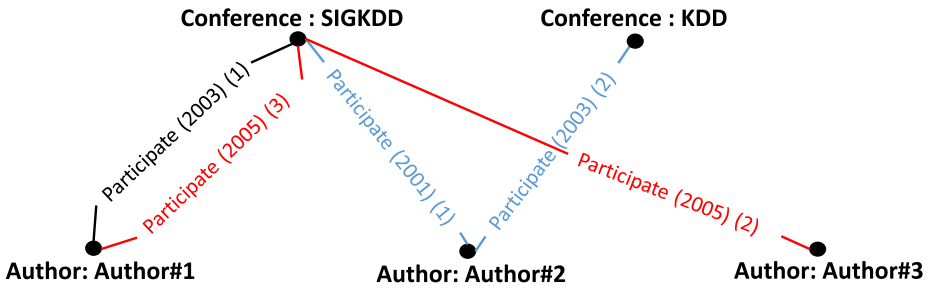


**Fig. 2.** An example showing a heterogeneous network of three authors that have participated in two conferences.

**Table 1**
Result of two similarity measure (PathSim and TSS) on the example of Fig. 2.

| Pair of authors | Similarity search measures | |
|---|---|---|
| | *PathSim* measure | *TSS* measure |
| (*Author*1, *Author*2) | 0.38 | 0.20 |
| (*Author*2, *Author*3) | 0.44 | 0.22 |
| (*Author*1, *Author*3) | 0.80 | 0.83 |

less correlated with a author's current interests or a paper's and conference's current characteristic. But in films fields, the weight of the less recent path drops slower than exponential function, so we select the linear one.

$$\begin{aligned} \triangle t &= \mid t_i - t_j \mid + 1, \\ \triangle d &= \mid min(t_i, t_j) - T \mid . \end{aligned} \tag{3}$$

Now, let us consider a toy example to discuss the effectiveness of our new measure. In this example, there are three authors Author1, Author2, and Author3, that have participated in two conferences *SIGKDD* and *KDD* between years 2000 and 2005. Fig. 2 shows the heterogeneous network of this example. In the example, we compare two measures PathSim as the state-of-art measure in this field and TSS, as our proposed measure. We perform the similarity search for ACA meta-path and use $f_3$ as a weight function with $\beta = 0.5$ and $T = 2009$. First we create the time commuting matrix for this example as:

$$MT_\rho = Y_{\{A\}\{CT\}} F_{\{CT\}\{CT\}} Y_{\{CT\}\{A\}} \tag{4}$$

We initialized the matrices as below:

$$MT_\rho = \begin{pmatrix} 0 & 1 & 3 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/9 & 1/11 & 1/13 & 0 & 0 \\ 1/11 & 1/7 & 1/9 & 0 & 0 \\ 1/13 & 1/9 & 1/5 & 0 & 0 \\ 0 & 0 & 0 & 1/9 & 1/12 \\ 0 & 0 & 0 & 1/12 & 1/6 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 3 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Finally, we have the time commuting matrix $M_\rho$ as:

$$MT_\rho = \begin{pmatrix} 2.61 & 0.32 & 1.42 \\ 0.32 & 0.55 & 0.15 \\ 1.42 & 0.15 & 0.8 \end{pmatrix}$$

Eq. (2) is then used to compute the similarity between pairs of authors. For example for author pair (Author1, Author2):

$$S(Author1, Author2) = \frac{2 \times 0.32}{2.61 + 0.55} = 0.20$$

Table 1 shows the results of similarity search between pairs of authors in the toy example. For example, the similarity of (Author1,Author2), in PathSim is higher than TSS measure. Note that PathSim counts instance paths between two authors without considering the time, of the interaction. Author1 participated in conference *SIGKDD* two times in years 2003 and 2005, while Author2 participated in the same conference only once in 2000. As mentioned before, for weighting the paths we consider the distance between the interaction times and the distance between each relation's time and $T$. For (Author1,Author2), both these quantities are high, and thus the score of this pair in TSS decreases compared to PathSim. (Author2,Author3) pair can be discussed same as (Author1,Author2). The value of TSS for (Author1,Author3) pair is higher than PathSim, because these two authors participated in a conference in the same year and what has short distance to $T$.

*4.1. Properties of TSS*

Similarity measures have some attributes such as symmetry, self-maximum and Unitary range. We mathematically introduce these attributes.

