Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7490656, 20 pages
https://doi.org/10.1155/2022/7490656

Research Article

Relative Entropy-Based Similarity for Patterns in Graph Data

Shihu Liu, Li Deng, Haiyan Gao, and Xueyu Ma

School of Mathematics and Computer Sciences, Yunnan Minzu University, Kunming 650504, China

Correspondence should be addressed to Li Deng; dengli713@126.com

Received 26 April 2022; Revised 17 June 2022; Accepted 27 June 2022; Published 26 July 2022

Academic Editor: Chao Zhang

How to make a correct similarity between patterns is a groundwork in data mining, especially for graph data. Despite these methods that can obtain great results, there may be still some limitations, for instance, the similarity of patterns in directed weighted graph data. Here, we introduce a new approach by taking the so-called the second-order neighbors into consideration. The proposed new similarity approach is named as relative entropy-based similarity for patterns in graph data, wherein the relative entropy provides a brand new aspect to make the difference between patterns in directed weighted graph data. The proposed similarity measure can be partitioned under three phases. First of all, strength set is given by degree and weight of patterns; in this phase, four variables holding the strength about out-degree, in-degree, out-weight, and in-weight are constructed. Then, with the help of Euclidean metric, pattern's probability set is constructed, which contains influence of similarity between pattern and its all one-order neighbors. Finally, relative entropy is used to measure the difference between patterns. In order to examine the validity of our approach as well as its advantage comparing with the state-of-art approach, two sorts of experiments are suggested for real-world and synthetic graph data. The outcomes of experiment indicate that the recommended method get handy execution done measuring similarity and gain accurate results.

1. Introduction

At present, many practical networks like Facebook social networks, protein interaction networks, aviation networks, and disease transmission networks can be presented as graph data. This type of data is no longer a straightforward portrayal of pattern's attribute information and composes possible topological information between patterns additionally, i.e., degree and weight of patterns. Due to the extensive use of graph data, many practical problems including pattern analysis, link prediction, and community detection can be abstracted into problems of graph data for research. Among these researches, how to calculate similarity between patterns in graph data is considered as one of the fundamental problems. Many researches on graph data are based on the pattern's similarity measure, for example, traffic networks [1], image classification [2], and pattern recognition [3, 4].

Over the past few decades, discovering similarity between patterns has attracted substantial consideration [5, 6]. Scholars proposed a range of methods to measure pattern's similarity, for example, shared neighbor-based similarity, random walk-based similarity, path-based similarity, and information theory-based similarity; these methods discuss the similarity of patterns from different perspective.

The shared neighbor-based similarity measure takes into account the shared information of the connected neighbors between patterns, and the greater the coincidence rate of shared neighbors means the higher similarity of two patterns. Cosine index [7], Sorensen index [8], Jaccard index [9], AA (Adamic-Adar) index [10], and WAA (weighted Adamic-Adar) index [11] are also common methods used in the research of similarity measure, which take into account the number of shared neighbors. Besides, LP (local path) [12] index is an improvement of CN index [13]; on the basis of CN index, the influence of neighbor with path length of 3 on the connection between patterns is added. These indicators reduce the computation time and earn good results in the identification of the most similar patterns. Unfortunately, they remain significant challenges, only topological information of first-order neighbors is

taken into account, and many patterns with high similarity have no common neighbors, which leads to certain limitations of such indicators.

Random walk-based similarity is widely used to measure the topological similarity of patterns, such as MLRW (Multiplex Local Random Walk) index [14], BRW (Biased Random Walk) index [15], and LRW (Local Random Walk) index [16]. In the process of calculation, these two methods measure similarity moving from one to other patterns through multistep random walk without the global information of graph data. They simplify the similarity measure to some extent, but these three similarity measures rely on large-degree patterns and most similar patterns may be large-degree patterns, which makes the similarity results sensitive to large pattern dependence.

Path-based similarity is an important method used to measure similarity of patterns, Katz index [17, 18] and ACT (average commute time) [19]. Compared with local index, global index requires the overall topological information. Besides, Aziz et al. in [20] proposed global and quasilocal extensions of some commonly used local similarity index. Although global index provides more accurate similarity than local index, the computation of global metrics is time-consuming and generally not applicable to large-scale graph data, and sometimes, global topology information is unavailable, especially when implemented in a decentralized manner.

In addition to the similarity measures mentioned above, information theory-based similarity is a kind of similarity measure that is often used. Hereinto, relative entropy is an important concept of information theory, which are used to measure similarity of patterns. Scholars have proposed pattern similarity measure based on relative entropy such as LRE (i.e., the abbreviation of local relative entropy) [21, 22], LRWE-SNM (Local Network Relative Weighted Entropy Based Similar Node Mining) [23], and RE-model (relative entropy model) [24]. These methods have advantages in their respective fields and can also measure the similarity of patterns to a certain extent. Although it is faster and simpler to measure, some pattern's information and complex relationships between patterns are lost, for example, information about second-order neighbors of patterns. That is to say, it is hard to distinguish differences between patterns with similar degree. In addition, there are many other ways to calculate pattern similarity, see [25–28] for details.

For the similarity of patterns in directed weighted graph data, similarity is affected by the direction of the edge and edges in different directions have an impact on its weight. Besides, each pattern has information such as out-degree and in-degree, out-weight, and in-weight, and the relationship between the pattern and its neighbors in different scales is complex. Therefore, the similarity measure of patterns in directed weighted graph data cannot start from a single direction. Generally speaking, the above measure of similarity has been used extensively. Nonetheless, there are still some inevitable limitations. These index that used mutual information are limited to the common neighbor structure or local information of patterns; so, it is easy to make the patterns of larger degree become the general patterns in the similarity calculation. Even if existing submethods simplify the measure of topological information, they ignore the directivity of the pattern's connection and its corresponding degree and weight diversity of the relationship between patterns. Under the circumstances, some edge information of pattern is lost, leading to their performance for calculating the similarity of patterns failing to get further enhancement. In particular, there may be a poor effect when the above indices are applied to ink prediction. To sum up, calculating similarity of patterns from the aspects of degree and weight diversity is still a hotspot [29–31].

In this paper, we aim at similarity of pattern in directed weighted graph data. To this, an extended version of the similarity measure approach from a relative entropy point of view is proposed. For more details, the comprehensive process can be considered as three stages. First, compute strength set. By using degree and weight of pattern's information in its first-order neighbors, four variables that contain the influence of topological information about degree and weight diversity are constructed. Second, generate probability set. To take advantage of the second-order neighbor information of patterns, Euclidean metric is presented to measure the similarity between pattern and its first-order neighbors. On this basis, the value of similarity is normalized to construct probability set of each pattern. Third, Quantify similarity of pattern. With the help of relative entropy, the dissimilarity of any two patterns is measured, and similarity can be gained subsequently. We numerically simulated the proposed similarity measure and verified its effectiveness and efficiency in similarity measure and link prediction. In this paper, there is a proposed relative entropy-based similarity for patterns in graph data with the following several contributions in mind.

(1) This paper presents a similarity measure based on relative entropy, which considers the information of second-order neighbors of patterns

(2) In the process of pattern's similarity measure, the proposed method considers both degree information and weight information

(3) Compared with most benchmark methods, the proposed similarity measure has a great advantage in measuring similarity of patterns and gets good performance the link prediction

To make a detailed description of the above proposed similarity approach, in this section, we will provide a brief introduction to the structure of this paper. Section 2 contains some preliminaries. Section 3 describes generation of strength set for patterns in detail. Section 4 proposes probability set calculated by similarity set. Section 5 constructs a measure to compute the similarity of pattern in graph data, and a novel algorithm is proposed. Section 6 carries out two type experiments to prove the effectiveness of the proposed method. Conclusion is given in section 7.

2. Preliminaries

In this section, we propose some basic concepts used in this paper, such as graph data [26], relative entropy [32], and pattern's neighbor [23].

2.1. Graph Data. A graph data G is defined as a set of patterns and a set of edges. Generally speaking, the so-named directed weighted graph data can be expressed as a 4-tuple $G = (V, E, D, W)$, formally, where

(i) $V = \{v_i | i = 1, 2, \cdots, n\}$ is the set of patterns, and $v_i \in V$ represents the i^{th} pattern

(ii) $E = \{e_{ij} | i, j = 1, 2, \cdots, n\}$ is the set of edges, and $e_{ij} \in E$ indicates the set of edges. Hereinto, $e_{ij} = 1$ if pattern v_i and v_j are connected; otherwise, $e_{ij} = 0$

(iii) $D = \{(d(v_i), d^+(v_i), d^-(v_i)) | i = 1, 2, \cdots, n\}$ is the set of corresponding weight with respect to patterns, thereinto $d^+(v_i)$ and $d^-(v_i)$ represent in-degree and out-degree of $v_i \in V$, respectively, and the value of them, take v_i for example, can be determined by equations

$$d^+(v_i) = \sum_{j=1}^{n} e_{ji},$$
$$d^-(v_i) = \sum_{j=1}^{n} e_{ji}. \tag{1}$$

Moreover, the degree $d(v_i)$ can be calculated by the sum of in-degree and out-degree, i.e.,

$$d(v_i) = d^+(v_i) + d^-(v_i) = \sum_{j=1}^{n} (e_{ij} + e_{ji}). \tag{2}$$

(iv) $W = \{(w(v_i), w^+(v_i), w^-(v_i)) | i = 1, 2, \cdots, n\}$ is the set of weights with respect to the corresponding edges. Analogously, $w(v_i), w^+(v_i), w^-(v_i)$ represent the weight, in-weight, and out-weight of pattern v_i, respectively. The value of in-weight and out-weight can be determined by following equations:

$$w^+(v_i) = \sum_{j=1}^{n} w_{ji},$$
$$w^-(v_i) = \sum_{i=1}^{n} w_{ij}, \tag{3}$$

Thereinto, w_{ij} represents weight on edge of v_i and v_j. Furthermore, the value of weight can be calculated by the sum of in-weight and out-weighted, i.e.,

$$w(v_i) = w^+(v_i) + w^-(v_i) = \sum_{j=1}^{n} (w_{ij} + w_{ji}). \tag{4}$$

2.2. Relative Entropy. As we known, relative entropy is an asymmetric measure and can be applied to measure the difference between two probability distributions. In general, its mathematical version can be expressed as

$$D_{\text{KL}}(P \| Q) = \sum_{i=1}^{m} P(i) \ln \frac{P(i)}{Q(i)}, \tag{5}$$

where P and Q are two probability distributions, and "m" in equation (5) represents the number of variables that P and Q depended on. Certainly, the greater value of $D_{\text{KL}}(P \| Q)$ reflects the smaller similarity of P and Q, and vice versa.

