World Wide Web (2018) 21:1575–1600
https://doi.org/10.1007/s11280-017-0510-9

# Durable relationship prediction and description using a large dynamic graph

**Ruili Wang**[1] · **Wanting Ji**[1] · **Baoyan Song**[2]

Received: 30 June 2017 / Revised: 27 September 2017 / Accepted: 25 October 2017 /
Published online: 25 November 2017


**Abstract** Dynamic graphs are a data structure widely used in representing changeable relationships or connections between different entities. This paper proposes a novel type of node similarity, based on the frequency of connections between nodes to describe the changeable relationships between entities over a period; this has not been considered before as an indication of similarity between two nodes. In other words, if two entities have a history of frequent connections, this means that they have something in common and have a durable relationship. In this paper, durable relationships describe the frequency of connections rather than only the continuous connection between two nodes. Thus, durable relationships are defined in two dimensions: (i) In the dimension of time, they can be categorized based on the length of duration as short-term, medium-term, or long-term relationships; (ii) Based on frequencies of connections over a period, they can be categorized into four statuses (No Relationship, Weak Relationship, In Relationship, and Strong Relationship). Based on this definition of durable relationships, a node similarity measurement algorithm is proposed, to study the status of relationships from a longitudinal study point of view. This method provides a new way to describe the semantics of relationships (such as collaborative relationships, or customer loyalty descriptions) and also gives a practical application of node similarity measurement in a real world, which is to provide a prediction of relationship. Our extensive

This article belongs to the Topical Collection: *Special Issue on Deep Mining Big Social Data*
Guest Editors: Xiaofeng Zhu, Gerard Sanroma, Jilian Zhang, and Brent C. Munsell

✉ Ruili Wang
Ruili.Wang@massey.ac.nz

Wanting Ji
jwt@escience.cn

Baoyan Song
bysong@lnu.edu.cn

[1] Institute of Natural and Mathematical Sciences, Massey University, Auckland 0632, New Zealand

[2] School of Information, Liaoning University, Shenyang 110036, China

experiments have shown that the proposed method can effectively describe durable relationships and especially predict future relationships.

**Keywords** dynamic graph · durable relationship · node similarity · relationship prediction · time snapshot

## 1 Introduction

Graph is a data structure widely used in representing social networks and other practical applications [12, 17, 21, 27, 28]. In general, many applications use a graph to represent entities (corresponding to nodes in the graph) and their dependency relationships (corresponding to edges in the graph). Using this graphical representation, practical applications can be processed easily in a mathematical manner.

For example, the cooperative relationships among authors in DBLP (Digital Bibliography & Library Project (http://dblp.uni-trier.de/)), which is a computer science bibliography website, listing computer science references in a chronological order, can be abstracted into a graph. The nodes in the graph can represent authors listed in DBLP, while the edges connecting two nodes represent a co-authoring relationship between these two authors (i.e., they have co-authored one or more publications together). In addition, some complex networks such as knowledge discovery and data mining [29], citation networks [14], and traffic networks [10] can be modeled using graphical structures.

In real world, relationships between different entities can change overtime, and thus a dynamic graph is used to represent such changes accordingly [1, 5, 8, 11, 23, 25, 26]. For instance, an author can have different co-authors in different years. There may be many reasons for this, such as research interests evolving over a period, or environmental reasons (e.g., moving to a new job) for either the author or co-authors. Thus, a graph showing this author and a set of co-authors in the DBLP co-authorship network for 1 year, corresponding to the snapshot of the graph of this year may be different from the snapshot of the graph of the following year. Such a graph becomes a dynamic graph.

In this paper, we intend to use "durable relationship" to describe this kind of changeable but reliable relationships in a dynamic graph over a long period. A connection (i.e., an edge) between two nodes does not always exist over a period in a dynamic graph. Thus, the connection of the two nodes may not be shown in every snapshot. However, if the connections are shown over a long period frequently, this means there is a durable relationship between the two nodes. Therefore, durable relationships describe the frequency of connections in practice rather than the connection of two nodes in every snapshot. With a massive amount of data (e.g., DBLP) accumulated, how to describe durable relationships between entities over a long period in a dynamic graph is proposed in this paper.

In addition, this paper aims to predict the evolution of a relationship in the future. For example, according to the co-authorship between two authors in the past several years, can we predict if they will continue their co-authorship in the following few years? A reliable prediction can be given by our proposed method to measure the possibility that a connection will remain in the future.

A novel type of node similarity based on the frequency of connections between nodes (corresponding to entities) is proposed in this paper to describe durable relationships between entities over a period, which has not been considered as an indication of similarity between

nodes before. In other words, if two entities have a history of frequent connections, this means that they have something in common and have a durable relationship. Thus, durable relationships can be defined in two dimensions: (i) In the dimension of time, they can be categorized based on the length of duration as short-term, medium-term, or long-term relationships; (ii) Based on frequencies of connections over a period, they can be categorized into four statuses (No Relationship, Weak Relationship, In Relationship, and Strong Relationship). The durable relationships can represent changeable but reliable relationships.

The relationship description method proposed in this paper represents a durable relationship in three different durations of time: (i) For a short period, there are four possible short-term relationships (i.e., short-term No Relationship, short-term Weak Relationship, short-term In Relationship, or short-term Strong Relationship), which can be described by the connection statuses shown in two consecutive snapshots of a dynamic graph; (ii) For a medium period, there are four possible medium-term relationships (i.e., medium-term No Relationship, medium-term Weak Relationship, medium-term In Relationship, or medium-term Strong Relationship), which can be described by the connection statuses shown in three consecutive snapshots of a dynamic graph; (iii) For a long period, there are also four possible relationships (i.e., long-term No Relationship, long-term Weak Relationship, long-term In Relationship, or long-term Strong Relationship), which can be described by the connection statuses shown in more than three consecutive snapshots of a dynamic graph. In this paper, we use six consecutive snapshots to describe the durable collaborative relationships in DBLP. This number of consecutive snapshots is viable depending on an application.

A short-term relationship (short-term Weak Relationship, short-term In Relationship, or short-term Strong Relationship) can show the status change and can show a causal relationship. A medium-term relationship (medium-term Weak Relationship, medium-term In Relationship, or medium-term Strong Relationship) can be an indicator of the potential to become a long-term relationship. A long-term relationship is to describe a sustained durable relationship in a longitudinal study.

In this paper, a novel node similarity measurement algorithm is proposed to measure durable relationships in a dynamic graph. By using the definition given above, the proposed algorithm searches all connections/interactions to a node over a period. For instance, when describing the relationship between two authors over a period, it can be translated into measuring the similarity between these two authors during this period. The relationship of these two authors can be estimated by two questions: (i) Connections: Have they co-authored papers together? (ii) Frequency of connections: How often do they co-author papers during a period? If they have co-authored papers frequently during a period, we can say that they have similar research interests or they have a durable relationship.

Such connections can be in any forms and between any two entities, not necessary in social networks. Here are some examples: How often does a customer visit a particular website? How often does a customer buy a product from a shop? How often do two students select the same courses?

We can see that the proposed method describes a new semantics of relationships and it can have practical significance when it is implemented in the real world. Besides processing the DBLP co-authorship network, it also can be used in other relationship analysis, such as customer relationship analysis [4], personal relationship social networks [13], Such as between people or organizations in social media.

For instance, this kind of durable relationship description can be used to describe the degrees of customer loyalty, which can be divided into four levels: cognitive loyalty, affective loyalty, and conative loyalty [30]. Initially, loyalty between a customer and a brand belongs to No Relationship before a loyalty is established. After that, cognitive loyalty can be defined as a Weak Relationship, which is very susceptible to external factors and an inferior relationship between a customer and a brand; affective loyalty can be known as a In Relationship which is an affective relationship between a customer and a brand; conative loyalty can be viewed as a Strong Relationship, which is based on the customers' understanding about product features, cost-effective factors and other specific information.

In this paper, we analyze actual relationships between DBLP authors as an example. This gives an example of a practical meaning of a durable relationship description for a longitudinal study. The novel type of node similarity based on the frequency of connections between two DBLP authors over a period is proposed to describe their durable co-authorship. Based on this node similarity, we will predict the evolution of this durable relationship in the future. In other words, our proposed method describes a current relationship between two DBLP authors and predicts whether this co-authorship will remain in the future.

The rest of this paper is organized as follows. Section 1 introduces the background and significance of this research. Section 2 describes related work such as existing similar nodes query methods and analyzes the applicable conditions of each existing method. Some useful concepts addressed in this paper are provided in Section 3. Section 4 proposes our method. The experimental analysis is described in Section 5. Section 6 provides the conclusions, which summarizes the research and shows that the proposed algorithm is effective and feasible.

## 2 Related works

Measuring node similarity has been intensively investigated in graph processing, which has been used in preferences query [9], community discovery [18], information retrieval [2], collaborative filtering [20], and other fields. The dominant algorithms can be divided into two categories: one is based on attributes; the other one is based on structures.

### 2.1 Key methods based on attributes

Node attributes are widely adopted as node similarity measures in many data mining applications. Existing algorithms based on attributes calculate node similarity by comparing nodes' attributes or their neighborhoods' attributes. In general, it constructs a series of attribute vectors for each node, which treats each attribute of a node as one dimension in a multi-dimensional space, to map a node to an n-dimensional space. The Euclidean distance [6] is commonly used directly or indirectly when using this kind of method to measure node similarity, as in the following:

$$dist(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{1}$$

where $n$ is the number of attributes of both node $a$ and node $b$. When the number of a node's attributes $n$ is very large, requiring mapping these attributes into a high-dimensional space, the computation cost will become very high. In addition, there are some attributes' which have

unknown or incomplete values in some cases. Thus, it is impossible to map all nodes completely to an n-dimensional space, which is a limitation of this kind of method [3].