(1) **Symmetry:**

$$S(a, b) = S(b, a)$$

**Proof.** Having:

$$\frac{2 \times MT_\rho[a, b]}{MT_\rho[a, a] + MT_\rho[b, b]} = \frac{2 \times MT_\rho[b, a]}{MT_\rho[a, a] + MT_\rho[b, b]},$$

and:

$$MT_\rho[a, b] = MT_\rho[b, a]$$

Finally, we should prove that $MT_\rho$ matrix is symmetric.

$MT_\rho$ is a result of multiplication of matrices:

$$MT_\rho = WXW^T; X = YFY^T$$

$F$ is time weight matrix that is symmetric and constant, thus we have:

$$X_{ij} = Y(i, :) \cdot Y^T(j, :) = Y^T(j, :) \cdot Y^T(i, :) = X_{ji}$$

The above expression shows that $X$ is symmetric. $MT_\rho = WXW^T$ is symmetric too, let us assume the dimensions of the matrices as $W_{n*m}$, $X_{m*m}$ and $W^T_{m*n}$; we have:

$$MT_\rho = \begin{pmatrix} W_1 & \dots & W_m \end{pmatrix} \begin{pmatrix} X_{11} & \dots & X_{1m} \\ . & \dots & . \\ . & \dots & . \\ . & \dots & . \\ X_{m1} & \dots & X_{mm} \end{pmatrix} \begin{pmatrix} W_1^T \\ . \\ . \\ . \\ W_m^T \end{pmatrix}$$

we know that;

$$\sum_i W_i X_{1i} W_1^T + \sum_i W_i X_{2i} W_2^T + \dots + \sum_i W_i X_{mi} W_m^T = \sum_i^n \sum_j^n W_i X_{ij} W_j^T$$

The sum of $W_i X_{ij} W_j^T$ depends whether $i$ and $j$ are equal or not. If $i = j$, this phrase is symmetric. If $i \neq j$ then are $W_i X_{ij} W_j^T$ and $W_j X_{ji} W_i^T$ and we know that matrix $X$ is symmetric ($X_{ij} = X_{ji}$),

$$X_{ij}(W_i W_j^T + W_j W_i^T)$$

Now, one should show that $(W_i W_j^T + W_j W_i^T)$ is symmetric. Thus, $W_i$ and $W_j$ are replaced by $M$ and $N$, respectively, resulting in:

$$(W_i W_j^T + W_j W_i^T) = MN^T + NM^T$$

We know that $(A^T)^T = A$ and $(AB)^T = A^T B^T$, thus

$$= MN^T + (MN^T)^T = B + B^T$$

At the end we have $U = B + B^T$ which is symmetric, because:

$$U_{ij} = b_{ij} + bji$$
$$U_{ji} = b_{ij} + bji$$
$$U_{ji} = U_{ij}$$

This proves that $MT_\rho$ is a symmetric matrix.

(2) **Self-maximum:**

$$S(a, a) = 1$$
$$S(a, a) = \frac{2 \times MT_\rho[a, a]}{MT_\rho[a, a] + MT_\rho[a, a]} = 1$$

(3) **Unitary range**

$$0 \le S(a,b) \le 1$$

$$0 \le \frac{2 \times MT_\rho[a,b]}{MT_\rho[a,a] + MT_\rho[b,b]} \le 1$$

If we assume that $MT_\rho$ is computed as below:

$$MT_\rho = WXW^T; X = YFY^T$$

and the dimensions of matrices are $MT_{\rho_{n*n}}$, $W_{n*m}$, $Y_{m*p}$, $F_{p*p}$, $Y^T_{p*m}$ and $W^T_{m*n}$; The values of $F$ are greater than zero:

$$F_{ab} \ge 0,$$
$$(1 \le a \le p, 1 \le b \le p)$$

**Proof.**

$$Y(a,:) = (r_1F_1, r_2F_2, \ldots, r_pF_p)$$
$$Y(b,:) = (u_1F_1, u_2F_2, \ldots, u_pF_p)$$

($r_k$ and $u_k$ are exists if $1 \le k \le p$.)