2.3. Pattern's Neighbor. For a graph data $G = (V, E, D, W)$, if there exists at least two patterns v_i and v_j such as $e_{ij} \neq 0$ or $e_{ji} \neq 0$. Then, one can say that v_i is the neighbor of v_j, and vice versa. All the neighbors of v_i constitute the so-named neighborhood with respect of v_j, in aspect of topological information. For the need of simplicity and uniformity, we summarize it as the following definition.

Definition 1. (first-order neighborhood). Given that $G = (V, E, D, W)$ is a directed weighted graph data, if $e_{ij} \neq 0$ or $e_{ji} \neq 0$ for $j = 1, 2, \cdots, n$, then the pattern v_j is a first-order neighbor of v_i. Certainly, all of the neighbors of v_i constitute the first-order neighborhood of it and can be expressed as

$$N(v_i) = \{v_j | e_{ij} \neq 0 \text{ or } e_{ji} \neq 0\}. \tag{6}$$

Generally speaking, if pattern v_i has $p$first neighbors, then $N(v_i)$ can be represented as $N(v_i) = \{v_i^1, v_i^2, \cdots, v_i^p\}$. Obviously, the elements of $N(v_i)$ reflect the topological information of v_i directly. For the case that $e_{ij} \neq 0$ and $e_{jk} \neq 0$ but $e_{ik} = 0$, how to depict the direct relationship of v_i and v_k in aspect of topological information is no longer an obvious question. To this, next definition gives the concept of second-order neighborhood to depict such situation.

Definition 2. (second-order neighborhood). Given that $G = (V, E, D, W)$ is a graph data and $v_i \in V$, the second-order neighborhood of a pattern v_i denoted as the set contained neighbors of its all first-order neighbors, which notes as $N(v_i, 2)$, which can be expressed as

$$N(v_i, 2) = \{N(v_i^1), N(v_i^2), \cdots, N(v_i^p)\}. \tag{7}$$

Definition 3. (local neighborhood). Given that $G = (V, E, D, W)$ is a graph data and $v_i \in V$, the so-named local neighborhood of v_i can be expressed as

$$L(v_i) = v_i \cup N(v_i) = \{v_i, v_i^1, \cdots, v_i^p\}, \tag{8}$$

where $v_i^k \in N(v_i)$ is the k^{th} first order neighbor of v_i, for $k = 1, 2, \cdots, p$.

3. Degree and Weight-Based Pattern's Strength Set

In this section, we investigate the problem of how to construct the pattern's strength set in terms of degree and weight.

For any pattern v_i in $G = (V, E, D, W)$, its first order neighborhood $N(v_i)$ depends on the corresponding topological connection. Whatever the connection, the topological information for each pattern can be described by four variables: in-degree $d^+(v_i)$, out-degree $d^-(v_i)$, in-weight $w^+(v_i)$, and out-weight $w^-(v_i)$.

In what follows, we introduce the concept of strength set for any pattern v_i in a graph data G.

Definition 4. (strength set). Given that $G = (V, E, D, W)$ is a graph data, for any pattern $v_i \in G$, its strength set $U(v_i)$ can be expressed by following equation:

$$U(v_i) = \left(u_{v_i}, u_{v_i^1}, u_{v_i^2}, \cdots, u_{v_i^p}\right), \tag{9}$$

where $k = 1, 2, \cdots, p$ and $p = |N(v_i)|$. Each variable in $U(v_i)$ contains four strength values consisting of in-degree, out-degree, in-weight, and out-weight, take v_i for example, $u_{v_i} = (u_{d^+}(v_i), u_{d^-}(v_i), u_{w^+}(v_i), u_{w^-}(v_i))$, in which case

(i) $u_{d^-}(v_i)$ represents the strength of out-degree and can be computed by equation

$$u_{d^-}(v_i) = \frac{d^-(v_i)}{d(v_i)} \cdot \frac{d^-(v_i)}{\sum_{i=1}^{n} d^-(v_i) + 1}. \tag{10}$$

(ii) $u_{d^+}(v_i)$ represents the strength of in-degree and can be computed by equation

$$u_{d+}(v_i) = \frac{d^+(v_i)}{d(v_i)} \cdot \frac{d^+(v_i)}{\sum_{i=1}^{n} d^+(v_i) + 1}. \tag{11}$$

(iii) $u_{w^-}(v_i)$ represents the strength of out-weight and can be computed by equation

$$u_{w^-}(v_i) = \frac{w^-(v_i)}{w(v_i)} \cdot \frac{w^-(v_i)}{\sum_{i=1}^{n} w^-(v_i) + 1}. \tag{12}$$

(iv) $u_{w+}(v_i)$ represents the strength of in-weight and can be computed by equation

$$w_w + (v_i) = \frac{w^+(v_i)}{w(v_i)} \cdot \frac{w^+(v_i)}{\sum_{i=1}^{n} w^+(v_i) + 1}. \tag{13}$$

Analogously, $u_{v_i^k}$ represents the strength of k^{th} first-order neighbor to v_i, which can be calculated by equations mentioned above. One can find that the above proposed strength fully depicts personal properties and topological information with respect to corresponding its first-order neighbors.

As discussed above, take v_i and v_j for example, if v_i and v_j are two different patterns, then $N(v_i) \neq N(v_j)$ is nothing unusual to some extent. In particular, there would be one extreme situation that $N(v_i) \neq N(v_j)$ if $v_i \neq v_j$, for $i, j = 1, 2, \cdots, n$ and $i \neq j$.

By making a deeper investigation of relative entropy, one can see that the patterns with more neighbors will lose certain information, for it only calculates the value of nonzero elements in probability set, and the information of nonzero elements in the probability set of its corresponding patterns will also be lost. Considering this deficiency, we introduce a concept, the scale of strength set, to depict the strength set. Before doing this, we suppose that for a graph data G, there exists at least one pattern v_i that having the most neighbors, in which we denote the number of it as $n_p = \max_{v_i \in V} |N(v_i)|$. To this, for the pattern v_i with $|N(v_i)| = p_1$ and pattern v_j with $|N(v_j)| = p_2$, we take the following cases into consideration:

Case 1. If $p_1 = p_2 = p = n_p$, then the $U(v_i)$ and $U(v_j)$ can be represented as

$$\begin{aligned} U^{new}(v_i) &= \left(u_{v_i}, u_{v_i^1}, u_{v_i^2}, \cdots, u_{v_i^p}\right), \\ U^{\text{new}}\left(v_j\right) &= \left(u_{v_j}, u_{v_j^1}, u_{v_j^2}, \cdots, u_{v_j^p}\right). \end{aligned} \tag{14}$$

Case 2. If $p_1 = p_2 = p < n_p$, then the $U(v_i)$ and $U(v_j)$ can be changed into

$$\begin{aligned} U^{new}(v_i) &= \left(u_{v_i}, u_{v_i^1}, u_{v_i^2}, \cdots, u_{v_i^p}, \overbrace{0, 0, \cdots, 0}^{n_p - p}\right), \\ U^{new}\left(v_j\right) &= \left(u_{v_j}, u_{v_j^1}, r_{v_j^2}, \cdots, u_{v_j^p}, \overbrace{0, 0, \cdots, 0}^{n_p - p}\right). \end{aligned} \tag{15}$$

In other words, append $n_p - p$ zeros, i.e., $0 = (0, 0, 0, 0)$ to the end of $U(v_i)$ and $U(v_j)$.

Case 3. If $p_1 < p_2 < n_p$, we append $p_2 - p_1$ average strength values $u_{v_i^*}$ of v_i to the end of $U(v_i)$ by the equation

$$u_{v_i^*} = \frac{u_{v_i^1} + u_{v_i^2} + \cdots + u_{v_i^{p_1}}}{p_1}. \tag{16}$$

Generally speaking, the $U(v_i)$ and $U(v_j)$ can be changed into

$$\begin{aligned} U^{\text{new}}(v_i) &= \left(u_{v_i}, u_{v_i^1}, u_{v_i^2}, \cdots, u_{v_i^{p_1}}, \overbrace{u_{v_i^*}, \cdots, u_{v_i^*}}^{p_2 - p_1}, \overbrace{0, 0, \cdots, 0}^{n_{\text{p}} - p_2}\right), \\ U^{\text{new}}\left(v_j\right) &= \left(u_{v_j}, u_{v_j^1}, u_{v_j^2}, \cdots, u_{v_j^{p_2}}, \overbrace{0, 0, \cdots, 0}^{n_{\text{p}} - p_2}\right), \end{aligned} \tag{17}$$

where the insufficient $p_2 - p_1$ locations of $U(v_i)$ will be appended by strength values $u_{v_i^*}$ calculated with the help of equation (16), and the rest location of $U(v_i)$ and $U(v_j)$ will be appended by $p - p_2$ zeros, i.e., $0 = (0, 0, 0, 0)$.

Case 4. If $p_2 < p_1 < n^p$, the $U(v_i)$ and $U(v_j)$ can be changed into

$$U^{\text{new}}(v_j) = \left(u_{v_j}, u_{v_j^1}, u_{v_j^2}, \cdots, u_{v_j^{p_2}}, \overbrace{u_{v_j^*}, \cdots, u_{v_j^*}}^{p_1 - p_2}, \overbrace{0, 0, \cdots, 0}^{n_p - p_1}\right),$$

$$U^{\text{new}}(v_i) = \left(u_{v_i}, u_{v_i^1}, u_{v_i^2}, \cdots, u_{v_i^{p_2}}, \overbrace{0, 0, \cdots, 0}^{n_p - p_1}\right), \tag{18}$$

where strength value $u_{v_j^*}$ of v_j can be calculated as the following equation:

$$u_{v_j^*} = \frac{u_{v_j^1} + u_{v_j^2} + \cdots + u_{v_j^{p_2}}}{p_2}. \tag{19}$$

4. Generating Probability Set

Relative entropy is applied to compare the difference of two probability set. To some extent, the similarity can be regarded as the difference. For this, we try to calculate the similarity between patterns in aspect of relative entropy. Before do this, how to construct the probability set of each pattern $v_i \in G$ constitutes the first step of similarity measure.