Another dominant method is the cosine similarity [19], which maps each node to a multivariate vector by using an evaluation function. An evaluation function is used to estimate the importance of an attribute, and map every node to a multivariate vector. When one component of this multivariate vector cannot be estimated, setting this component with the default value 0, this can overcome the limitation of the Euclidean distance. Thus, the node similarity of two nodes can be measured by calculating the cosine of their vectors:

$$\cos(a,b) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{2}$$

where $n$ is the number of attributes of both node $a$ and node $b$.

Besides, Jaccard similarity coefficient [15] is another common method, which is used to measure similarity between two nodes. It is defined as the number of the common attributes of both nodes divided by the number of attributes of both nodes in total. The formal definition is as follows:

$$J(a,b) = \frac{\sum_{i=1}^{n} \min(a_i, b_i)}{\sum_{i=1}^{n} \max(a_i, b_i)} \tag{3}$$

where $n$ is the number of attributes of both node $a$ and node $b$.

### 2.2 Key methods based on connections

Current algorithms based on connections measure node similarity by analyzing the connections between two nodes. Dominant methods include: Dice's similarity coefficient [7], and SimRank similarity [16].

Normally, the inner product of two vectors is used to measure the similarity between two nodes. For example, as shown in the formula of cosine similarity, the similarity between node $a$ and node $b$ is proportional to the cosine of vector $a$ and vector $b$. The larger the cosine similarity score, the more similar these two nodes are. However, this kind of calculation method has some limitations. For example, it cannot reflect similarity properly when the attributes have different scales of values since the cosine similarity will treat all attributes equally. Thus, Dice's similarity coefficient [7] has been proposed to calculate the similarity between nodes whose attributes have different scales of values. Dice's similarity coefficient uses an arithmetic mean rather than a geometric mean when computing vectors of nodes:

$$Dice(a,b) = \frac{2\sum_{i=1}^{n} a_i \cdot b_i}{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2} \tag{4}$$

where $n$ is the number of attributes of both node $a$ and node $b$.

SimRank [16] measures the similarity between two nodes based on the topology of a graph. The notion is that if two nodes in a graph have similar neighbor structures, then these

two nodes are similar to each other. In other words, other nodes that are known to be directly connected to these two nodes are similar, thus these two nodes are similar to each other. This method uses known similarity between nodes to determine the similarity of other nodes which are connected to these nodes directly. Because the SimRank algorithm is defined in a recursive way, the initial condition is that each node is most similar to itself. The formal definition is as follows:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s\left(I_i(a), I_j(b)\right) \tag{5}$$

where $C$ is an attenuation factor, the value range of which is [0, 1]; $I(a)$ is a collection of the edges that connect to node $a$; and $s(a, b)$ donates the similarity between node $a$ and node $b$, the value range of which is [0, 1]. Since SimRank is defined in recursion, if we want to calculate the similarity between node $a$ and node $b$, we need to know the similarity between node $c$ and node $d$, while node $c$ connects to node $a$ and node $d$ connects node $b$, respectively. In some cases, this will cause difficulties in measuring node similarity.

### 2.3 Key methods based on the combination of attribute and connection similarity

Recently, a combination of both an attribute-based similarity and a connection-based similarity to get a combined similarity has been proposed [12, 21]. The combined similarity of two nodes in [12, 21] was achieved by the weighted sum of both the attribute-based similarity and the connection-based similarity.

For attribute-based similarity, reference [21] provided each node with a set of keywords to represent its attributes. If a node has the $i^{\text{th}}$ keyword in its keyword set, then the $i^{\text{th}}$ vector of this node is set to 1. Otherwise, it will be 0. After that, the cosine similarity is used to compute node similarity. A higher value of the cosine similarity indicates that two nodes are more similar because they have more common node attributes/keywords, regardless of whether they have any connection. This approach is used to investigate complex community detection [21].

Similarly, reference [12] used this kind of attribute-based similarity on a citation network, which can be seen as a super-graph. In a super-graph, a publication can be represented by a super-node and an edge denotes a citation relationship between two publications. In a super-node that is a publication, nodes of the super node are the key words of the publication. A key word represents one attribute of a super-node. For a super-graph, each super-node contains a set of attributes. Accordingly, the cosine distance is used to measure the super-node attribute similarity in [12].

Generally, a connection-based similarity indicates that if two nodes in a graph have many common neighbors, they are likely to be similar to each other. Based on this, [21] proposed a definition about topology-based node similarity as follows. Similar to Dice's similarity coefficient, it is a ratio of the number of common neighbors to the number of all neighbors between two nodes. A high number of common neighbors indicate that these two nodes are likely to belong to the same community.

For structure-based node similarity computation between two super-nodes, reference [12] proposed using a random walk-based distance measure. Given a pair of super-nodes, if they have more the same nodes (i.e. keywords), these two super nodes are more similar.

Finally, the final similarity is calculated by the weighted sum of the attribute similarity and connection similarity.

## 3 Preliminaries

In this section, some useful concepts in a dynamic graph and the problem descriptions addressed in this paper are discussed below.

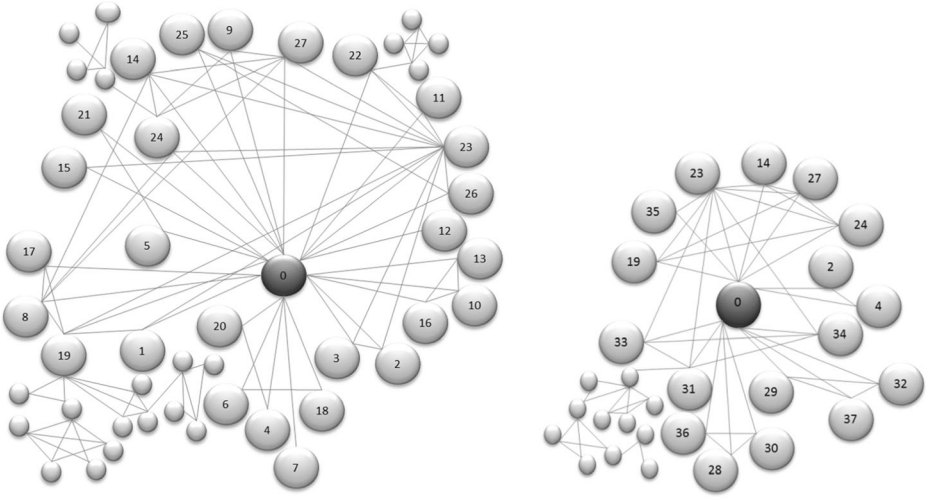
Let $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$ records the evolution of a dynamic graph $G$ during the period $T=[t_1, t_n]$, where $V_i$ is the set of all nodes (i.e., entities) in graph $G$ at time $t_i$, and $E_i$ is the set of all edges (i.e., connections between entities) in graph $G$ at time $t_i$. In other words, this dynamic graph is collected by a series of sequentially and non-overlapping time snapshots [22].

The DBLP co-authorship network can be seen as a large dynamic graph, where nodes represent the authors listed in DBLP and edges represent the co-authorships between two authors. Figure 1 uses a series of time snapshots to show the evolution of the co-authorship network over 2 years, 2006 and 2007, respectively. Due to limited space, only a few of them are shown in Figure 1, where Figure 1a is the time snapshot of 2006, and Figure 1b is the time snapshot of 2007. The set of time snapshots shows the evolution of dynamic graph $G$. Different time snapshots are mutually exclusive and independent. The relationships shown in one snapshot have no cause and effect of the other snapshots. The correspondence between all authors' names and the number is listed in Table 1.

As shown in Figure 1, node $v_0$ (the author's name is Shichao Zhang) is a subject node as an example of durable relationship description in this paper. Our method only retains node $v_0$ and the nodes and edges which connect to node $v_0$ directly. We have six snapshots over 6 years (2006-2011), which are shown in Figure 2. Table 1 listed the correspondence between all authors' names and the number used in Figure 2.

### 3.1 Short-term relationships

The connection between two nodes does not always exist over a period in a dynamic graph. A time snapshot is to collect all connections between nodes exist at a particular time. Changing




(a) DBLP snapshot in 2006 (b) DBLP snapshot in 2007

**Figure 1** Evolution of a DBLP co-authorship network

**Table 1** The list of node numbers and their corresponing authors' names in Figure 2

| Author's name | No. | Author's name | No. | Author's name | No. | Author's name | No. | Author's name | No. |
|---|---|---|---|---|---|---|---|---|---|
| K. Cai | 28 | H. Huan | 8 | W. Qian | 45 | Z. Xu | 20 | X. Zhang | 61 |
| J. Cao | 52 | F. Huang | 7 | Y. Qin | 14 | Z. Xu | 33 | L. Zhao | 25 |
| M. Chang | 65 | Z. Huang | 9 | Z. Qin | 15 | B. Yan | 34 | Y. Zhao | 26 |
| F. Chen | 1 | Z. Jin | 38 | S. Sheng | 16 | X. Yan | 35 | J. Zhao | 49 |
| Q. Chen | 2 | X. Li | 68 | W. Shu | 46 | J. Yang | 21 | X. Zhao | 50 |
| Y. Chen | 3 | T. Liang | 56 | W. Song | 31 | Q. Yang | 47 | Z. Zheng | 36 |
| D. Cheng | 71 | C. X. Ling | 10 | K. Su | 17 | X. You | 42 | Z. Zheng | 62 |
| Y. Dai | 55 | C. Liu | 30 | Y. Sun | 57 | X. Yu | 22 | Z. Zhong | 54 |
| L. Davidson | 29 | L. Liu | 39 | K. Sun | 74 | J. Yu | 75 | X. Zhu | 27 |
| Z. Deng | 72 | H. Liu | 43 | Z. Tang | 58 | D. Yuan | 48 | X. Zhu | 37 |
| W. Ding | 66 | Y. Liu | 73 | R. Wang | 18 | C. Yuan | 60 | M. Zhu | 51 |
| Y. Dong | 4 | J. Lu | 11 | T. Wang | 40 | M. J. Zaki | 12 | Y. Zhu | 63 |
| A. Ekart | 64 | X. Luo | 69 | R. Wang | 41 | S. Zhang | 0 | Y. Zhu | 76 |
| N. Gu | 5 | T. M. Khoshgoftaar | 32 | X. Wu | 19 | C. Zhang | 23 | M. Zong | 77 |
| Y. Guo | 6 | Y. Mo | 44 | J. Wu | 59 | J. Zhang | 24 | | |
| M. He | 67 | A. Ni | 13 | W. Wu | 70 | Z. Zhang | 53 | | |

relationships are described by the changes of the connection between two nodes in several time snapshots.