For matrix $X$:

$$X_{ab} = \sum_{k=1}^{p} r_k u_k F_k \ge 0$$

$$X_{aa} = \sum_{k=1}^{p} {r_k}^2 {F_k}^2 \ge 0$$

$$X_{bb} = \sum_{k=1}^{p} {u_k}^2 {F_k}^2 \ge 0$$

From above we conclude that the values of matrix $X$ are larger than zero. Now, for $MT_\rho$:

$$W(a,:) = (q_1X_1, q_2X_2, \ldots, q_mX_m)$$
$$W(b,:) = (z_1X_1, z_2X_2, \ldots, z_mX_m)$$

($z_d$ and $q_d$ are exists if $1 \le d \le m$.)

In way:

$$MT_{\rho_{ab}} = \sum_{d=1}^{m} q_d z_d X_d \ge 0$$

$$MT_{\rho_{aa}} = \sum_{d=1}^{m} {q_d}^2 {X_d}^2 \ge 0$$

$$MT_{\rho_{bb}} = \sum_{d=1}^{m} {z_d}^2 {X_d}^2 \ge 0$$

From three above phrases:

$$S(a,b) = \frac{2 \times MT_\rho[ab]}{MT_\rho[aa] + MT_\rho[bb]} = \frac{2 \times \sum_{d=1}^{m} q_d z_d X_d}{\sum_{d=1}^{m} {q_d}^2 {X_d}^2 + \sum_{d=1}^{m} {z_d}^2 {X_d}^2} \ge 0$$

$$\text{so } S(a,b) \ge 0$$

We need to show $S(a,b) \le 1$:

$$\sum_{d=1}^{m} {q_d}^2 {X_d}^2 + \sum_{d=1}^{m} {z_d}^2 {X_d}^2 - \sum_{d=1}^{m} 2 \times q_d z_d X_d \ge 0$$

$$\sum_{d=1}^{m} {q_d}^2 {X_d}^2 + \sum_{d=1}^{m} {z_d}^2 {X_d}^2 \ge 2 \times \sum_{d=1}^{m} q_d z_d X_d$$

$$\text{so } S(a,b) \le 1$$

Finally we have $0 \le S(a,b) \le 1$.

The proposed similarity measure has all these above attributes.

### 4.2. Ranking similarity score based on proposed measure TSS

In this section we introduce an algorithm for similarity search based on meta-path and time. Our goal in this algorithm is to return an ordered list of $k$ objects that have the highest similarity to query $q$. In general, this algorithm has some steps as below:

(1) Matrix–Matrix multiplication to get $MT_\rho$.

(2) Computing similarity score between query $q$ and all of available objects $j$ with $S(q,j) = \frac{2*MT_\rho[q,j]}{MT_\rho[q,j]+MT_\rho[j,j]}$.

(3) Extracting top-k similar objects for query $q$.

A pseudo-code of algorithm is presented in Algorithm 1. Time complexity of matrix multiplication to compute C = AB, if we assume matrix A and B with $m \times n$ and $n \times m$ dimensions, is $O(m * n^2)$. According to nature of sparsity in heterogeneous networks, by using sparse matrices we can decrease runtime matrix multiplication. We use CSC format for storing sparse matrices in most sparse matrix packages that is column-by-column method and time complexity $O(flops+nnz(B)+n+m)$ [23], where flops is defined as the number of nonzero arithmetic operations required to compute the output matrix C and $nnz(B)$ are non-zero elements of matrix B. This is time complexity for part one of algorithm 1 and for part two in worst case the query has all other objects in its neighbor, $O(m)$.

**Algorithm 1:** Similarity Ranking Algorithm based on TSS Measure.

**Input**: Query $x_i$, Commuting Matrix **W**, Time Info Matrix**Y**, Time Weight Matrix **F**, K
**Output**: $Top-k\ List$, $SortList$
**begin**
  $MT_\rho = [W][Y][F][Y]^T[W]^T$
  $neighbors(x_i) = \{y_k | MT_\rho(x_i, y_k) \neq 0\}$
  $x_j =$ All of $MT_\rho.neighbors(x_i)$
  $SimilarityScore = []$
  **forall** $j \in x_j$ **do**
    $values = 2 * MT_\rho(i,:)MT_\rho(j,:)/(MT_\rho(i,i) + MT_\rho(j,j))$;
    $SimilarityScore.add(values)$
  **end**
**end**
$SimilarityScore.sort()$;
$SortList = SimilarityScore.topk(K)$;
**return** $SortList$;

## 5. Performance analysis

For our experiment, we use two available bibliographic datasets including DBLP and ACM and a multimedia dataset that is IMDB.