We have known that the strength set $U(v_i)$, take v_i for example, and its one order-neighbors can be determined in terms of degree and weight. To make full use of relative entropy for the purpose of similarity measure, in what follows, we construct an approach to generate the probability set of patterns $v_i \in V$ for $i = 1, 2, \cdots, n$. Each strength value of the j first-order neighbors is composed by four variables; here, the Euclid metric can be applied to compute the similarity between v_i and $v_j \in N(v_i)$ for $j = 1, 2, \cdots, p$ with respect to its strength set. The concrete formula can be depicted as the following equation:

$$s\left(v_i, v_i^j\right) = \sqrt{< u_{v_i}, u_{v_i^j} >}. \tag{20}$$

Obviously, the value $s(v_i, v_i^j)$ describes the similarity between v_i and its j^{th} first-order neighbor, and it is only a local description in view point of $N(v_i)$. With the help of equation (20), we can make a global description of the similarity of pattern v_i by the following equation:

$$S(v_i) = \left[s\left(v_i, v_i^1\right), s\left(v_i, v_i^2\right), \cdots, s\left(v_i, v_i^p\right)\right]. \tag{21}$$

Up to now, the caring thing, that is, creating probability set, can be realized by the following equation:

$$P(v_i) = \left[p\left(v_i, v_i^1\right), p\left(v_i, v_i^2\right), \cdots, p\left(v_i, v_i^p\right)\right], \tag{22}$$

where

$$p\left(v_i, v_i^j\right) = \frac{s\left(v_i, v_i^j\right)}{\sum_{k=1}^{p} s\left(v_i, v_i^k\right)}. \tag{23}$$

By above aforementioned, the relative entropy between the probability set $P(v_i)$ and $P(v_j)$, take v_i, v_j in G for example, it can be determined by the equation

$$D_{\text{KL}}\left(P(v_i) \| P\left(v_j\right)\right) = \sum_{k=1}^{m'} p\left(v_i, v_i^k\right) \ln \frac{p\left(v_i, v_i^k\right)}{p\left(v_j, v_j^k\right)}, \tag{24}$$

where m' represents the maximal neighbors of patterns v_i and v_j; that is, $m' = \max\{|N(v_i)|, |N(v_j)|\}$.

It can be analyzed that, in the process of calculating relative value of pattern v_i and v_j, strength set of first-order neighbors of pattern v_i and v_j is constructed, which contain second-order neighbor information. That is to say, with the help of Euclidean metric, the information of pattern's second-order neighbors is indirectly used during similarity calculation process.

5. Similarity and Algorithm

The calculation of relative entropy among the patterns has been discussed in detail. In this section, the calculated value of relative entropy will be used to compute the similarity between patterns. And then, an algorithm is proposed.

5.1. Quantify Similarity of Pattern. From the process mentioned above, relative entropy of any two patterns is obtained based on the sorted probability sets. Therefore, the relative entropy matrix R of graph data with respect to any two patterns can be represented as

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}. \tag{25}$$

And then, the similarity matrix S of graph data G can be given as follows.

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}. \tag{26}$$

For the value of relative entropy is asymmetric, take v_i and v_j for example, both $D_{\text{KL}}(P(v_i)\|P(v_j))$ and $D_{\text{KL}}(P(v_j)\|P(v_i))$ describe dissimilarity of pattern v_i and v_j. To obtain more accurate similarity of pattern, the value of relative

Input: A directed weighted graph data G.
Output: Similarity matrix S of patterns in G.
1: **for** each $v_i \in G$**do**
2: Calculate first-order neighbors $L(v_i)$
3: Calculate $U(v_i)$ by equation (10)- (13)
4: **end**
5: **for** each $v_i \in G$ with p_1 neighbors and $v_j \in G$ with p_2 neighbors **do**
6: **if** $p_1 > p_2$ **then**
7: $(p_1 - p_2)u_{v_j^*} \longrightarrow U(v_j)$
8: **else**
9: $(p_2 - p_1)u_{v_i^*} \longrightarrow U(v_i)$
10: **end**
11: **end**
12: **for** $v_i \in V, v_i^j \in L(v_i)$ **do**
13: Compute $S(v_i)$ by equation (20) and (21)
14: Compute $P(v_i)$ by equation (22)
15: **end**
16: **for** each $v_i, v_j \in G$ **do**
17: Compute $D_{KL}(P(v_i)\|P(v_j))$ by equation (24)
18: Compute entropy matrix R by equation (25)
19: Compute similarity matrix s_{ij} by equation (27)
20: **end**
21: **return** Similarity matrix S

Algorithm 1: Relative entropy-based similarity for patterns in graph data.

entropy can be calculated by taking both $D_{\mathrm{KL}}(P(v_i)\|P(v_j))$ and $D_{\mathrm{KL}}(P(v_j)\|P(v_i))$ into consideration, and the specific calculations of it are shown as follows:

$$s_{ij} = 1 - \frac{r_{ij}}{\max\ (R)}, \tag{27}$$

$$r_{ij} = \frac{D_{KL}\left(P(v_i)\middle\|P\left(v_j\right)\right) + D_{KL}\left(P\left(v_j\right)\|P(v_i)\right)}{2}. \tag{28}$$

5.2. Algorithm. With the purpose of a better understanding for the proposed pattern similarity measure, this section will give an algorithm containing detailed description of this similarity measure. Notice that for briefness, "Relative entropy-based similarity for patterns in graph data" can be summarized as "RESG." In terms of this algorithm, the similarity of any two patterns will be computed, after which the most similar patterns can be obtained. One can easily see that there are four states of this algorithm. The input of the RESG algorithm is a weighted directed graph data G, and the output is a matrix S composed of similarity between any two patterns in G.

The first state of the RESG algorithm is lines 1-4, strength set U of each pattern in G is generated, and each strength set has four variables in terms of in-degree, out-degree, in-weight, and out-weight. The second state is lines 5-11, to fully utilize information of pattern's first-order neighbors, the pattern with less neighbors will append average strength value in the end of strength set. The third state of the RESG algorithm is lines 12-15, the similarity between patterns and its one-order will be computed, and similarity set is generated. With the help of similarity set, pattern's probability set will be obtained. It is not hard to find; during the process of generating similarity set, the information of pattern's second-order neighbors will be used indirectly. The last state is lines 16-20; by taking the above information into account, the relative entropy and similarity of patterns are measured.

Table 1: Experimental environment.

Parameter	Parameter value
RAM	16 GB
Speed	2.10 GHz
Programming	MATLAB 2018a
CPU	AMD Ryzen 54600 U
GUP	AMD Radeon Graphics
System	Windows 10 system with 8 cores

6. Experimental Materials

In this section, we introduce some experimental materials such as experimental environment, the graph data used in experiment, and benchmark algorithms. The experimental environment we used is listed in Table 1.

6.1. Data. This subsection will give a detailed description about the directed weighted graph data used in experiments.

A synthetic graph data *Datal* generated by means of the graph generator Gephi will be applied in first experiment. *Datal* contains 21 patterns and 31 edges and can be used

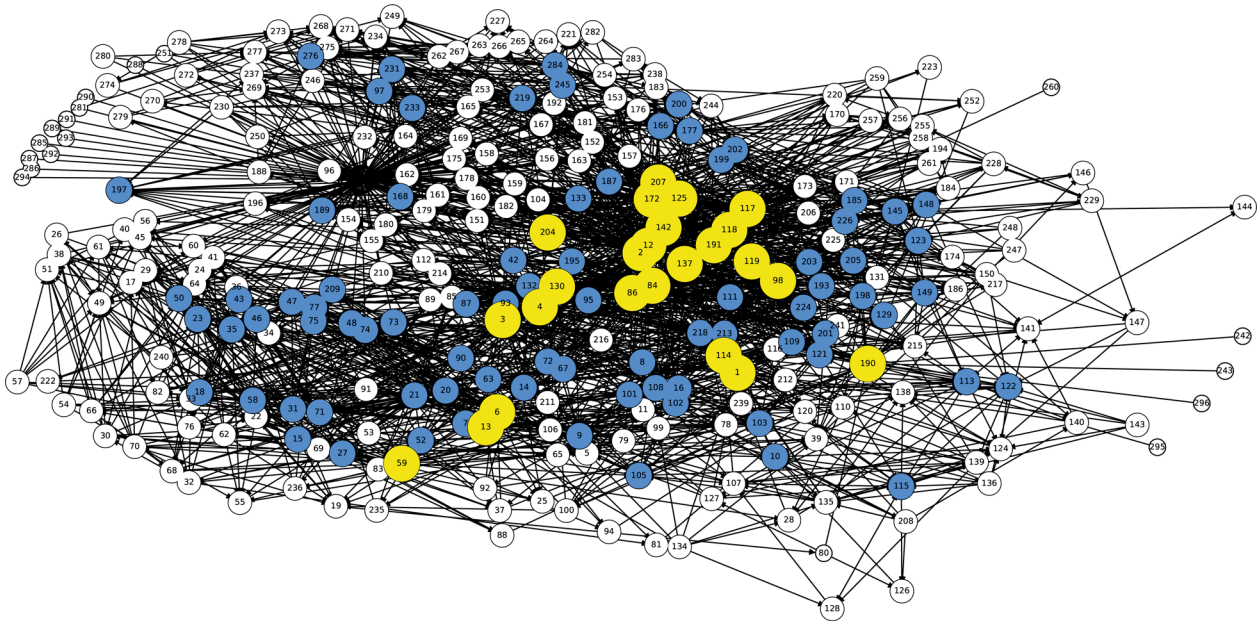


Figure 1: *Data2*.

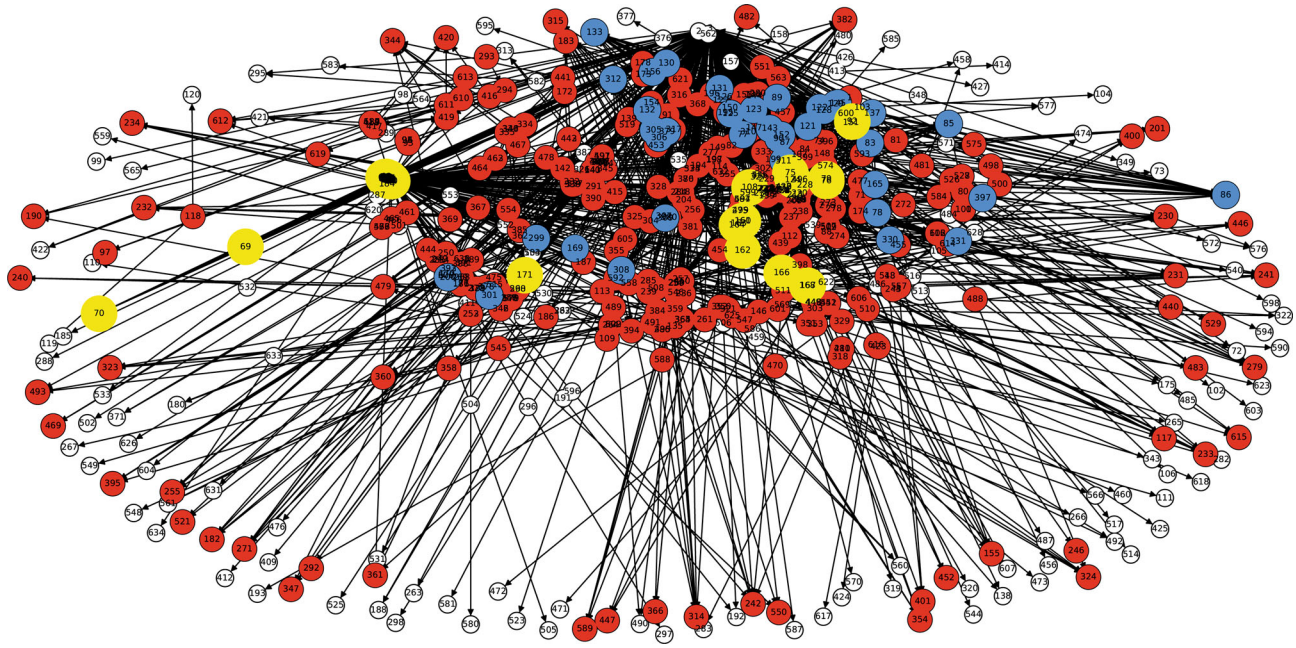


Figure 2: *Gene*.

to illustrate the feasibility of the proposed RESG algorithm in the following illustrative example.