The smallest measure unit of a relationship in this paper is called a short-term relationship (also called as causal relationship), which is used to describe a changeable connection in a short term (i.e., two consecutive time snapshots equivalent to consecutive 2 years). Thus, a short-term relationship is defined as: Given dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, node $v_j^i$ and node $v_k^i$ are two nodes in the snapshot of $DG$ at time $t_i$ (i.e., $v_j^i, v_k^i \in V_i$), node $v_j^{i+1}$ and node $v_k^{i+1}$ are the two nodes in the snapshot of $DG$ at time $t_{i+1}$ (i.e., $v_j^{i+1}, v_k^{i+1} \in V_{i+1}$), a short-term relationship of node $v_j$ and node $v_k$ at time $t_{i+1}$ is defined by the connection between node $v_j$ and node $v_k$ in the time snapshots at time $t_i$ and time $t_{i+1}$.

**Definition 1 (short-term No Relationship)** Given dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes never appear in both the $i^{\text{th}}$ and $i+1^{\text{th}}$ time snapshots, the relationship between node $v_j$ and node $v_k$ at time $t_{i+1}$ is a short-term No Relationship.

As shown in Figure 2a and b, there are no connections between node $v_0$ and node $v_{38}$. Thus, the relationship between $v_0$ and $v_{38}$ in 2007 is a short-term No Relationship.

**Definition 2 (short-term Weak Relationship)** Given dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes existing in the $i^{\text{th}}$ time snapshot but not existing in the $i+1^{\text{th}}$ time snapshot, the relationship between node $v_j$ and node $v_k$ at time $t_{i+1}$ has a short-term Weak Relationship.

As shown in Figure 2a and b, node $v_0$ and node $v_{13}$ have co-authored a paper in 2006, but did not work together in 2007. Thus, the relationship between $v_0$ and $v_{13}$ in 2007 has a short-term Weak Relationship.

**Definition 3 (short-term In Relationship)** Given dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of

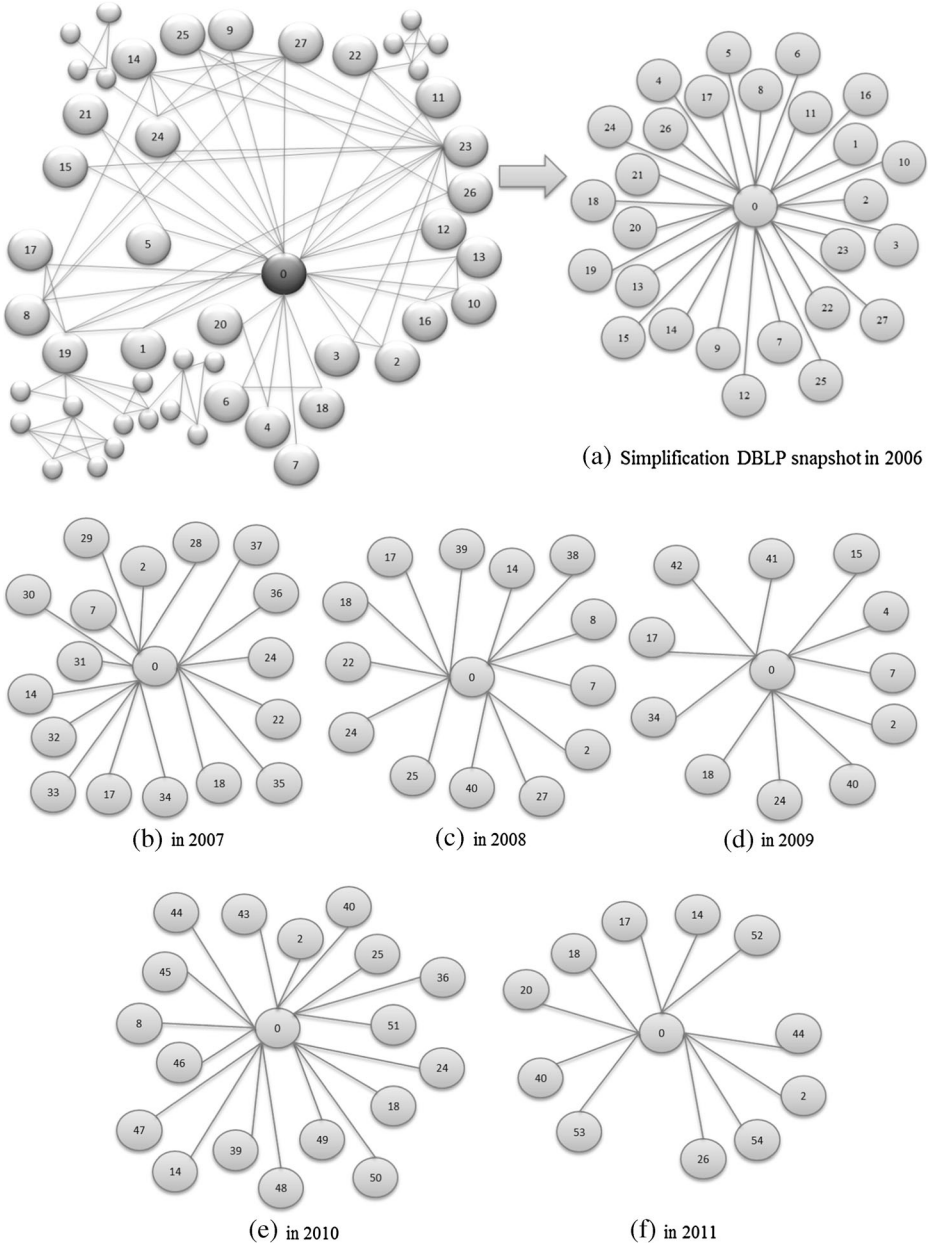



**Figure 2** Evolution relationship between the subject node $v_o$ and its subject-related nodes over 6 years (2006-2011)

nodes not existing in the $i^{\text{th}}$ time snapshot but existing in the $i+1^{\text{th}}$ time snapshot, the relationship between node $v_j$ and node $v_k$ at time $t_{i+1}$ is a short-term In Relationship.

As shown in Figure 2a and b, there is no connection between node $v_0$ and node $v_{30}$ in 2006. However, they have co-authored a paper(s) in 2007. Thus, the relationship between $v_0$ and $v_{30}$ in 2007 is a short-term In Relationship.

**Definition 4 (short-term Strong Relationship)** Given dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes exist in both the $i^{th}$ and $i+1^{th}$ time snapshots, the relationship between node $v_j$ and node $v_k$ at time $t_{i+1}$ is a short-term Strong Relationship.

As shown in Figure 2a and b, the connection between node $v_0$ and node $v_{19}$ existed in both these 2 years. Thus, the relationship between $v_0$ and $v_{19}$ in 2007 is a short-term Strong Relationship.

### 3.2 The 0-1 representations

In a dynamic graph, a durable relationship is used to describe the changeable connections between two nodes over a period. We use 0 (unconnected) or 1 (connected) to represent the connection status between a pair of nodes. For example, when describing the connection of node $v_0$ and node $v_k$, where node $v_0$ is a subject node, and node $v_k$ is one of subject-related nodes (i.e., have connections with subject node $v_0$), if there is a connection between node $v_k$ and subject node $v_0$ at the $i^{th}$ time snapshot, then the connection status value of node $v_k^i$ is set to 1, otherwise is set to 0.

The 0-1 representation indicates the status of connection between two nodes. According to the above definitions, the statues of short-term relationships can be represented as sequence of connection statuses: (i) A short-term No Relationship can be represented as 00; (ii) A short-term Weak Relationship can be represented as 10; (iii) A short-term In Relationship can be represented as 01; (iv) A short-term Strong Relationship can be represented as 11.

### 3.3 Medium-term relationships

A medium-term relationship can be an indicator of the potential to become a long-term relationship. Therefore, we provide a series of definitions for a relationship in a medium term, which is described by the connection statues shown in three consecutive snapshots of a dynamic graph in this paper.

**Definition 5 (medium-term No Relationship)** Given dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes never appear in the $i^{th}$, $i+1^{th}$ and $i+2^{th}$ time snapshots, the relationship between node $v_j$ and node $v_k$ at time $t_{i+2}$ is a medium-term No Relationship. The connection between node $v_j$ and node $v_k$ can be represented as 000.

As shown in Figure 2a-c, for the subject node $v_0$, the status of node $v_{38}$ over this 3-year period can be represented as a sequence of connection statuses of 000. Thus, the relationship between $v_0$ and $v_{38}$ in 2008 is a medium-term No Relationship.