### 5.1. Datasets

In each dataset we choose some queries to assess the performance of the metrics. Depending on the dataset our queries are authors or actors and their selection is based on the number of papers or films that they have participated. For each dataset we perform an analysis of the effect of time, when shows that co-authors or co-actors of a particular author or actor in recent years are more common than those in further years. meta-paths that we study in this paper are APA (Author–Paper–Author), ACA (Author–Conference–Author), ADA (Actor–Director–Actor) and AMA (Actor–Movie–Actor).

(1) **DBLP − 1**: This dataset is a biblographic network [2]. It contains 46122 authors, 54,181 papers and 404 conferences during years 1963 to 2009. The number of queries for this data is 92.

(2) **DBLP − 2**: This dataset contains 1321 authors, 14,901 papers and 252 conferences during years 2000 to 2009. In this dataset all authors have participated at least in 10 papers. The number of queries for this data is 444.

(3) **ACM**: The dataset contains 9658 authors, 339,581 papers and 4277 conferences between years 2000 to 2013. The authors of this dataset have participated in at least 30 papers. We use 97 queries for this dataset.

(4) **IMDB**: This dataset is from IMDB networks [24] and contains 5305 actors, 2036 films and 1375 directors during 2000 to 2007. In this data actors have participated at least in 5 films. The number of queries for this data is 804.

### 5.2. Performance indices

For evaluating the performance of the measures, first we divide our data into two parts: train data and test data. Then we assess the performance in term of a number of performance indices. Various metrics have been proposed for this purpose and here we use three of them that have been frequently used in the literature: Discount Cumulative Gain (DCG), Mean Reciprocal Rank(MRR) and Precision.

(1) **nDCG@p:** This indice is used for evaluating quality of the rankings. Computation of DCG@p with query $Q$ and rank $p$ is performed as [25,26]:

$$DCG@p = rel_1(x) + \sum_{i=2}^{p} \frac{rel_i(x)}{log_2(i)} \tag{5}$$

where $rel_i$ denotes the relevance score for an object $x$ at position $i$ and which is defined in Eq. (6):

$$rel(x) = \begin{cases} 0, & \text{if } N(Q,x) = 0 \\ 1, & \text{if } N(Q,x) \neq 0 \end{cases} \tag{6}$$

where $N(Q,x) = 0$ denotes the number of papers, conferences or directors (depending on the dataset) that are common between $Q$ and $x$ in the test data. Finally, nDCG@p can be computed as:

$$nDCG@p = \frac{DCG@p}{IDCG@p} \tag{7}$$

where IDCG@p is the ideal DCG for a perfect ranking.

(2) **MRR@p:** Having a set of queries $Q$, MRR@p is defined as [27]:

$$MRR@p = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{8}$$

where $rank_i$ is the first position where there is a relevance to our query.

(3) **Precision@p:** For each query, we have [25]:

$$Precision@p = \frac{|(rel_q) \cap (ret_q)|}{p} \tag{9}$$

where $rel_q$ is real ranked objects and $ret_q$ is retrieved ranked objects similar to the query $q$.

These indices are applied on each query and then averaged over all queries to give the final performance indices.

### 5.3. Results

In this section, we represent the results of our proposed measure TSS in similarity search and compare its performance with other methods.

#### 5.3.1. Impact of time weight functions on accuracy of similarity search

In this section, we study the impact of the time weight functions in the proposed similarity measure in different meta-paths.

Figs. 3 and 4 show nDCG@p for APA (AMA) and ACA (ADA) meta-paths, respectively. Other meta-paths can be used, but we chose a few of them. It is observed that for APA meta-path, the proposed similarity metrics with an exponential weight function ranks better than the one with a linear function as well as PathSim. However, there is no single exponential function that has the best performance for all cases. For example in DBLP-1 and ACM, $f_1$ often has better performance, while $f_2$ ranks better in DBLP-2. In IMDB, which is a different dataset than the exponential functions, and for the values of $p = 15$ and $p = 20$, the proposed metric with a linear weight function results in almost a similar performance as PathSim. For ACA meta-path and in DBLP-1 and DBLP-2 datasets, $f_1$ with $\alpha = 0.5$ is often the top-performer followed by $f_2$ with $\alpha = 0.8$ (Fig. 4). In ACM dataset, $f_1$ with $\alpha = 0.8$ has the best performance. In IMDB and for ADA meta-path, the linear weight function performs better than others.