The following is a detailed description of graph data used in second experiment. *Data2* and *Gene* [33] will be used to demonstrate the similarity of our proposed RESG index and other similarity measures. For the edge connections of *Data 2* and *Gene* [33], see Figures 1 and 2 for detail. *Stmarks*, *FWEW*, *FWMW*, *FWFW*, *Celegans*, and *Email167* are directed weighted graph data collected from Stanford Dataset. Each of them will be used to show the effectiveness of the proposed RESG algorithm in link prediction. The topology information of these eight graph data are shown in Table 2, where n is the number of patterns, m is the number of edges, $\langle d \rangle$ is the average shortest distance, $\langle \rho \rangle$ is the density, $\langle k \rangle$ is the average degree, and $\langle c \rangle$ is the clustering coefficient.

6.2. Benchmark Algorithms. Here, we introduce several benchmark pattern's similarity indices, which are usually used for similarity measure and link prediction. Adamic-

Table 2: Topological properties of graph data.

Graph data	n	m	$\langle d \rangle$	$\langle \rho \rangle$	$\langle k \rangle$	$\langle c \rangle$
Data2	297	2358	6.0000	0.0492	14.5000	0.3092
Gene	636	3959	4.0000	0.0226	12.0000	0.5701
Stmarks	54	350	3.0000	0.2446	12.9630	0.4128
EW	69	880	3.0000	0.3751	25.5072	0.5521
FWMW	97	1446	3.0000	0.3106	29.9144	0.4683
FWFW	128	2075	3.0000	0.2540	32.4129	0.3364
Celegans	297	2148	2.4550	0.0489	14.4646	0.3079
Email167	167	3250	5.0000	0.2345	38.9222	0.6864

Adar (AA) [10], weighted Adamic-Adar (WAA) [11], local relative entropy (LRE) [21], common neighbors (CN) [13], Katz [17], local path (LP) [12], and Local Random Walk (LRW) [16] are often used for the purpose of comparing results with the RESG algorithm. The basic definitions of these indexes are given below.

AA index is the extended version of CN index, which is defined as

$$s_{\mathrm{ij}}^{\mathrm{AA}} = \sum_{v_z \in N(v_i) \cap N(v_j)} \frac{1}{\log d(v_k)}. \tag{29}$$

WAA index is the weighted version of AA index, which is defined as

$$s_{\mathrm{ij}}{}^{\mathrm{WAA}} = \sum_{v_z \in N(v_i) \cap N(v_j)} \frac{w(v_i) + w(v_j)}{\log (1 + a)}, \tag{30}$$

where a may be smaller than 1; so, we use $\log (1 + a)$ to avoid a negative value.

CN index directly takes the number of all common neighbors between patterns as similarity into consideration, which is defined as

$$s_{ij}{}^{CN} = \left| N(v_i) \cap N(v_j) \right|. \tag{31}$$

LRE index is a similarity measure based on relative entropy and local structure of patterns, which is defined as

$$s_{\mathrm{ij}}^{\mathrm{LRE}} = 1 - \frac{D_{\mathrm{KL}}(P_i \| P_j) + D_{\mathrm{KL}}(P_j \| P_i)}{m \left(D_{\mathrm{KL}}(P_i \| P_j) + D_{\mathrm{KL}}(P_j \| P_i) \right)}, \tag{32}$$

whereinto

$$D_{\mathrm{KL}}(P_i \| P_j) = \sum_{k=1}^{\Delta(G)} p_i(k) \ln \frac{p_i(k)}{p_j(k)},$$

$$p_i(k) = \begin{cases} \dfrac{d(v_i)}{\sum_{v_k \in N(v_i)} d(v_k)}, & k \le d(v_i), \\ 0, & d(v_i) + 1 \le k \le \Delta(G). \end{cases} \tag{33}$$

Hereinto, $\Delta(G)$ is the maximum degree of the graph data, and G, P_i is the probability set of pattern v_i with respect to degree.

Katz index is based on the global information of graph data, which is defined as

$$s_{\mathrm{ij}}^{\mathrm{Katz}} = \sum_{k=1}^{\infty} \cdot \beta^l \cdot \left| \mathrm{path}\, s_{\mathrm{ij}} < l > \right|, \tag{34}$$

where $|\mathrm{path}\, s_{ij} < l > |$ represents the set of all paths with distance l between pattern v_i and v_j, β is the damping factor used to control the path weight.

LP index considers the third-order paths on the basis of common neighbors, which is defined as

$$s_{\mathrm{ij}}{}^{\mathrm{LP}} = A^2 + \alpha A^3, \tag{35}$$

where A is the adjacency matrix of graph data [34], $(A^3)_{ij}$ represents the number of paths with length of 3 between patterns v_i and v_j, and α is adjustable parameter.

LRW index is proposed based on the local random walk of particles between two patterns, which is defined as

$$s_{\mathrm{ij}}{}^{\mathrm{LRW}} = \frac{d(v_i)}{2 \cdot |E|} \cdot \pi_{ij}(t) + \frac{d(v_j)}{2 \cdot |E|} \cdot \pi_{ji}(t), \tag{36}$$

where $|E|$ is the number of the edges in the graph data, $\pi_{ij}(t)$ is obtained according to the density vector evolution equation: $\vec{\pi}_i(t+1) = P^T \cdot \vec{\pi}_i(t)$, the P is the transition probability matrix, and T is the matrix transpose.

7. Experimental Analysis

In this section, we evaluated the proposed RESG index into different real-world graph data, and two different forms of experiments are used to demonstrate experimental results, which aims to further prove the effectiveness and efficiency of proposed RESG index.

7.1. Illustrative Example. *Data1* is used to illustrate the proposed RESG index, for the edge connections of *Data1*, see Figure 3 for detail. Taking pattern v_{15} and v_{10} for example,

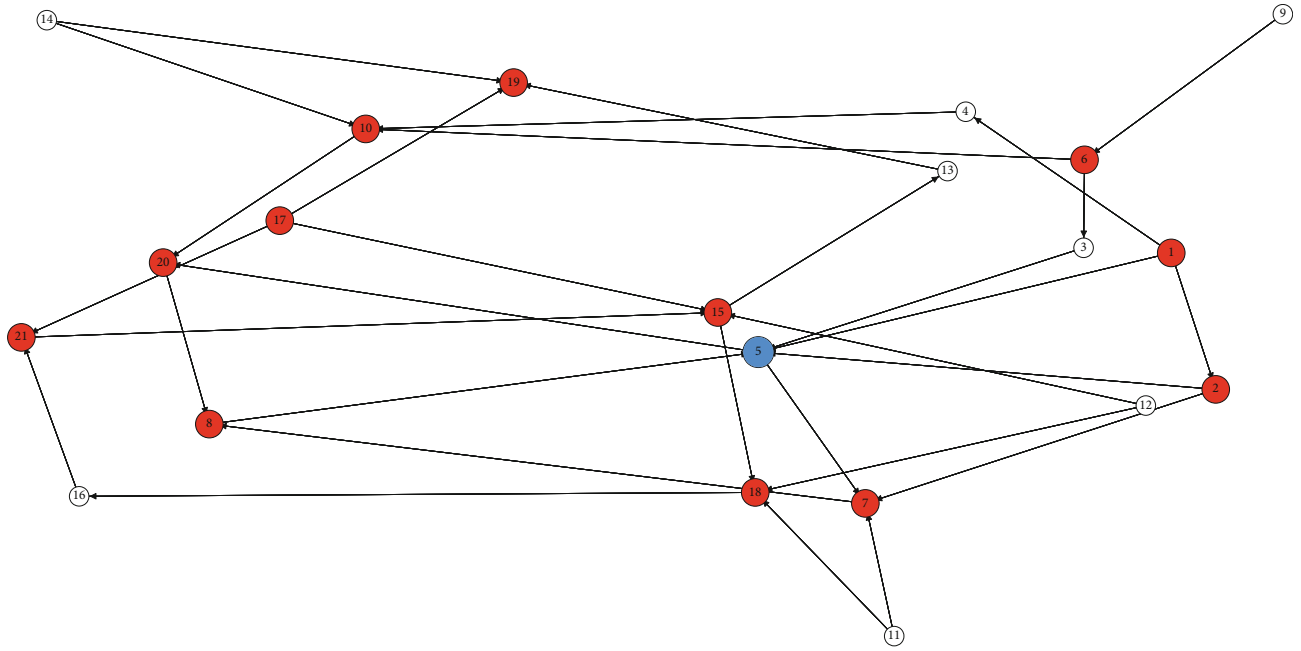


Figure 3: *Data1*.

Table 3: The $L(v_i)$ and $U(v_i)$ of pattern v_{15}.

Serial number	k_{out}	k_{in}	w_{out}	w_{in}	$u_{k_{\text{out}}}$	$u_{k_{\text{in}}}$	$u_{w_{\text{out}}}$	$u_{w_{\text{in}}}$
v_{15}	2	3	5	10	0.0727	0.1800	0.0556	0.2564
v_{12}	2	0	5	0	0.1818	0.0000	0.1667	0.0000
v_{13}	1	1	2	3	0.0455	0.0500	0.0267	0.0692
v_{17}	1	3	0	12	0.2727	0.0000	0.0333	0.0000
v_{18}	1	3	2	6	0.0227	0.2250	0.0667	0.1731
v_{21}	1	2	3	6	0.0303	0.1333	0.0333	0.1538

in terms of RESG index, next, we deal with the problem of pattern similarity step by step.