**Definition 6 (medium-term Weak Relationship)** Given a dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes only existing one of three snapshots (the $i^{th}$, $i+1^{th}$ and $i+2^{th}$ time snapshots), the relationship between node $v_j$ and node $v_k$ at time $t_{i+2}$ is medium-term Weak Relationship. The connection between node $v_j$ and node $v_k$ can be represented as 100, 010 or 001.

As shown in Figure 2a-c, for the subject node $v_0$, the status of node $v_{13}$ over this 3-year period can be represented as a sequence of 100. Thus, the relationship between $v_0$ and $v_{13}$ in 2008 is a medium-term Weak Relationship.

**Definition 7 (medium-term In Relationship)** Given a dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes only not existing in one of three snapshots (the $i^{th}$, $i+1^{th}$ and $i+2^{th}$ time snapshots), the relationship between node $v_j$ and node $v_k$ at time $t_{i+2}$ is a medium-term In Relationship. The connection between node $v_j$ and node $v_k$ can be represented as 110, 101, or 011.

As shown in Figure 2b-d, for the subject node $v_0$, the status of node $v_{34}$ over this 3-year period can be represented as a sequence of 101. Thus, the relationship between $v_0$ and $v_{34}$ in 2009 is a medium-term In Relationship.

**Definition 8 (medium-term Strong Relationship)** Given a dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, if the connection between this pair of nodes still existing in the $i^{th}$, $i+1^{th}$ and $i+2^{th}$ time snapshots, the relationship between node $v_j$ and node $v_k$ at time $t_{i+2}$ is a medium-term Strong Relationship. The connection between node $v_j$ and node $v_k$ is 111.

As shown in Figure 2a-c, for the subject node $v_0$, the statuses of node $v_{19}$ over this 3-year period can be represented as a sequence of 111. Thus, the relationship between $v_0$ and $v_{19}$ in 2008 is a medium-term Strong Relationship.

## 4 Durable relationship descriptions

A novel type of node similarity based on the frequency of connections between two nodes (i.e., entities) is proposed in this paper to describe a relationship between entities over a period, which has not been considered as an indication of similarity between the two nodes before. Our proposed method defines node similarity in three levels: short-term node similarity, medium-term node similarity, and long-term node similarity.

### 4.1 Short-term relationship descriptions

Given a dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, where node $v_j$ is a subject node and node $v_k$ is a subject-related node of node $v_j$, the short-term node similarity between node $v_j$ and node $v_k$ can be defined as:

$$short\text{–}sim\left(v_j^i, v_k^i\right) = \frac{\left(v_j^{i-1}, v_j^i\right) \cdot \left(v_k^{i-1}, v_k^i\right)^T}{\left\| \left(v_j^{i-1}, v_j^i\right) \right\| \left\| \left(v_k^{i-1}, v_k^i\right) \right\|} \tag{6}$$

where $(v_j^{i-1}, v_j^i)$ is the 0-1 representation of subject node $v_j$, and $(v_k^{i-1}, v_k^i)$ is the 0-1 representation of subject-related node $v_k$.

It is very similar to the calculation of cosine similarity. In other words, if the connections between these two nodes are longer and more frequency, these two nodes are more similar to each other.

Table 2 listed the similarities of different 0-1 sequences to describe different types of short-term relationship between two nodes.

### 4.2 Medium-term relationship descriptions

For medium-term relationships, when describing the relationship/similarity between a subject author and subject-related authors over a medium period, a subject-related author who has more frequently co-authored papers with the subject author is more similar (i.e., more strong relationship) to the subject author. At the same time, if a co-authorship between the subject author and this subject-related author is more close to the current time, the co-authorship indicates that the subject related author is more similar (i.e., more strong relationship) to the subject author's current research interests. The connection between two authors happened in different times will have different weights on similarity.

Therefore, the weight of node $v_j$ at time $t_i$ can be defined as:

$$w_j^i = \sqrt{\frac{i}{n!}} \tag{7}$$

where $n$ is the number of snapshots over a medium period (i.e., $n = 3$), and $i$ is the $i^{\text{th}}$ snapshot at time $t_i$ in a medium-term relationship.

Because of the medium-term relationship between a pair of nodes has four statuses: a medium-term No Relationship (which can be represented as a sequence of 000), a medium-term Weak Relationship (i.e., the sequence is 001, 010, or 100), a medium-term In Relationship (i.e., the sequence is 110, 101, or 011), or a medium-term Strong Relationship (i.e. the sequence is 111). A medium-term relationship can be seen as a continuation of a short-term relationship. Therefore, for a dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and node $v_j$ and node $v_k$ are two nodes in $DG$, where node $v_j$ is a subject node and node $v_k$ is a subject-related node of node $v_j$, the medium-term node similarity of $DG$ can be defined as:

$$medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sum_{i=1}^{n} w_j^i v_j^i w_k^i v_k^i}{\sqrt{\sum_{i=1}^{n}\left(w_j^i v_j^i\right)^2}\sqrt{\sum_{i=1}^{n}\left(w_k^i v_k^i\right)^2}} \tag{8}$$

where $w_j^i$ and $w_k^i$ are the weights of the connection between node $v_j$ and node $v_k$ at time $t_i$, respectively, and $n$ is the duration of medium term (i.e., $n = 3$). In other words, if the

**Table 2** The similarity values of short-term relationships

| Short-term relationship | 0-1 sequence of subject-related node | Short-term node similarity |
|---|---|---|
| Short-term No Relationship | 00 | $short{-}sim\left(v_j^i, v_k^i\right) = \frac{(1,1)\cdot(0,0)^T}{\|(1,1)\|\|(0,0)\|} = 0$ |
| Short-term Weak Relationship | 01 | $short{-}sim\left(v_j^i, v_k^i\right) = \frac{(1,1)\cdot(0,1)^T}{\|(1,1)\|\|(0,1)\|} = \frac{\sqrt{2}}{2}$ |
| Short-term In Relationship | 10 | $short{-}sim\left(v_j^i, v_k^i\right) = \frac{(1,1)\cdot(1,0)^T}{\|(1,1)\|\|(1,0)\|} = \frac{\sqrt{2}}{2}$ |
| Short-term Strong Relationship | 11 | $short{-}sim\left(v_j^i, v_k^i\right) = \frac{(1,1)\cdot(1,1)^T}{\|(1,1)\|\|(1,1)\|} = 1$ |

connections between node $v_j$ and node $v_k$ are more frequent and more recent over a medium term time, these two nodes are more similar to each other. Table 3 listed different values of similarity to describe different types of medium-term relationships between two nodes.

### 4.3 Long-term relationship descriptions

Similar to the medium-term relationship description, the long-term relationship description is also based on the frequency of connections over a long period. In this paper, a durable relationship is described by six time snapshots, which can be also represented as a 0-1 sequence with six digits. The number of snapshots for a long term is selected six for this DBLP application for the following reasons: (i) It is three-time long of a short term; (ii) It is two-time long of a medium term; (iii) It is common that authors can collaborate more than 6 years. The number of snapshots for a long term can be either nine or twelve.

Thus, the four statuses of long-term relationships can be described as: (i) For a long-term No Relationship, there are no 1s in the six-digit sequence; (ii) For a long-term Weak Relationship, there is only one 1 or two 1s in the six-digit sequence; (iii) For a long-term In Relationship, there is three or four 1s in the six-digit sequence; (iv) For a long-term Strong Relationship, there is five or six 1s in the six-digit sequence. The calculation of long-term node similarity (i.e., a long-term relationship) can be seen as a continuation of a medium-term relationship.

### 4.4 Sliding window in relationship descriptions

In a dynamic graph, a durable relationship is used to describe the changeable connections between two nodes over a long period. It is to describe a sustained relationship in practice rather than the connection of two nodes in each snapshot. With a massive amount of changeable data (e.g., DBLP) accumulated over the time, a long-term relationship is to describe durable relationships from a longitudinal study point of view.

For example, as shown in Figure 2, the statuses of the connection between the subject node $v_0$ and subject-related node $v_{24}$ over 6 years (2006-2011) can be represented as a sequence of 111100. It can be seen as an example for a long-term In Relationship. When describing the changes happened during this period, compared with medium-term relationships, more snapshots (i.e., the statuses of connections) need to be considered, which will have more variable than medium-term relationships. Thus, we use a medium-term relationship as an evaluate unit for the durable relationship description.

The evaluate unit is implemented by a sliding window. Setting the sliding window size to 3 (i.e., the same length as a medium term), while the sliding step is one snapshot. Whenever the window has 3 snapshots, we calculate the node similarity by using the proposed formula above. Then the window slides one step.

### 4.5 Algorithm

In this section, according to the definitions above, we propose an algorithm for node similarity, which can be used for durable relationship descriptions.

Given a dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, and a subject node and its subject-related nodes are the nodes in $DG$. Firstly, we collect this dynamic graph $DG$ with $n$ time snapshots and select all subject-related nodes in each snapshot related to the subject node.