#### 5.3.2. Comparing TSS with other meta-path based similarity measures

In this section, we compare TSS with PathSim and JoinSim [14] measures that do not consider the time and Temporal measure [17] as a temporal one.

We use $p = 5$ in this experiment and the optional $\alpha$ and $\beta$. Table 2 compares the performance of the methods in term of different performance indices and meta-paths. As it is shown, TSS is the top-performer. In some cases, it improves the performance of the second-top method by more than 10%. In general, the other temporal method ranks better than the other two, which does not consider the temporal information in its calculations.

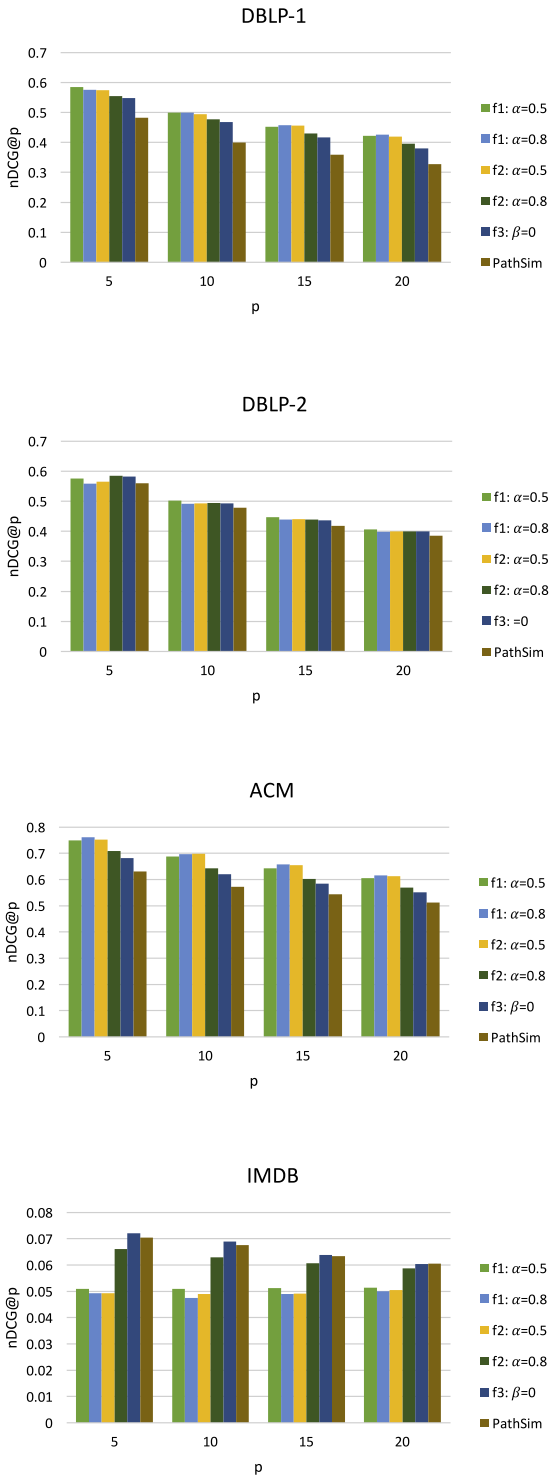


**Fig. 3.** nDCG@p for different values of p in four datasets. The results show the performance of PathSim and TSS with linear weight function $f_3$ and exponential weight functions $f_1$ and $f_2$ with different values of $\alpha$.

In IMDB dataset, accuracy values of results are very few, it is because of nature of this data. In this case, there are a lot of actors and directors, with different areas for acting, it is difficult to forecast which actor will work with which