Firstly, we find pattern's first-order neighbors of them, respectively, and put them in $L(v_i)$, and relevant strength value about topological information $u_{d_-}, u_{d_+}, u_{w_-}, u_{w_+}$ of v_{15} and v_{10} is calculated and shown in Tables 3 and 4, respectively. However, it can be easily found that $d_{15} = 5$ and $d_{10} = 4$. Based on this, a pattern v_0 with the average value of v_{10} for $u_{d_-}, u_{d_+}, u_{w_-}$ and u_{w_+} is added as the one-neighbor of v_{10}. After that, the neighbors of the two patterns reached the same number, which avoided the partial information loss of v_{15} in the subsequent calculation of relative entropy.

Secondly, the similarity sets are generated in the process of calculating the similarity between patterns and its first-order neighbors, and the details of $S(v_{15})$ and $S(v_{10})$ are shown as

$$S(v_{15}) = [0.5115, 0.2313, 0.7248, 0.1860, 0.1225], \quad (37)$$

$$S(v_{10}) = [0.2821, 0.5212, 0.6099, 0.3965, 0.3533]. \quad (38)$$

The details of probability set based on strength set of v_{15} and v_{10} can be calculated and arranged each element in descending order, which can be shown as

$$\begin{aligned} P(v_{15}) &= [0.4801, 0.2880, 0.1302, 0.1047, 0.0690], \\ P(v_{10}) &= [0.2820, 0.2410, 0.1833, 0.1633, 0.1304]. \end{aligned} \quad (39)$$

Then, with the help of equation (24), the relative entropy $r_{10,15}$ of pattern v_{15} and v_{10} can be computed as follows.

$$r_{10,15} = \frac{D_{\text{KL}}(P(v_{15}) \| P(v_{10})) + D_{\text{KL}}(P(v_{10}) \| P(v_{15}))}{2} = 0.0986. \quad (40)$$

Finally, by computing pattern's similarity of the graph data G, the maximum value of pattern similarity can be found from Figure 4; in terms of equation (25), similarity of v_{15} and v_{10} is 0.8901. Obviously, the similarity calculation process of v_{15} and v_{10} can help better understand RESG index. The details of relevance matrix of graph data G are

Table 4: The $L(v_i)$ and $U(v_i)$ of pattern v_{10}.

Serial number	k_{out}	k_{in}	w_{out}	w_{in}	$u_{k_{out}}$	$u_{k_{in}}$	$u_{w_{out}}$	$u_{w_{in}}$
v_{10}	1	3	1	8	0.0313	0.2500	0.0056	0.4183
v_4	1	2	2	4	0.0417	0.1481	0.0333	0.1569
v_6	2	1	5	1	0.1667	0.0370	0.2083	0.0098
v_{14}	2	0	6	0	0.2500	0.0000	0.3000	0.0000
v_{20}	1	2	5	3	0.0417	0.1481	0.1563	0.0662
v_0					0.1063	0.1166	0.1407	0.1302

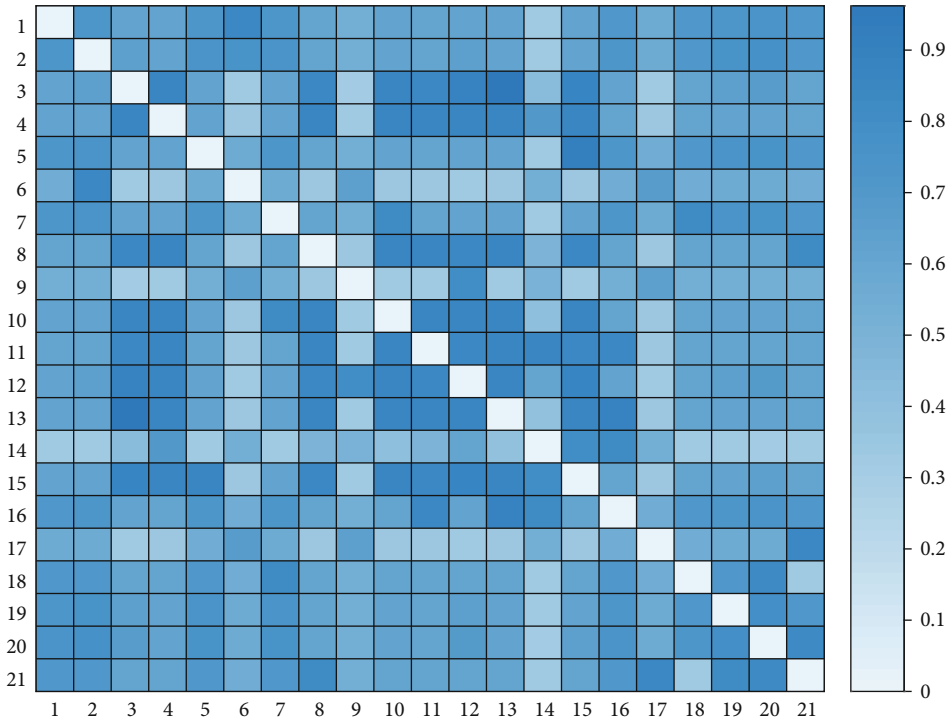


Figure 4: The relevance value and similarity of G.

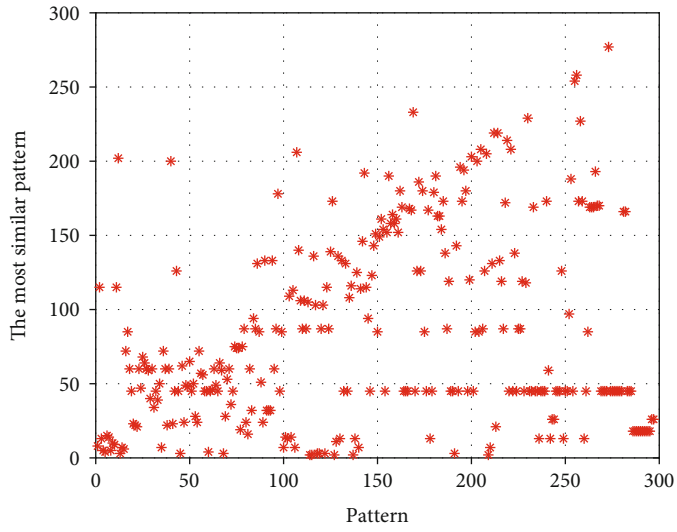
Table 5: The most similar pattern of each pattern in G.

Pattern	Most similar pattern	Pattern	Most similar pattern	Pattern	Most similar pattern
v_1	v_6	v_8	v_{21}	v_{15}	v_5
v_2	v_6	v_9	v_{12}	v_{16}	v_{13}
v_3	v_{13}	v_{10}	v_7	v_{17}	v_{21}
v_4	v_3	v_{11}	v_{16}	v_{18}	v_7
v_5	v_{15}	v_{12}	v_{10}	v_{19}	v_{21}
v_6	v_1	v_{13}	v_3	v_{20}	v_{21}
v_7	v_{18}	v_{14}	v_{16}	v_{21}	v_{20}

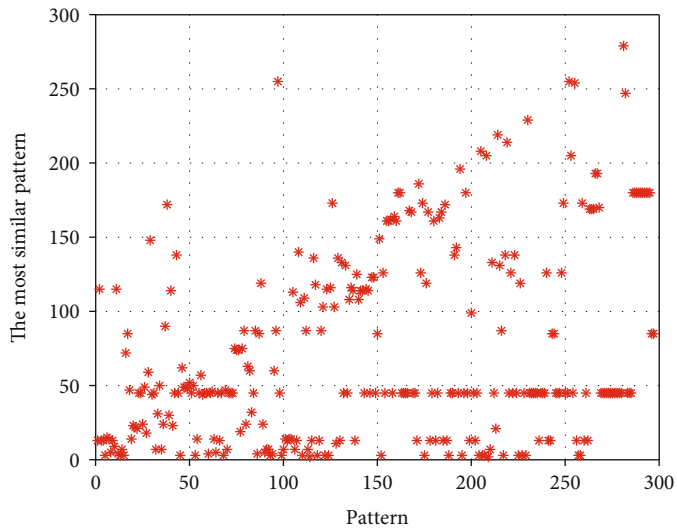
shown in Figure 4, and the most similar pattern in G is shown in Table 5.

According to Figure 3, we can find that compared with patterns v_{10}, v_5 and v_{15}, they have more similar topological structures. Depending on Table 5, the most similar pattern of v_{15} is exactly identified as pattern v_5. Illustrative example given shows that RESG index is simple, efficient, and reliable with highly satisfactory accuracy.

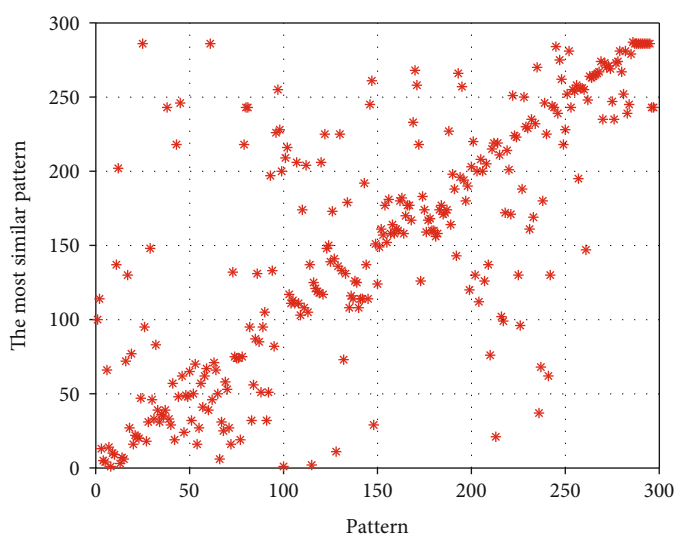
7.2. Result Analysis. To further illustrate the efficiency of the proposed RESG algorithm in measuring pattern's similarity, this subsection gives comparative experiments with serval