**Table 3** The value of medium-term relationship

| Medium-term relationship | 0-1 sequence of subject-related node | Medium-term node similarity |
|---|---|---|
| Medium-term No Relationship | 000 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*0+\sqrt{2/6}*1*\sqrt{2/6}*0+\sqrt{3/6}*1*\sqrt{3/6}*0}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*0\right)^2+\left(\sqrt{2/6}*0\right)^2+\left(\sqrt{3/6}*0\right)^2}} = 0$ |
| Medium-term Weak Relationship | 100 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*1+\sqrt{2/6}*1*\sqrt{2/6}*0+\sqrt{3/6}*1*\sqrt{3/6}*0}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*0\right)^2+\left(\sqrt{3/6}*0\right)^2}} = 0.408$ |
| | 010 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*0+\sqrt{2/6}*1*\sqrt{2/6}*1+\sqrt{3/6}*1*\sqrt{3/6}*0}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*0\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*0\right)^2}} = 0.577$ |
| | 001 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*0+\sqrt{2/6}*1*\sqrt{2/6}*0+\sqrt{3/6}*1*\sqrt{3/6}*1}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*0\right)^2+\left(\sqrt{2/6}*0\right)^2+\left(\sqrt{3/6}*1\right)^2}} = 0.707$ |
| Medium-term In Relationship | 110 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*1+\sqrt{2/6}*1*\sqrt{2/6}*1+\sqrt{3/6}*1*\sqrt{3/6}*0}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*0\right)^2}} = 0.707$ |
| | 101 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*1+\sqrt{2/6}*1*\sqrt{2/6}*0+\sqrt{3/6}*1*\sqrt{3/6}*1}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*0\right)^2+\left(\sqrt{3/6}*1\right)^2}} = 0.816$ |
| | 011 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*0+\sqrt{2/6}*1*\sqrt{2/6}*1+\sqrt{3/6}*1*\sqrt{3/6}*1}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*0\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}} = 0.913$ |
| Medium-term Strong Relationship | 111 | $medium{-}sim\left(v_j^i, v_k^i\right) = \frac{\sqrt{1/6}*1*\sqrt{1/6}*1+\sqrt{2/6}*1*\sqrt{2/6}*1+\sqrt{3/6}*1*\sqrt{3/6}*1}{\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}\sqrt{\left(\sqrt{1/6}*1\right)^2+\left(\sqrt{2/6}*1\right)^2+\left(\sqrt{3/6}*1\right)^2}} = 1$ |

Next, using the 0-1 representation to represent every subject-related node's connection statuses over the $n$ time snapshots and calculate the node similarity between each subject-related node and the subject node by using a sliding window. Finally, we sort the values of node similarity in a descending order and print out them in a descending order with their node similarity. According to the result, we can predict which subject-related nodes are more likely to connect with the subject node in the future.

**Algorithm 1** Durable Node Similarity
**Input:** dynamic graph $DG = \{<V_i, E_i, t_i> \mid 1 \le i \le n\}$, the subject node $v_0$
**Output:** every subject-related node and their node similarity and its label
1: $DG = \{\langle V_i, E_i, t_i\rangle \mid 1 \le i \le n\}$;
2: **for** $i$ = 1 to $n$ **do**
3: select subject-related nodes in $n$ time snapshots;
4: get the 0-1 representation for $m$ subject-related nodes;
5: **end for**
6: **if** $n$ = 2
7: **for** $k$ = 1 to $m$ **do**
8: $short\text{–}sim\left(v_0^2, v_k^2\right) = \frac{\left(v_0^1, v_0^2\right)\cdot\left(v_k^1, v_k^2\right)^T}{\left\|\left(v_0^1, v_0^2\right)\right\|\left\|\left(v_k^1, v_k^2\right)\right\|}$;
9: $long\text{–}sim(v_0, v_k) = short\text{–}sim\left(v_0^2, v_k^2\right)$;
10: set a label for $v_k$;
11: **end for**
12: **else**
13: $w^1 = \sqrt{\frac{1}{3!}}$;
14: $w^2 = \sqrt{\frac{2}{3!}}$;
15: $w^3 = \sqrt{\frac{3}{3!}}$;
16: **for** $k$ = 1 to $m$ **do**
17: **for** $i$ = 3 to $n$ **do**
18: $medium\text{–}sim\left(v_0^i, v_k^i\right) = \frac{w^1 v_0^{i-2} w^1 v_k^{i-2} + w^2 v_0^{i-1} w^2 v_k^{i-1} + w^3 v_0^i w^3 v_k^i}{\sqrt{\left(w^1 v_0^{i-2}\right)^2 + \left(w^2 v_0^{i-1}\right)^2 + \left(w^3 v_0^i\right)^2}\sqrt{\left(w^1 v_k^{i-2}\right)^2 + \left(w^2 v_k^{i-1}\right)^2 + \left(w^3 v_k^i\right)^2}}$;
19: save $medium\text{-}sim(v_0{}^i, v_k{}^i)$ as the value of connection at time $t_{i-1}$ or future time;
20: change the label of $v_k$;
21: **end for**
22: $long\text{–}sim(v_0, v_k) = medium\text{–}sim\left(v_0^n, v_k^n\right)$;
23: **end for**
24: **end if**
25: sort subject-related nodes in each label set in a descending order;
26: return every subject-related node and their node similarity and its label;

## 5 Experiments

We evaluate the efficiency and effectiveness of our newly proposed node similarity measurement method on real-world networks. The algorithm is implemented in Java, and the experiments are conducted on the platform of Windows 7 Enterprise with Intel Core i5 CPU and 726GB main memory.

### 5.1 Running time

A randomly generated dynamic graph is used to evaluate the proposed method and demonstrate the maximum running time when processing different scale of the graph firstly. The random dynamic

graph can generate time snapshots to simulate the evolution of the dynamic graph. Based these time snapshots, the running time of the proposed method can be computed. The flowchart of the experiment is shown in Figure 3.

Our experiments on randomly generated dynamic graphs are conducted in two settings: (i) When the scale of a randomly generated dynamic graph (i.e., the total number of nodes in a generated graph) is a fixed, we test the relationship between the running time and the number of subject-related nodes. (ii) When the number of subject-related nodes is fixed, we test the relationship between the running time and the scales of a randomly generated dynamic graph. The scale of a randomly generated dynamic graph is the total number of nodes in the graph.

For the first setting, the total number of nodes in our generated graph is fixed to 100,000. Experimental results show the relationship between running time and different numbers of subject-related nodes. As shown in Figure 4, when the number of subject-related nodes increases, the running time of the proposed method increases nonlinearly (i.e., the trend line of running time can be fitted into a quadratic polynomial function). The maximum running time appears when the number of subject-related nodes is the largest. In other words, when all nodes in a graph are connected to the subject node (i.e., fully connected with the subject node), the proposed method has the maximum running time.

In the second setting, we intend to find the relationship between the running time and the scale of a dynamic graph. According to the result obtained in the first experiment, when all nodes in a graph are connected to the subject node, the proposed method has the maximum running time when the scale of the graph is fixed. Thus, we set all nodes in the randomly generated dynamic graph to be fully connected firstly. Then the experimental results show the relationship between the running time and the scale of a randomly generated dynamic graph. As shown in Figure 5, when the scale of the dynamic graph increases, the running time of the proposed method increase nonlinearly (i.e., the trend line of running time can be fitted into a quadratic polynomial function). The overall trend is that larger scale of dynamic graph has longer running time.

From the experimental results above, even when there are 1,000,000 nodes, the running time is only about 500s. As this test is conducted using a single computer, when the proposed method is used to deal with a big dataset (i.e., a larger scale of dynamic graph), it is possible to use parallel computing or other types of calculation to reduce the running time. Therefore, the proposed method can be used for processing a big dataset.

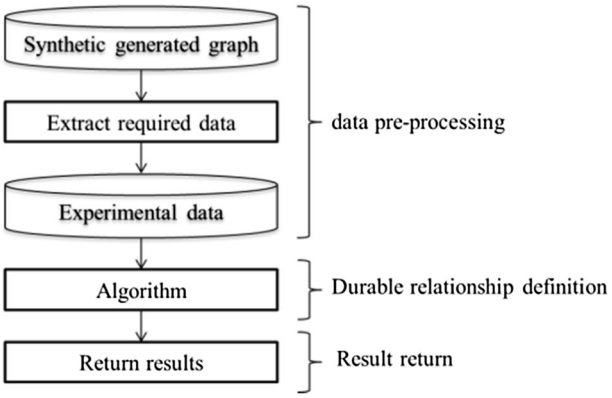



**Figure 3** Flowchart of the experiment

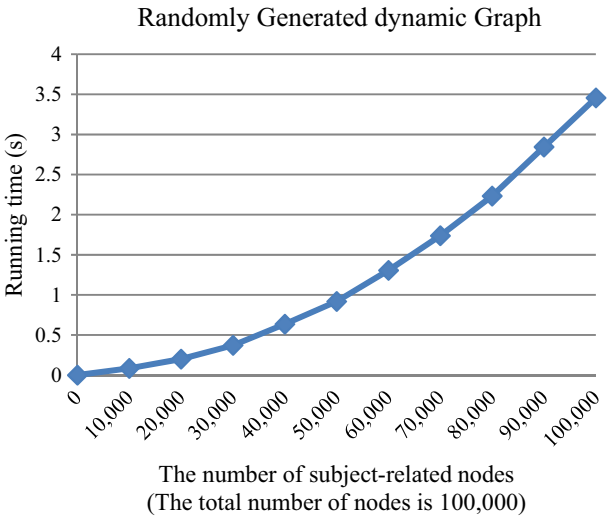



**Figure 4** Running time on randomly generated dynamic graph

## 5.2 Prediction accuracy

In order to test the effectiveness and efficiency of the proposed method in prediction, a real-world network, the DBLP co-authorship network, is used. DBLP provides a comprehensive list of research papers in Computer Science. Until June 2017, this DBLP co-authorship network includes 1,919,861 authors and 3,799,107 publications.