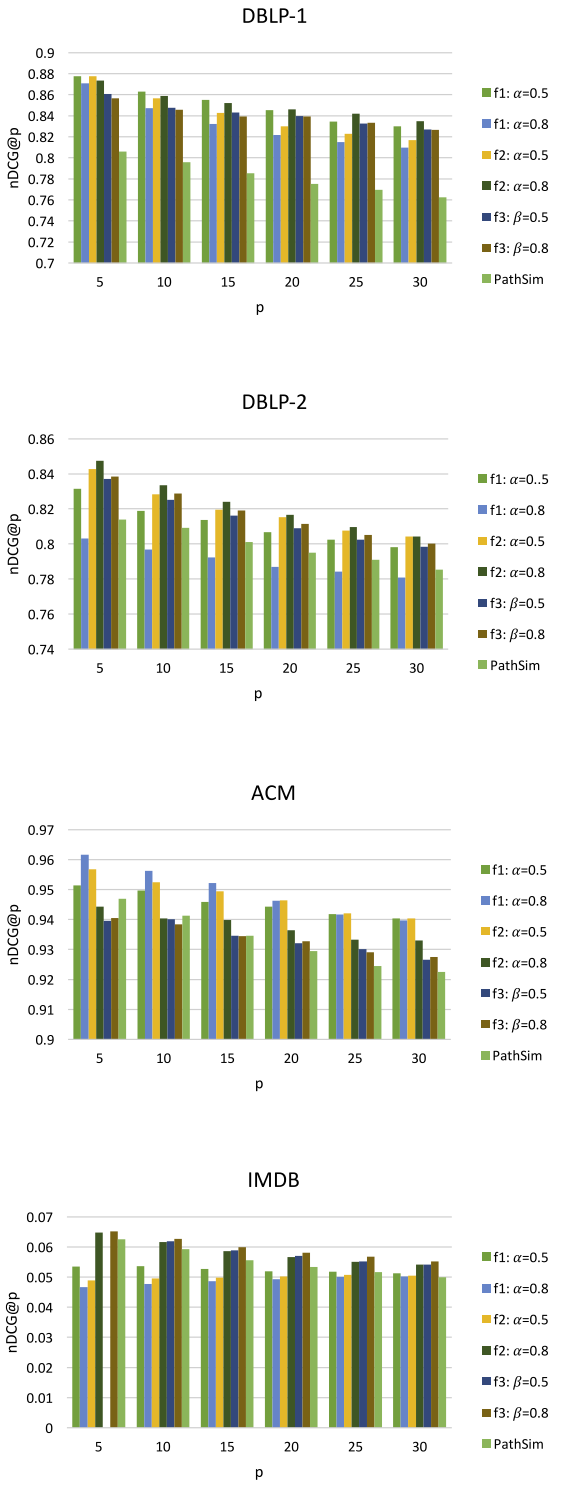


**Fig. 4.** nDCG@p for different values of p in four datasets. The results show the performance of PathSim and TSS with linear weight function $f_3$ and exponential weight functions $f_1$ and $f_2$ with different values of $\alpha$ and $\beta$.

directors. Although our measure has little improve in ranking accuracy against PathSim and JoinSim for both AMA and ADA meta-path.

**Table 2**
nDCG@5, Precision@5 and MRR@5 for different similarity measures including the proposed TASS. For different metapaths APA and ACA for DBLP-1, DBLP-2 and ACM datasets, AMA and ADA for IMDB dataset.

| Indices | Datasets | DBLP-1 | | | | DBLP-2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Metapath | Pathsim | JoinSim | Temporal | TimSim | Pathsim | JoinSim | Temporal | TimSim |
| nDCG@5 | APA | 0.4824 | 0.4266 | 0.5545 | **0.5849** | 0.5598 | 0.5491 | **0.5848** | 0.5752 |
| | ACA | 0.8060 | 0.6337 | $N/A$ | **0.8749** | 0.8139 | 0.8172 | $N/A$ | **0.8459** |
| Precision@5 | APA | 0.4521 | 0.3891 | 0.5260 | **0.5478** | 0.5254 | 0.5215 | **0.5529** | 0.5411 |
| | ACA | 0.7999 | 0.6217 | $N/A$ | **0.8673** | 0.8067 | 0.8090 | $N/A$ | **0.8445** |
| MRR@5 | APA | 0.7464 | 0.7102 | 0.7793 | **0.8164** | 0.7830 | 0.7611 | 0.7813 | **0.7904** |
| | ACA | 0.8925 | 0.7969 | $N/A$ | **0.9365** | 0.8987 | 0.9180 | $N/A$ | **0.9225** |
| | Datasets | ACM | | | | IMDB | | | |
| | Metapath | Pathsim | JoinSim | Temporal | TimSim | Pathsim | JoinSim | Temporal | TimSim |
| nDCG@5 | APA-AMA | 0.6308 | 0.6324 | 0.7094 | **0.7617** | 0.0704 | 0.0653 | 0.0660 | **0.0721** |
| | ACA-ADA | 0.9468 | 0.9412 | $N/A$ | **0.9615** | 0.05931 | 0.0482 | 0.0609 | **0.0626** |
| Precision@5 | APA-AMA | 0.61030 | 0.6061 | 0.6783 | **0.7195** | 0.0686 | 0.0629 | 0.0644 | **0.0703** |
| | ACA-ADA | 0.9440 | 0.9373 | $N/A$ | **0.9599** | 0.0581 | 0.0470 | 0.0589 | **0.0618** |
| MRR@5 | APA-AMA | 0.8305 | 0.8274 | 0.9012 | **0.9179** | 0.1052 | 0.1051 | 0.1098 | **0.1118** |
| | ACA-ADA | 0.9749 | 0.9833 | $N/A$ | **0.9835** | 0.1070 | 0.1008 | 0.1002 | **0.1136** |