(a) AA

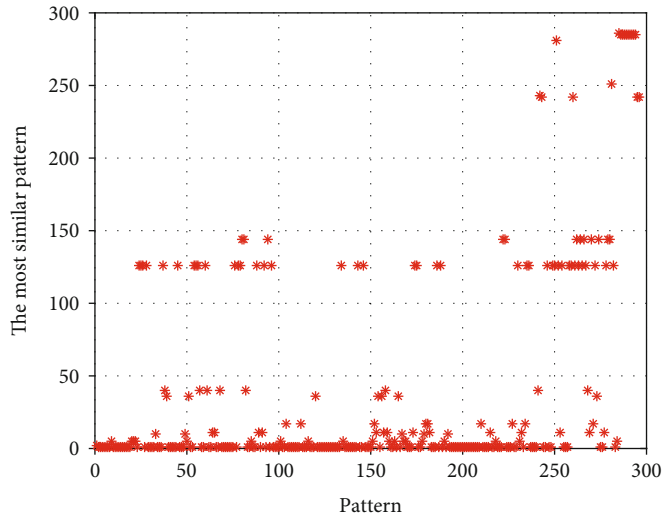


(b) WAA

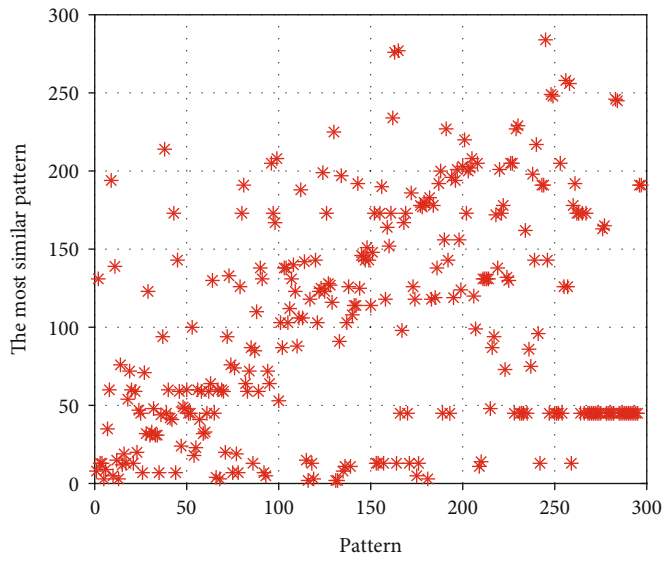


(c) CN

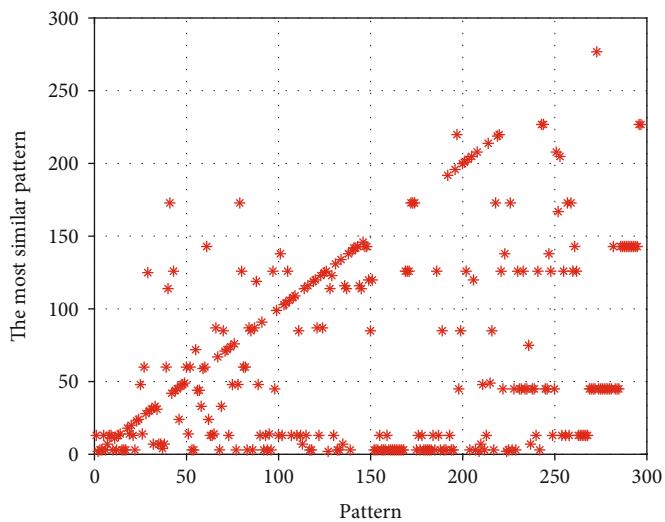
Figure 5: Continued.



(d) LRE

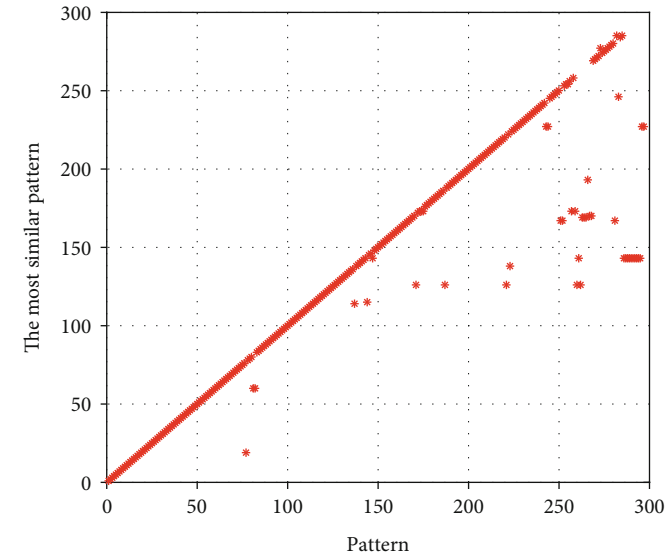


(e) Katz

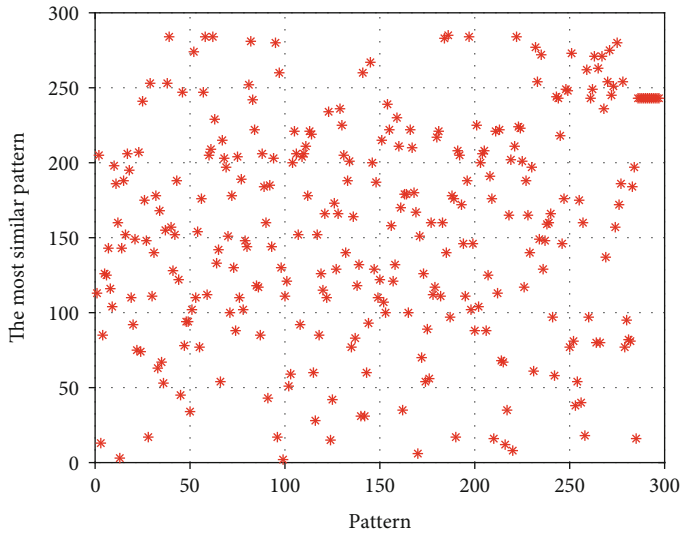


(f) LRW

Figure 5: Continued.



(g) LP



(h) RESG

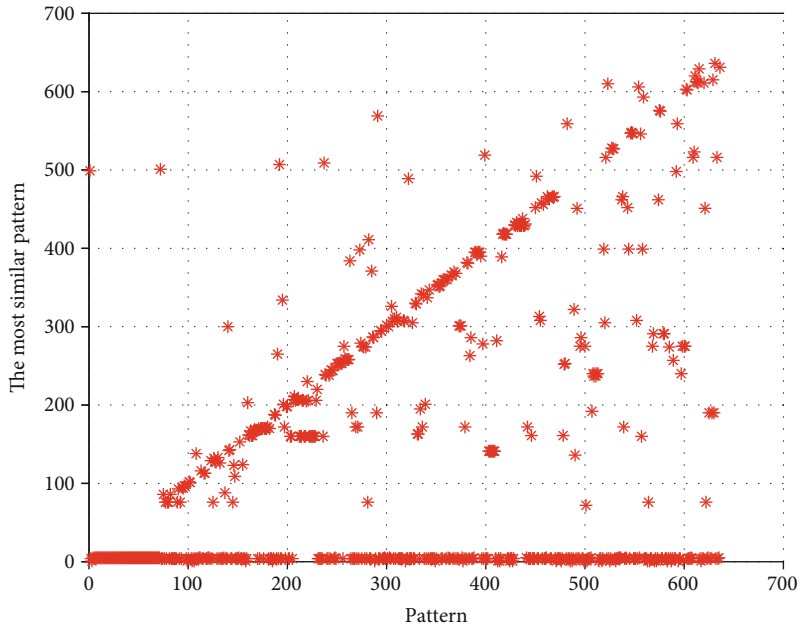
Figure 5: The scatter plots under *Data2*.

proposed similarity measures. In order to make a detailed description of experimental results, two ways are given. The first way of comparative experiment is to show the experimental results through scatter plots and table of the most similar pattern. The second way is to demonstrate the effectiveness by applying the RESG to link prediction.

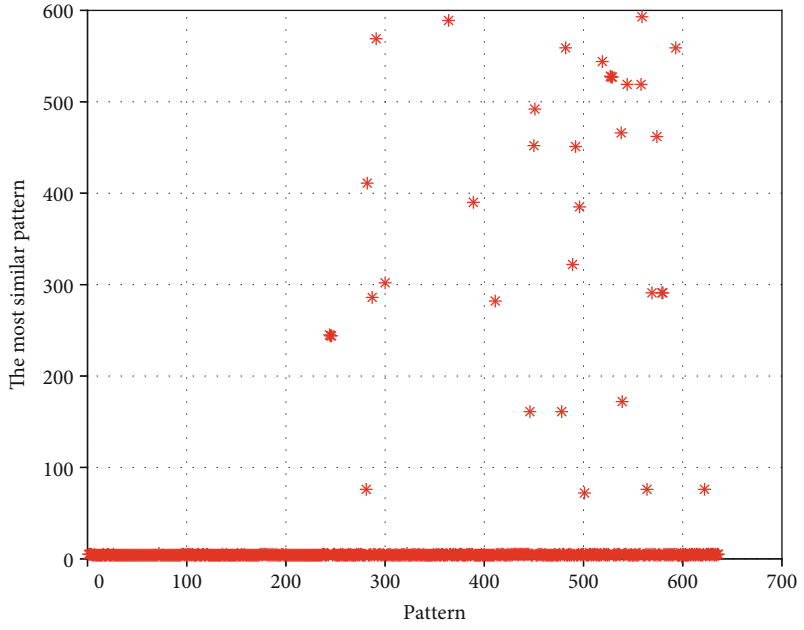
The scatter plot reflects the distribution of similarity between patterns. For example, if the most similar pattern of v_1 is v_{12}, then there exists draw points on $(1, 12)$ in plane. There is a good similarity measure, whose scatter plot is dispersed on the plane, rather than concentrated on both sides of diagonal. The reason is that if the points are concentrated on both sides of the diagonal line, it shows that this method is easier to identify its neighbors as most similar patterns, which is not accurate enough. In the following, under *Data2* and *Gene*, the scatter plots of the proposed RESG index and other seven similarity indices are used to further validate the performance of similarity measure, which are vividly shown in Figures 5 and 6, respectively.

Figures 5(a), 6(a), 5(b), 6(b), 5(c), and 6(c) show scatter plots formed by AA index, WAA index, and CN index, respectively. As we can see, the most similar patterns are concentrated near to diagonal. There is no denying that these three indices are low computational complexity; nevertheless, it uses very limited information. Generally speaking, similarity is determined by the number of common neighbors between patterns. Accordingly, the most similar patterns are distributed near the corresponding patterns. Although the symmetry of patterns is good, it is difficult to accurately describe the similarity between patterns when only one path is considered.