We will collect six time snapshots from 2006 to 2011 of the DBLP co-authorship network to explain the evolution of this dynamic graph over these 6 years. Then the co-authorship from 2012 to 2014 will be used to test the accuracy of our proposed method. An edge between two

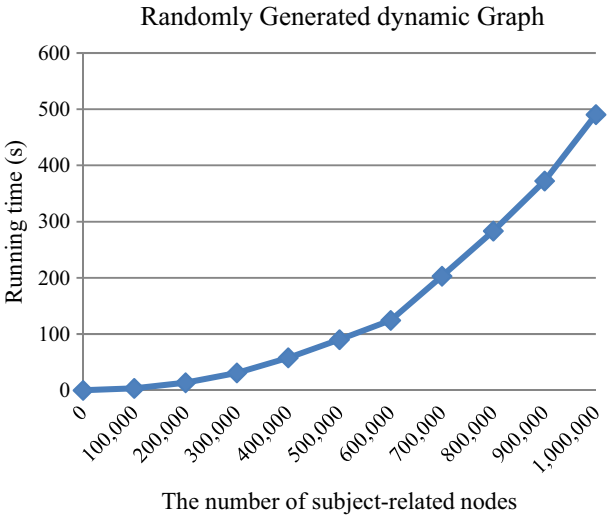



**Figure 5** Running time on randomly generated dynamic graph

nodes/authors in a time snapshot of the DBLP co-authorship network represents the two authors who have co-authored at least one paper or papers in the year.

Our newly proposed node similarity measurement is based on the frequency of connections between a subject node and subject-related nodes over a period (i.e. across several sequential snapshots), thus the durable relationship description only focuses on the connection between the subject node and the subject-related nodes. The frequency corresponds to how many connection/edges between two nodes appear in a sequence of snapshots. For example, as shown in Figure 1, node $v_0$ (author's name is Shichao Zhang) is the subject node, the relationships that the subject node has include all nodes which have edges connected to the subject node. Because our proposed method is to describe the relationship between two nodes, thus some papers like [31], which are published only by Shichao Zhang himself, are ignored in the experiments.

In order to test the effectiveness and efficiency of the proposed method, two different subject nodes are selected for two experiments, respectively. The first experiment is used to test the effectiveness of the proposed method in terms of prediction. As shown in Figure 2, these two nodes are node $v_0$ (author's name is Shichao Zhang) and node $v_{19}$ (author's name is Xindong Wu). The second experiment is used to test the efficiency of the proposed method. We compared the predictive accuracy of our proposed method in a medium term and a long term.

The reason that these two authors are chosen is because they have a broad research area and a large amount of co-authors. For example, the research area of Shichao Zhang includes: sparse coding [35], data mining [24, 33, 34], and statistics [36].

The first experimental results by using node $v_0$ show there are 54 authors who have co-authored papers with node $v_0$ during the 6 years. As shown in Figure 2, there are 10 authors who have more than two connections with node $v_0$ over these 6 years. The nodes like node $v_{41}$ which only co-authored paper with node $v_0$ only one time has been ignored [32]. In other words, in this durable relationship, there are 10 subject-related nodes have long-term In Relationship or long-term Strong Relationship with the subject node $v_0$. These 10 authors' co-authorship information has been listed in Table 4. According to Algorithm 1 defined above, the values of long-term node similarity of these 10 subject-related nodes with the subject node (node $v_0$) have been listed in the fourth column from the right in Table 4.

The similarity values of these nodes with node $v_0$ are in the descending order: $v_{19}$, $v_{23}$, $v_{38}$, $v_2$, $v_{27}$, $v_{14}$, $v_{11}$, $v_{15}$, $v_{24}$, $v_3$. The node that has a higher similarity value is more similar to the subject node. The durable relationship can be described as: (i) In the dimension of durations of time, a higher similarity value of a subject-related node means that it has longer and more recent connection with the subject node; (ii) Based on the frequencies of connections, the connection between this node and the subject node is more frequent than other subject-related nodes.

In conjunction with Figure 2, node $v_{19}$ has the highest value of similarity with the subject node $v_0$. Node $v_{19}$ has the largest number of connections with the subject node $v_0$ over this period. Therefore, we can predict that node $v_{19}$ is more likely to connect with the subject node in the future. It means that according to the durable relationship description above, Xindong Wu has the largest possible to continue the co-authorship with Shichao Zhang in the following 3 years (2012-2014). The last three columns in Table 4 show the evolution relationship between the subject node $v_0$ and its subject-related nodes in the following 3 years. We can find that node $v_{19}$ still has connection with node $v_0$ in these three time snapshots.

**Table 4** Subject-related nodes which connected with node $v_o$ more than 3 years

| Node No. | Co-authorship in 6 years | | | | | | Node similarity | Co-authorship continuity | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | | 2012 | 2013 | 2014 |
| 19 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 38 | 0 | 0 | 1 | 1 | 1 | 1 | 0.9946 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 1 | 1 | 0.824 | 0 | 0 | 0 |
| 27 | 1 | 1 | 1 | 1 | 0 | 1 | 0.7567 | 0 | 1 | 0 |
| 14 | 1 | 1 | 1 | 1 | 1 | 0 | 0.7071 | 0 | 0 | 0 |
| 11 | 1 | 0 | 1 | 0 | 1 | 0 | 0.4091 | 0 | 0 | 0 |
| 15 | 1 | 0 | 1 | 0 | 1 | 0 | 0.4091 | 1 | 1 | 0 |
| 24 | 1 | 1 | 1 | 1 | 0 | 0 | 0.2887 | 1 | 1 | 0 |
| 3 | 1 | 1 | 1 | 1 | 0 | 0 | 0.2887 | 0 | 0 | 0 |

The first experimental results on Xindong Wu show there are 175 authors who have co-authored paper with node $v_{19}$ (author's name is Xindong Wu) during the 6 years. The information about the subject-related nodes has been listed in Appendix Table 8. In order to describe more concise, we only focus on the subject-related nodes which have long-term In Relationship or long-term Strong Relationship with the subject node $v_{19}$ over this period. There are 23 authors have more than two connections with node $v_{19}$ which has been listed in Table 5. According to Algorithm 1 defined above, the value of node similarity of each subject-related node with the subject node is listed in the right third column in Table 5.

**Table 5** Subject-related nodes which connected with node $v_{19}$ more than 3 years

| NodeNo. | Co-authorship in 6 years | | | | | | Node similarity | Co-authorship continuity | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | | 2012 | 2013 | 2014 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 37 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 104 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 107 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 136 | 0 | 1 | 1 | 1 | 1 | 1 | 0.9996 | 1 | 1 | 1 |
| 281 | 0 | 1 | 1 | 1 | 1 | 1 | 0.9996 | 1 | 1 | 1 |
| 220 | 0 | 1 | 1 | 1 | 1 | 1 | 0.9996 | 1 | 1 | 0 |
| 124 | 0 | 0 | 1 | 1 | 1 | 1 | 0.9946 | 0 | 1 | 1 |
| 189 | 0 | 1 | 0 | 1 | 1 | 1 | 0.9691 | 1 | 1 | 1 |
| 139 | 0 | 0 | 0 | 1 | 1 | 1 | 0.9677 | 1 | 1 | 1 |
| 181 | 0 | 0 | 0 | 1 | 1 | 1 | 0.9677 | 1 | 1 | 0 |
| 318 | 0 | 0 | 0 | 1 | 1 | 1 | 0.9677 | 0 | 0 | 0 |
| 285 | 1 | 1 | 1 | 0 | 1 | 1 | 0.824 | 0 | 0 | 0 |
| 311 | 0 | 0 | 1 | 0 | 1 | 1 | 0.8018 | 1 | 1 | 1 |
| 297 | 0 | 1 | 1 | 1 | 0 | 1 | 0.756 | 0 | 0 | 0 |
| 221 | 0 | 0 | 1 | 1 | 0 | 1 | 0.7474 | 0 | 0 | 0 |
| 225 | 0 | 0 | 1 | 1 | 0 | 1 | 0.7474 | 0 | 0 | 0 |
| 197 | 1 | 1 | 0 | 0 | 1 | 1 | 0.6223 | 1 | 1 | 1 |
| 145 | 1 | 1 | 0 | 0 | 1 | 1 | 0.6223 | 1 | 1 | 0 |
| 293 | 1 | 0 | 1 | 0 | 1 | 0 | 0.4091 | 0 | 0 | 0 |
| 282 | 1 | 1 | 1 | 1 | 0 | 0 | 0.2887 | 0 | 1 | 1 |
| 249 | 1 | 1 | 1 | 1 | 0 | 0 | 0.2887 | 0 | 0 | 0 |

According to Table 5, node $v_{124}$ co-authored papers with node $v_0$ from 2008 to 2011 while node $v_{249}$ co-authored papers with node $v_0$ from 2006 to 2009. Both of these two nodes co-authored papers with node $v_0$ four times. However, the publications co-authored by node $v_{124}$ and node $v_0$ are more close to current time (i.e., the connection between node $v_0$ and node $v_{124}$ is more recent). For prediction, from the last three columns in Table 5 we can find that node $v_{124}$ co-authored paper with node $v_0$ in 2013 and 2014 while node $v_{40}$ not. Thus, the experimental results test that the subject-related node which has higher similarity is more likely to cooperate with the subject node in the future.

The results listed in Table 5 verified that: (i) the subject-related nodes which connect to the subject node longer and more frequent have higher similarity values (i.e., more strong relationships); (ii) a subject-related node which has a higher similarity value is more likely to cooperate with the subject node in the future.

Table 6 listed the relationships between the subject node $v_0$ and all its related nodes in the medium term from 2006 to 2014. The first column shows all types of medium-term relationships except No Relationship. The second column shows the corresponding similarity values. The third column shows the sum of the numbers of a type in each sliding window divided by the number of sliding windows. For example, there are 7 sliding windows from 2006 to 2014 (i.e. 2006-2008, 2007-2009, … , 2012-2014). The forth column shows the average number of cooperation occurs in the following year of a sliding window. For example, in the first row, the third column is 4.0, while the fourth column is 2.7. The percentage in the fifth column is obtained by (2.7 divided by 4.0). The sixth column is corresponding to the average cooperation occurred during the following 2 years of the sliding windows. This number should be large than the number shown in the fourth column. Table 7 listed the relationships between the subject node $v_{19}$ and all its related nodes in the medium-term from 2006 to 2014. Comparing with the results listed in the last two columns in Tables 4 and 5 respectively, we can find that the accuracy of prediction based on a medium-term relationship description is higher than based on a long-term relationship description. We can also quantify this conclusion by using more subject nodes. However, these two experiments can support this conclusion.