## 6. Conclusion

In this manuscript, we have introduced a novel temporal measure for similarity search in heterogeneous networks. The proposed measure uses meta-path framework and accounts for interaction time. Real social networks are dynamics, and their attributes of nodes and/or edges might change over time. For example, in a heterogeneous network such as DBLP, the authors' interest or conferences' popularity might change over time. We proved that the proposed temporal similarity measures has all properties required for a similarity measure. We then tested the effectiveness of the proposed measure with a number of state-of-the-art methods including a temporal one. To this end, four datasets were chosen. Our experiments showed that having comparable computational complexity, the proposed similarity measure outperformed others in term of various performance indices. This method can be used in various application such as designing efficient recommendation systems where are the need to compute similarity between different network entities and TSS can be used to in the link prediction problem, in which the aim is to predict missing or spurious relationships in networks. For future work, we can examine other metapaths with different semantics by modeling other phenomena as networks and see effectiveness of time in them.

## Acknowledgement

This research was in part supported by a grant from IPM, Iran (No. CS1397-4-143). Mahdi Jalili is supported by Australian Research Council through project No DP170102303.

## References

[1] Molaei Soheila, Babei, Sama, Salehi, Mostafa, Jalili, Mahdi, Information spread and topic diffusion in heterogeneous information networks, Sci. Rep. 8 (2018) 9549.
[2] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu, Ianyi pathsim: meta-path-based top-k similarity search in heterogeneous information networks, in: VLDB' 11, 2011.
[3] Gregorio D'Agostino, Antonio Scala (Eds.), Networks of Networks: The Last Frontier of Complexity, Springer Berlin / Heidelberg, 2014.
[4] Wang Huijuan, Li, Qian, G.D. Agostino, Gregorio, Havlin, Shlomo, Stanley, H. Eugene, Piet Van Mieghem, Effect of the interconnected network structure on the epidemic threshold, Phys. Rev. E 88 (2) (2013) 022801.
[5] Ghariblou Saeed, Salehi, Mostafa, Magnani, Matteo, Jalili, Mahdi, Shortest paths in multiplex networks, Sci. Rep. 7 (2017) 2142.
[6] Ramezanian Rasool, Salehi, Mostafa, Magnani, Matteo, Montesi, Danilo, Diffusion of innovations over multiplex social networks, in: International Symposium on Artificial Intelligence and Signal Processing, Iran, 2015.
[7] Salehi Mostafa, Sharma, Rajesh, Marzolla, Moreno, Magnani, Matteo Magnani, Siyari, Payam, Montesi, Danilo, Spreading processes in multilayer networks, IEEE Trans. Netw. Sci. Eng. 2 (2015) 65–83.