Figures 5(d) and 6(d) show scatter plots formed by LRE, respectively. It can be found that the most similar patterns are not distributed near to diagonal nevertheless. From

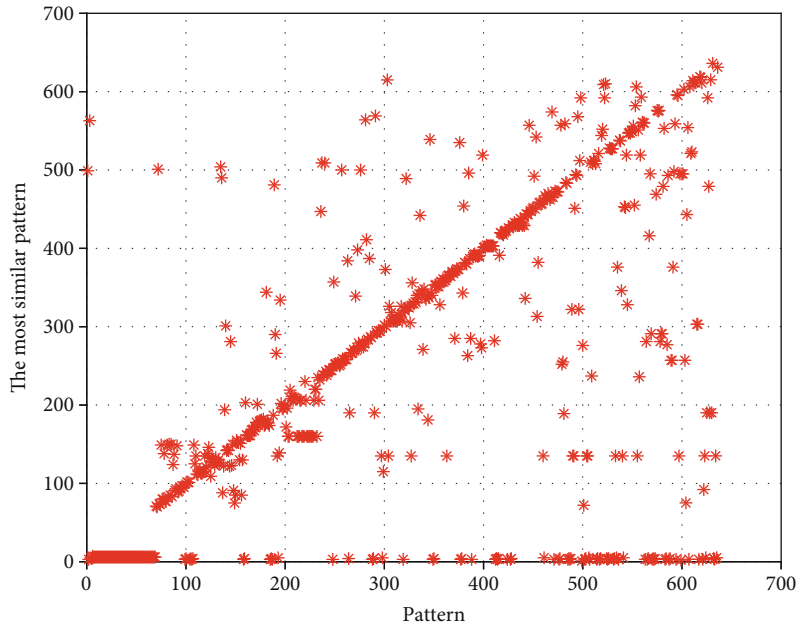


(a) AA

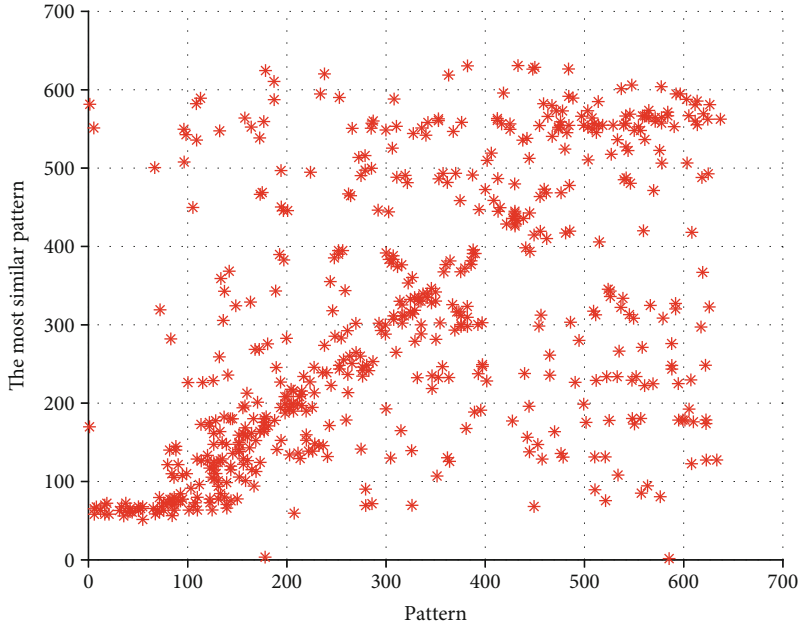


(b) WAA

Figure 6: Continued.

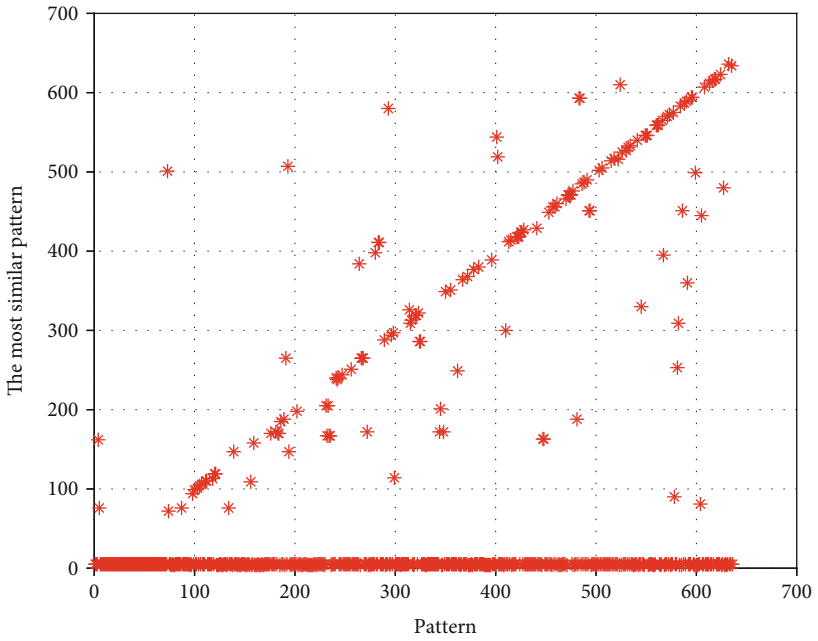


(c) CN

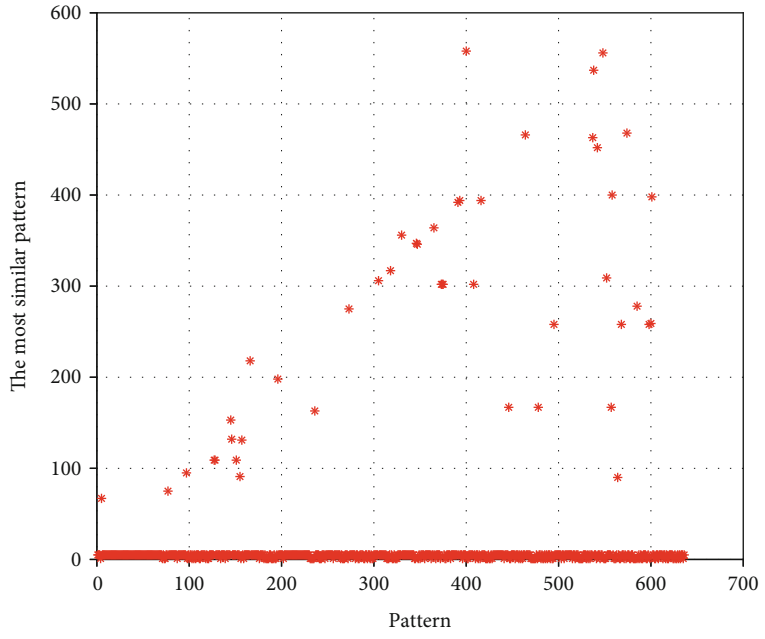


(d) LRE

Figure 6: Continued.

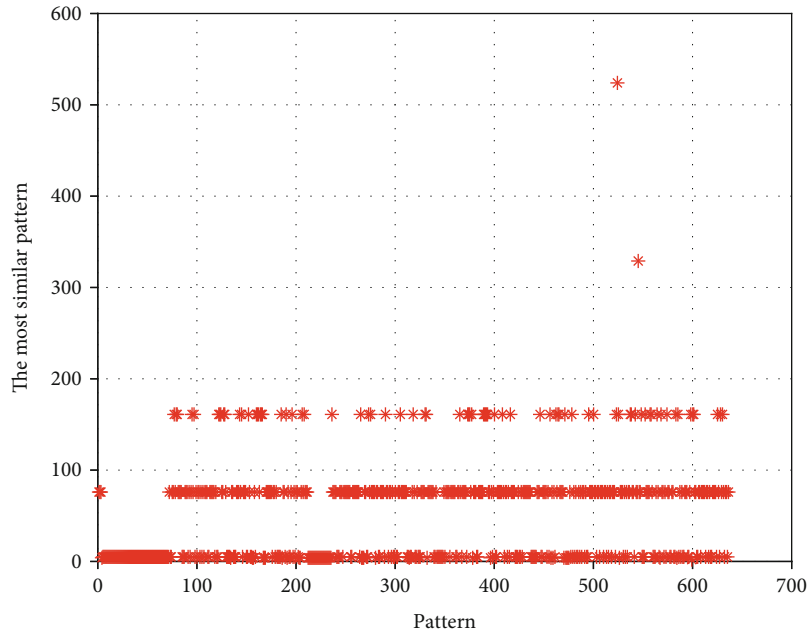


(e) Katz

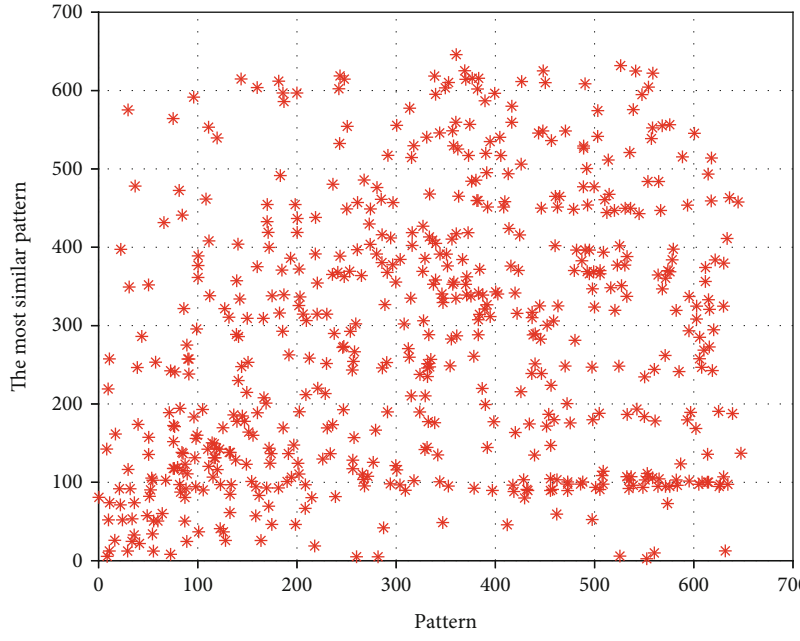


(f) LRW

Figure 6: Continued.



(g) LP



(h) RESG

Figure 6: The scatter plots under *Gene*.

overall view, LRE takes information of the local structure into consideration, and it remains a daunting challenge on obtaining accurate similarity value. In addition, with the size of graph data increasing, the symmetry between patterns decreases significantly.

Figures 5(e) and 6(e) show scatter plots formed by Katz. It is worth noting that the adjustable parameter of Katz index is set to $\alpha = 0.001$, whereby the scatter plot under Katz index is concentrated around the diagonal line. Furthermore, the symmetry is not desirable. Katz index relies more on path among patterns in graph data, and patterns with larger degree are more likely to be in the path between different patterns; so, there is a greater probability that most patterns are similar to the patterns with greater degree in graph data.