Based on these two experiments above, the effectiveness and efficiency of the proposed method has been tested completely. The similarity value between a subject node and a subject-related node can effectively predict and describe durable relationship in a large dynamic graph over a period. For the prediction of durable relationship evolution, a medium-term relationship description appears more accurate than a long-term relationship description.

The proposed method provides a new way to describe and predict node similarity/relationship in a large dynamic graph, which also can be used in real-world applications such as entity relationship definition, and entity relationship prediction. Except for processing the DBLP co-authorship network, it also can be used in other relationship analysis.

## 6 Conclusions

In this paper, a novel type of node similarity based on the frequency of connections between two nodes (i.e., entities) is proposed to describe the changeable relationships between entities over a shot-, medium- or long- period, which has not been considered as an indication of similarity between two nodes before. In this paper, durable relationships describe the frequency of connections rather than just the continuous connection between two nodes. According to this definition, durable relationships can be defined in two dimensions (durations of time and

**Table 6** The relationship between subject-related nodes of node $v_0$ and node $v_0$ in medium-term

| Relationship Type | Similarity Value | Subject-related nodes number (in average) | Cooperate in the following year (in average) | Percentage | Cooperate in the following 2 years (in average) | Percentage | Cooperate in following 3 years (in average) | Percentage |
|---|---|---|---|---|---|---|---|---|
| 111 | 1 | 4.0 | 2.7 | 64.96% | 3.2 | 76.23% | 3.3 | 79.56% |
| 011 | 0.913 | 1.2 | 0.3 | 25.00% | 0.3 | 25.00% | 0.5 | 30.56% |
| 101 | 0.816 | 2.2 | 0.5 | 15.28% | 1.5 | 68.06% | 1.5 | 68.06% |
| 110 | 0.707 | 1.5 | 0.5 | 25.00% | 0.7 | 33.33% | 0.7 | 33.33% |
| 001 | 0.707 | 6.8 | 0.8 | 14.33% | 1.7 | 25.71% | 2.0 | 32.26% |
| 010 | 0.577 | 8.2 | 1.8 | 23.33% | 2.2 | 28.33% | 2.2 | 28.33% |
| 100 | 0.408 | 8.5 | 0.8 | 9.37% | 0.8 | 9.37% | 1.2 | 11.33% |

**Table 7** The relationship between subject-related nodes of node $v_{19}$ and node $v_{19}$ in a medium term

| Relationship Type | Similarity Value | Number of subject-related nodes (in average) | Cooperate in the following year (in average) | Percentage | Cooperate in the following 2 years (in average) | Percentage | Cooperate in the following 3 years (in average) | Percentage |
|---|---|---|---|---|---|---|---|---|
| 111 | 1 | 12.2 | 9.3 | 79.13% | 10.5 | 88.33% | 10.8 | 90.49% |
| 011 | 0.913 | 7.7 | 3.5 | 39.32% | 4.5 | 56.54% | 4.5 | 56.54% |
| 101 | 0.816 | 5.0 | 1.2 | 24.14% | 1.5 | 35.25% | 2.0 | 50.13% |
| 110 | 0.707 | 6.8 | 1.8 | 24.19% | 2.3 | 32.52% | 3.0 | 43.65% |
| 001 | 0.707 | 33.5 | 7.0 | 20.53% | 10.8 | 32.20% | 13.2 | 38.63% |
| 010 | 0.577 | 28.2 | 3.8 | 14.25% | 5.2 | 18.67% | 6.0 | 21.66% |
| 100 | 0.408 | 24.5 | 1.7 | 7.12% | 2.8 | 11.74% | 3.5 | 15.25% |

frequencies of connections). A similar node query method is proposed, based on our new definition of node similarity, to study the statues of durable relationships from a longitudinal study point of view. This method provides a new way to describe the semantics of durable relationships, and also gives a practical application of node similarity measurement in the real world. In the meantime, it is also a useful method to predict the evolution of a relationship in the future.

Our extensive experiments show that the proposed method can effectively describe durable relationships and especially predict future relationships. The experimental results shows the running time of our proposed method is not long even in processing a big dataset (i.e., a larger scale of dynamic graph) on a single computer. The value of similarity between a subject node and subject-related node can effectively describe durable relationship in a large dynamic graph over a period. The medium-term relationship description can predict durable relationship evolution more accurately than the long-term relationship description.

Furthermore, the experimental results also show that the proposed method can be used in real-world applications. The proposed method can be used in entity relationship definition, and also in entity relationship prediction. Besides in processing the DBLP co-authorship network, it also can be used in other relationship analysis.

In the future, we intend to consider the quality of connections, which can be defined as the number of co-authored papers in a year and/or the quality of the co-authored papers. The quality of connections may give more insight in durable relationships and may even provide the prediction of the quality of a relationship.

**Acknowledgements** This work is supported in part by the New Zealand Marsden Fund, the Chinese Scholarship Council, and the National Natural Science Foundation of China under Grant (No.61472169).

## Appendix

**Table 8** The list of the subject node $v_{19}$ and its subject-related nodes

| Author's name | No. | Author's name | No. | Author's name | No. | Author's name | No. | Author's name | No. |
|---|---|---|---|---|---|---|---|---|---|
| A. A. Hamed | 79 | D. Guo | 106 | L. Li | 201 | Z. Ren | 321 | G. Wei | 130 |
| D. A. Simovici | 105 | L. Guo | 200 | L. Li | 202 | A. Rubin | 81 | S. Wei | 245 |
| S. Abdullah | 240 | S. Hao | 244 | P. Li | 220 | T. S. Chua | 254 | X. Wu | 19 |
| M. Ali | 213 | Z. Hao | 319 | R. Li | 239 | V. S. Sheng | 261 | G. Wu | 136 |
| A. An | 80 | M. He | 67 | X. Li | 276 | P. S. Yu | 223 | Y. Wu | 301 |
| Y. An | 303 | D. He | 107 | X. Li | 280 | Y. Saygin | 309 | T. Xiang | 250 |
| J. Bailey | 161 | J. He | 173 | X. Li | 282 | J. Shen | 168 | Y. Xiao | 288 |
| B. Baruque | 94 | T. He | 257 | Y. Li | 284 | Y. Shen | 291 | B. Xie | 86 |
| P. Beinat | 218 | Y. He | 302 | Z. Li | 311 | V. Sheng | 260 | F. Xie | 124 |
| P. Berkhin | 219 | A. Herrero | 82 | Q. Liang | 225 | J. Shi | 190 | Z. Xie | 312 |
| Y. Bi | 304 | R. Hong | 233 | Y. Lin | 287 | Y. Shi | 297 | H. Xiong | 157 |
| F. Bonchi | 128 | X. Hong | 275 | Y. Lin | 290 | A. Siebes | 84 | G. Xu | 137 |
| J. Bongard | 183 | T. Khoshgoftaar | 32 | C. Ling | 96 | M. Song | 211 | J. Xu Yu | 164 |
| K. C. C. Chan | 194 | W. Hsu | 271 | H. Liu | 43 | Y. Song | 292 | J. Xuan | 176 |
| P. C. K. Hung | 217 | D. Hu | 117 | B. Liu | 90 | A. Srivastava | 85 | F. Xue | 126 |
| H. C. Lau | 154 | M. Hu | 210 | B. Liu | 91 | M. Steinbach | 209 | X. Xue | 272 |