[8] Sun Yizhou, Han, Jiawei, Mining heterogeneous information networks: A structural analysis approach, SIGKDD Explor. Newsl. 14 (2) (2013) 20–28.
[9] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, Angela Y. Wu, An optimal algorithm for approximate nearest neighbor searching fixed dimensions, J. ACM 45 (6) (1998) 891–923.
[10] Jalili Mahdi, Orouskhani, Yasin, Asgari, Milad, Alipourfard, Nazanin, Perc, Matjaž, Link prediction in multiplex online social networks, in: Royal Society Open Science, in: The Royal Society, vol. 4, 2017.
[11] Javari Amin, Jalili, Mahdi, Accurate and novel recommendations: An algorithm based on popularity forecasting, in: ACM Trans. Intell. Syst. Technol., ACM, New York, NY, USA, 2015, pp. 56:1–56:20.
[12] Shi Chuan, Kong, Xiangnan, Huang, Yue, S. Yu, Philip, Wu, Bin, Hetesim: A general framework for relevance measure in heterogeneous networks, IEEE Trans. Knowl. Data Eng. 26 (10) (2014) 2479–2492.
[13] Zhang Mingxi, Hu, Hao, He, Zhenying, Wang, Wei, Top-k similarity search in heterogeneous information networks with x-star network schema, Expert Syst. Appl. 42 (2) (2015) 699–712.
[14] Xiong Yun, Zhu, Yangyong, Yu, S. Philip, Top-k similarity join in heterogeneous information networks, IEEE Trans. Knowl. Data Eng. 27 (6) (2015) 1710–1723.
[15] Meng Xiaofeng, Shi, Chuan, Li, Yitong, Zhang, Lei, Wu, Bin, Relevance measure in large-scale heterogeneous networks, in: Lei Chen, Yan Jia, Timos Sellis, Guanfeng Liu (Eds.), Web Technologies and Applications - 16th Asia-Pacific Web Conference, APWeb 2014, Changsha, China, September 5-7, 2014. Proceedings, in: Lecture Notes in Computer Science, vol. 8709, Springer, 2014, pp. 636–643.
[16] U. Leong Hou, Yao, Kun, Mak, Hoi Fong, Pathsimext: Revisiting pathsim in heterogeneous information networks, in: Feifei Li, Guoliang Li, Seung-won Hwang, Bin Yao, Zhenjie Zhang (Eds.), Web-Age Information Management - 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings, in: Lecture Notes in Computer Science, vol. 8485, Springer, 2014, pp. 38–42.
[17] He Jiazhen, Bailey, James, Zhang, Rui, Exploiting Transitive Similarity and Temporal Dynamics for Similarity Search in Heterogeneous Information Networks, Springer International Publishing, Cham, 2014, pp. 141–155.
[18] Jeh Glen, Widom, Jennifer, Scaling personalized web search, in: Proceedings of the 12th International Conference on World Wide Web, in: WWW '03, ACM, New York, NY, USA, 2003, pp. 271–279.
[19] Jeh Glen, Widom, Jennifer, Simrank: A measure of structural-context similarity, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '02, ACM, New York, NY, USA, 2002, pp. 538–543.
[20] Tedesco Jonathan, AsymSim: Meta-Path-Based Similarity with Asymmetric Relations (Master's thesis), University of Illinois at Urbana-Champaign, 2013.
[21] Hermann Christoph, Time-based recommendations for lecture materials, in: J. Herrington, C. Montgomerie (Eds.), Proceedings of EdMedia: World Conference on Educational Media and Technology, 2010, pp. 1028–1033.
[22] Nathan N. Liu, Min Zhao, Evan Xiang, Qiang Yang, Online evolutionary collaborative filtering, in: Proceedings of the Fourth ACM Conference on Recommender Systems, in: RecSys '10, ACM, New York, NY, USA, 2010, pp. 95–102.
[23] A. Buluc, J.R. Gilbert, On the representation and multiplication of hypersparse matrices, in: 2008 IEEE International Symposium on Parallel and Distributed Processing, 2008, pp. 1–11, 1530-2075.
[24] Cantador Iván, Brusilovsky, Peter, Kuflik, Tsvi, 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011), in: Proceedings of the 5th ACM Conference on Recommender Systems, RecSys 2011, ACM, New York, NY, USA, 2011.
[25] Vuong Ba-Quy, Lim, Ee-Peng, Sun, Aixin, Le, Minh-Tam, Lauw, Hady Wirawan, Chang, Kuiyu, On ranking controversies in wikipedia: Models and evaluation, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, in: WSDM '08, ACM, New York, NY, USA, 2008, pp. 171–182.
[26] Järvelin Kalervo, Kekäläinen, Jaana, Cumulated gain-based evaluation of ir techniques, ACM Trans. Inf. Syst. 20 (4) (2002) 422–446.
[27] Xu Ying, Gao Zhiqiang, Wilson Campbell, Zhang Zhizheng, Zhu Man, Ji Qiu, in: Lin Xuemin, Manolopoulos Yannis, Srivastava Divesh, Huang Guangyan (Eds.), Web Information Systems Engineering – WISE 2013: 14th International Conference, first ed., in: Lecture Notes in Computer Science, vol. 8180, Springer-Verlag Berlin Heidelberg, Nanjing, China, 2013.