**Table 8** (continued)

| Author's name | No. | Author's name | No. | Author's name | No. | Author's name | No. | Author's name | No. |
|---|---|---|---|---|---|---|---|---|---|
| V. C. Leung | 259 | W. Hu | 269 | C. Liu | 101 | D. Steinberg | 108 | B. Yang | 92 |
| J. Cao | 52 | X. Hu | 274 | W. Liu | 265 | K. Su | 192 | F. Yang | 127 |
| C. Cao | 103 | X. Hu | 281 | X. Liu | 278 | Y. Sui | 306 | P. Yang | 222 |
| L. Cao | 204 | E. Hua | 122 | Y. Liu | 295 | J. Sun | 178 | Y. Yang | 293 |
| T. Cao | 256 | F. Huang | 7 | Y. Liu | 298 | Z. Sun | 314 | H. Yao | 151 |
| R. Caruana | 232 | H. Huang | 158 | J. Lu | 11 | R. T. White | 234 | M. Ye | 212 |
| F. Chen | 1 | J. Huang | 179 | C. Lu | 95 | L. T. Yang | 198 | W. Ye | 270 |
| D. Chen | 115 | Y. Huang | 299 | H. Lu | 144 | X. Tang | 277 | J. Yin | 175 |
| G. Chen | 135 | Zi Huang | 325 | K. Lu | 193 | D. Tao | 104 | X. You | 42 |
| P. Chen | 224 | G. I. Webb | 132 | Z. Lu | 318 | K. Thompson | 196 | K. Yu | 197 |
| Q. Chen | 226 | S. Islam | 242 | W. Lv | 268 | B. Thuraising | 88 | J. Z. Huang | 184 |
| S. Chen | 248 | D. J. Cook | 113 | J. Ma | 170 | W. Tian | 267 | Z. Zha | 317 |
| X. Chen | 279 | G. J. F. Jones | 131 | G. Mao | 138 | H. Toivonen | 142 | S. Zhang | 0 |
| J. Cheng | 188 | D. J. Hand | 109 | T. Mei | 253 | R. Vilalta | 231 | C. Zhang | 23 |
| Y. Cheung | 296 | S. J. Maybank | 249 | G. Melli | 129 | J. Vreeken | 177 | J. Zhang | 24 |
| C. Chi | 98 | G. J. McLachlan | 133 | F. Min | 123 | C. W. Clifton | 99 | Z. Zhang | 53 |
| E. Corchado | 120 | M. J. Zaki | 12 | H. Motoda | 149 | R. Wang | 41 | C. Zhang | 102 |
| Q. Dai | 227 | R. Ji | 237 | Y. Mu | 286 | C. Wang | 97 | J. Zhang | 181 |
| W. Ding | 66 | S. Ji | 246 | A. N. Arslan | 78 | D. Wang | 111 | L. Zhang | 199 |
| R. Duan | 235 | H. Jiang | 146 | Qiang Yang | 47 | D. Wang | 112 | P. Zhang | 221 |
| M. Ester | 206 | J. Jiang | 187 | H. Ning | 156 | D. Wang | 116 | R. Zhang | 238 |
| A. F. M. Ng | 83 | X. Jiang | 283 | Z. Niu | 315 | F. Wang | 125 | Y. Zhang | 285 |
| T. F. Stepinski | 258 | Z. Jiang | 313 | E. Oja | 121 | H. Wang | 140 | Y. Zhang | 300 |
| C. Faloutsos | 100 | Z. Jin | 38 | L. P. Bandeira | 205 | H. Wang | 141 | Y. Zhang | 307 |
| W. Fan | 264 | N. K. | 215 | J. P. Bond | 163 | H. Wang | 145 | Z. Zhang | 310 |
| T. Fandy | 252 | E. K. Park | 119 | J. Pan | 171 | H. Wang | 155 | Y. Zhao | 289 |
| M. Fang | 207 | H. Kargupta | 148 | J. Pei | 169 | J. Wang | 172 | Z. Zhao | 322 |
| Y. Fei | 308 | G. Karypis | 134 | H. Peng | 150 | J. Wang | 180 | T. Zhou | 255 |
| J. Feng | 166 | B. Kitts | 93 | M.Spilopoulou | 214 | J. Wang | 186 | Z. Zhou | 320 |
| J. Feng | 167 | R. Kumar | 229 | V. Piuri | 262 | L. Wang | 203 | X. Zhu | 27 |
| J. Gao | 189 | V. Kumar | 263 | K. Q. Zhu | 195 | M. Wang | 208 | X. Zhu | 37 |
| Y. Gao | 305 | D. L. Small | 110 | H. Qi | 153 | R. Wang | 236 | M. Zhu | 51 |
| R. Ghani | 230 | B. Li | 89 | J. Qiang | 182 | S. Wang | 247 | Q. Zhu | 228 |
| J. Ghosh | 185 | H. Li | 139 | J. R. Fingar | 162 | W. Wang | 266 | Z. Zhu | 316 |
| H. Glotin | 147 | H. Li | 143 | J. R. Quinlan | 160 | X. Wang | 273 | Z. Zhu | 324 |
| B. Goethals | 87 | H. Li | 152 | O. R. Zaiane | 216 | Y. Wang | 294 | D. Zong | 118 |
| J. Gui | 174 | H. Li | 159 | S. Ranka | 241 | Z. Wang | 323 | | |
| D. Gunopulos | 114 | J. Li | 191 | J. Ren | 165 | T. Washio | 251 | | |

## References


1. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 44-54. ACM (2006)
2. Berlingerio, M., Pinelli, F., Calabrese, F.: Abacus: frequent pattern mining-based community discovery in multidimensional networks. Data Mining and Knowledge Discovery. **27**(3), 294–320 (2013)
3. Qian, G., Sural, S., Gu, Y., Pramanik, S.: Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In: Proceedings of the 2004 ACM symposium on Applied computing, pp. 1232-1237. ACM (2004)
4. Coogan, S., Arcak, M.: Dynamical properties of a compartmental model for traffic networks. In: Proceedings of the 2014 American Control Conference (ACC), pp. 2511-2516. IEEE (2014)
5. Cui, W., Xiao, Y., Wang, H., Lu, Y., Wang, W.: Online search of overlapping communities. In: Proceedings of the 2013 ACM SIGMOD international conference on Management of data, pp. 277-288. ACM (2013)
6. Deza, M. M., Elena D.: Encyclopedia of distances. Encyclopedia of Distances, 1-583. Springer Berlin Heidelberg, 2009
7. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology. **26**(3), 297–302 (1945)
8. Ding, L., Li, Z., Ji, W., Song, B.: Reachability query of large scale dynamic graph based on improved Huffman coding. Acta Electronica Sinica. **45**(2), 359–367 (2017)
9. Ellison, N.B., Vitak, J., Gray, R., Lampe, C.: Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. Journal of Computer-Mediated Communication. **19**(4), 855–870 (2014)
10. Freitas, A. A. Data mining and knowledge discovery with evolutionary algorithms. Springer Science & Business Media (2013)
11. Gao, Y., Lin, M., Wang, R.: Adaptive support framework for wisdom web of things. World Wide Web. **16**(4), 379–398 (2013)
12. Guo, T., Wu, J., Zhu, X., Zhang, C.: Combining Structured Node Content and Topology Information for Networked Graph Clustering. ACM Transactions on Knowledge Discovery from Data (TKDD). **11**(3), 29 (2017)
13. Horace, W. II, and John B. N.: Method and system for implementing a custom workspace for a social relationship management system. U.S. Patent Application, 14/828, 466 (2015)
14. Ipsen, M., Mikhailov, A.S.: Evolutionary reconstruction of networks. Physical Review E. **66**(4), 046109 (2002)
15. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat. **37**, 547–579 (1901)
16. Kusumoto, M., Maehara, T., Kawarabayashi, K. I.: Scalable similarity search for SimRank. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 325-336. ACM (2014)
17. Li, X., Chang, L., Zheng, K., Huang, Z., Zhou, X.: Ranking weighted clustering coefficient in large dynamic graphs. World Wide Web. **20**(5), 855–883 (2017)
18. Makvana, K., Shah, P., Shah, P.: A novel approach to personalize web search through user profiling and query reformulation. In: Proceedings of the 2014 Data Mining and Intelligent Computing (ICDMIC), pp. 1-10. IEEE (2014)
19. Ochiai, A.: Zoogeographic studies on the solenoid fishes found in Japan and its neighboring regions. Bull. Jpn. Soc. Sci. Fish. **22**(9), 526–530 (1957)
20. Ren, K., Qian, X.: Research on User Similarity Measurement Method in Collaborative Filtering Algorithm. Computer Engineering. **41**(8), 18–22 (2015)
21. Shang, J., Wang, C., Wang, C., Guo, G., Qian, J.: An attribute-based community search method with graph refining. J. Supercomput. 1-28 (2017). https://doi.org/10.1007/s11227-017-1976-z
22. Song, B., Ji, W., Ding, L.: Similarity nodes query algorithm on large dynamic graph based on the snapshots. Journal of Computer Applications. **36**(2), 358–363 (2016)
23. Tantipathananandh, C., Berger-Wolf, T., Kempe, D.: A framework for community identification in dynamic social networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 717-726. ACM (2007)
24. Wang, T., Qin, Z., Zhang, S., Zhang, C.: Cost-sensitive classification with inadequate labeled data. Information Systems. **37**(5), 508–516 (2012)
25. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386-1393. (2014)
26. Wang, P., Gao, L., Ma, X.: Dynamic community detection based on network structural perturbation and topological similarity. Journal of Statistical Mechanics: Theory and Experiments. **2017**(1), 013401 (2017)

27. Wu, T., Chen, L., Zhong, L., Xian, X.: Predicting the evolution of complex networks via similarity dynamics. Physica A: Statistical Mechanics and its Applications. **465**, 662–672 (2017)
28. Yang, Y., Pei, J., Al-Barakati, A.: Measuring in-network node similarity based on neighborhoods: a unified parametric approach. Knowl Inf Syst, **53**, 43-70 (2017)
29. Yu, Y., Gao, Y., Wang, H., Wang, R. Joint user knowledge and matrix factorization for recommender systems. In International Conference on Web Information Systems Engineering, 77-91. Springer International Publishing, 2016
30. Yuksel, A., Yuksel, F., Bilim, Y.: Destination attachment: Effects on customer satisfaction and cognitive, affective and conative loyalty. Tourism management. **31**(2), 274–284 (2010)
31. Zhang, S.: Shell-neighbor method and its application in missing data imputation. Applied Intelligence. **35**(1), 123–133 (2011)
32. Zhang, S., Chen, F., Jin, Z., Wang, R.: Mining class-bridge rules based on rough sets, Expert Systems with Applications, Vol.36 (3). Part. **2**, 6453–6460 (2009)
33. Zhang, S., Jin, Z., Zhu, X.: Missing data imputation by utilizing information within incomplete instances. Journal of Systems and Software. **84**(3), 452–459 (2011)
34. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z.: Missing value estimation for mixed-attribute data sets. IEEE Transactions on Knowledge and Data Engineering. **23**(1), 110–121 (2011)
35. Zhu, X., Li, X., Zhang, S., Ju, C., Wu, X.: Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Transactions on Neural Networks and Learning Systems. **28**(6), 1263–1275 (2017a)
36. Zhu, X., Li, X., Zhang, S., Xu, Z., Yu, L., Wang, C.: Graph PCA Hashing for Similarity Search. IEEE Transactions on Multimedia. (2017b)