Figures 5(f) and 6(f) show scatter plots formed by LRW index, and the number of random walks in this experiment is set to 3. One can see that the scatter plots are unevenly distributed, and the accuracy of similarity obtained by this similarity measure needs to be further improved. Moreover,

LRW index considers the random walk with finite number of steps, and the computational complexity of this measure is higher.

The scatter plots of LP index are vividly showed in Figures 5(g) and 6(g). The advantage of LP index is low computational complexity. However, due to the limited information used, the distribution of similarity values is too concentrated, which makes distinguishable similarity between patterns.

Figures 5(h) and 6(h) show scatter plots formed by RESG index, respectively. As we can see that the most similar pattern is not distributed near to diagonal, and with the size of graph data increasing, the scatter plot formed RESG index still maintains good symmetry. RESG index measures the similarity between patterns using influence of pattern degree and weight and takes the information of first-order neighbors and second-order neighbors of patterns into account, which can get more accurate similarity of any two patterns. Under the circumstances, most patterns avoid becoming general patterns and avoid being identified as certain patterns with common structure that are most similar to multiple patterns. Moreover, in terms of runtime, RESG index is higher to LRW index, CN index, and AA index. However, compared with the same type of relative entropy-based similarity LRE index, the running time of RESG index is only 1/4 of it. In addition, comparing with the normal algorithm, RESG index is simple and efficient and can satisfy measure the similarity of patterns in large graph data efficiently.

For a different method, a quantification named most similar pattern listed in Table 6 is used to demonstrate the difference between RESG index and three existing measures: LRE index, CN index, and EI index [35], so as to verify the good effect of RESG index from another perspective. The first line of Table 6 is the pattern's label, three of every 100 patterns are selected randomly, and a total of 20 will be used as experimental patterns listed. "/" represents that the pattern does not have the most similar pattern. Since there is such a situation that pattern in graph data has more than one of the most similar patterns, only the same pattern sequence numbers are listed, and the rest of most similar patterns are shown in the table with abbreviation numbers. Take pattern v_5 under the EI index for example, (148) represents pattern v_5 which has 148 most similar patterns.

As it shows in Table 6, pattern v_7 is identified as the most similar pattern of 7 different patterns under LRE index, including pattern v_5, v_{52}, v_{172}. LRE index takes the degree of patterns into consideration simply; so, it is possible that most patterns may have the same degree distribution, which leads to the same similarity of patterns. Analogously, under the EI index, several patterns have more than one most similar pattern. For example, a number of 148 most similar patterns are identified by patterns v_5, v_{104}, v_{297} and so on. However, there are also patterns without the most similar pattern, for instance, patterns v_{52}, v_{81}.

As we can see, there is no situation that multiple patterns identify the same most similar patterns under RESG index. RESG index takes information of pattern's one-order and second-order neighbors into account, which can accurately calculate the similarity. Meanwhile, the weight of patterns also contains a lot of topological information, and there may be a situation that the degree distribution is the same but the weight is different. RESG index starts from the perspective of degree and weight, which may make it exact to distinguish the similarity. As a result, RESG index is feasible and effective.

Table 6: The most similar pattern in *Gene*.

Pattern	RESG	LRE	EI	CN
v_5	v_{76}	v_{76}	(148)	(15)
v_{52}	v_{32}	v_{76}	/	v_6
v_{81}	v_{127}	v_{85}	/	v_6
v_{104}	v_{376}	v_{557}	(148)	/
v_{157}	v_9	v_{126}	/	/
v_{172}	v_{156}	v_{76}	/	(2)
v_{201}	v_{129}	v_{76}.	/	v_{172}
v_{253}	(2)	v_{251}	(7)	(2)
v_{297}	v_{192}	v_{557}	(148)	/
v_{314}	v_{578}	v_{520}	(18)	/
v_{348}	v_{324}	v_{346}	(148)	v_{340}
v_{397}	v_{480}	v_{251}	/	/
v_{429}	v_{87}	v_{434}	(148)	(7)
v_{459}	v_{193}	v_{76}	/	v_{457}
v_{498}	v_{181}	v_{76}	/	v_{592}
v_{518}	v_{100}	v_{557}	(7)	/
v_{552}	v_{482}	v_{76}	/	/
v_{590}	v_{100}	v_{154}	(148)	/
v_{616}	v_{241}	v_{483}	/	/
v_{636}	v_{459}	v_{181}	(148)	v_{631}

To further verify the feasibility of the proposed similarity measure, RESG index is applied to link prediction and compared the prediction performance with CN index, LP index, Katz index, and LRW index. The experiment is carried out on six graph data collected from Stanford Dataset, and AUC is selected as an index to evaluate the prediction performance of effective path topology stability. For the more information of AUC, see reference [34] for details.

Figure 7 shows the comparison of AUC results on RESG and other four similarity measures. Among them, CN index only considers the degree information of patterns, LRW index, LP index, and Katz index either consider the local path or the global path of graph data; so, their time complexity is relatively high. As we can see from Figure 7, compared with RESG index and LRW index, the AUC value of CN index, LP index, and Katz value on *Stmarks* and *FWEW* is not ideal. However, compared with RESG index, LRW index has higher time complexity. The AUC of RESG index is the highest on four graph data: *FWMW*, *FWFW*, *Celegans*, and *Email167*, second only to LRW index on *Stmarks* and *FW EW*. Meanwhile, compared with the AUC of other four

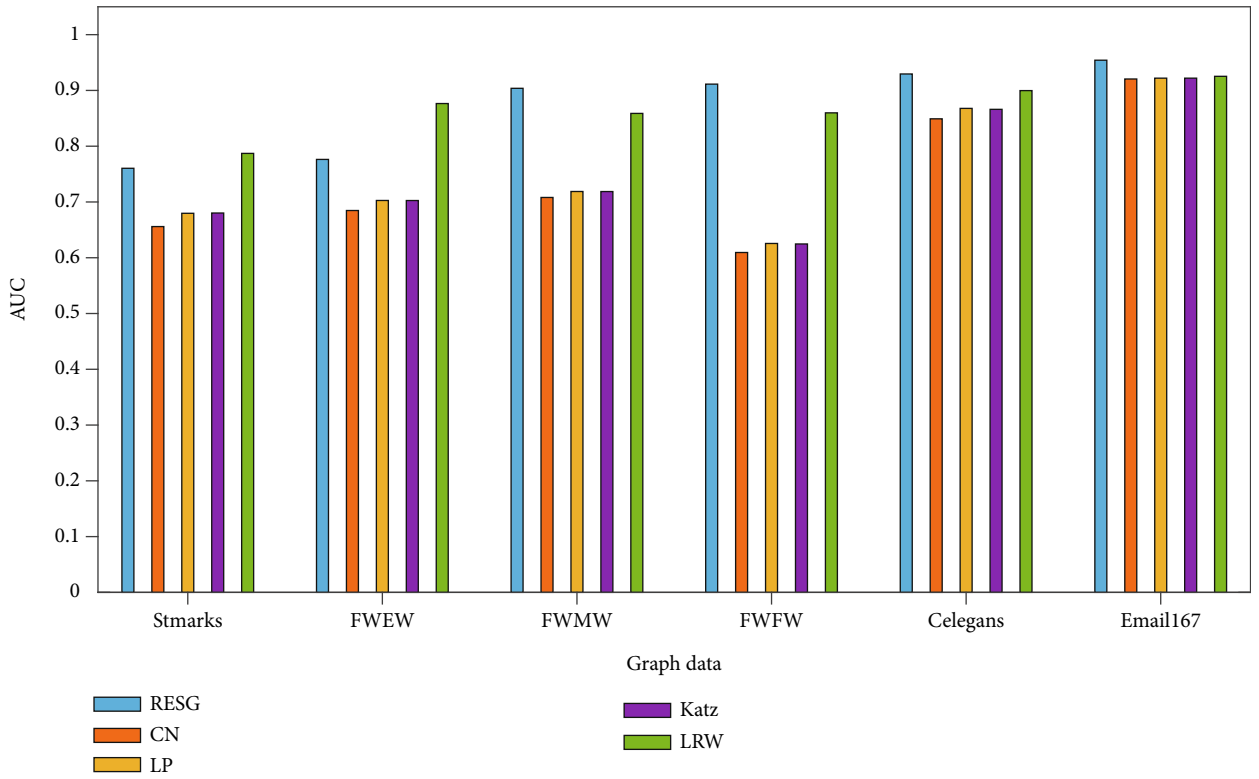


Figure 7: The AUC comparison of each index on six graph data.

measures, the improvement rate can reach 2% – 21%. The experiment suggests that RESG index can achieve the highest AUC value in four graph data; to some extent, it shows the effectiveness and feasibility of RESG index.

However, it deserves our attention that the proposed RESG index also has limitations, and it can achieve better link prediction effect on graph data with small clustering coefficients. For graph data with large clustering coefficient, the effect of this measure needs to be further improved and optimized.

8. Conclusion

Measuring similarity of patterns in graph data is a significant work in many fields. In this paper, to overcome the shortcomings and limitations of existing similarity measures, a relative entropy-based similarity for patterns in graph data abbreviated as RESG index is constructed. Our main work is divided into three aspects. Firstly, strength set is given by degree and weight, which proposed four variables that contains the information of topological relationship in first-order neighbors. Then, in order to generate probability set, patterns with smaller neighbors are redefined by appending empty neighbors up to the same neighbors as another. Finally, relative entropy is computed, and pattern's similarity will be calculated. In addition, two sets of comparison experiments with serval classic similarity measure are used to show effectiveness and feasibility of the proposed RESG index algorithm. Experiments indicate that by taking pattern's degree, weight and second-order neighbors into consideration, the RESG index algorithm can better identify similarity between patterns. To some extent, our purposed approach can enrich the research in area of pattern's similarity in graph data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

What is more, we thank the National Natural Science Foundation of China (No. 61966039). Also, this work is partially supported by the Scientific Research Foundation of Education Department of Yunnan Province (No. 2021Y670).

References

[1] M. Aslani, M. S. Mesgari, and M. Wiering, "Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 732–752, 2017.

[2] Q. Liu, Z. Wu, L. Sun, Y. Xu, L. du, and Z. Wei, "Kernel low-rank representation based on local similarity for hyperspectral image classification," *IEEE Journal of Selected Topics*

in Applied Earth Observations and Remote Sensing, vol. 12, no. 6, pp. 1920–1932, 2019.

[3] J. Wang, J. Y. Liang, and W. P. Zheng, "A graph clustering method for detecting protein complexes," *Journal of Computer Research and Development*, vol. 52, pp. 1784–1793, 2015.

[4] B. Saoud and A. Moussaoui, "Node similarity and modularity for finding communities in networks," *Physica A*, vol. 492, pp. 1958–1966, 2018.

[5] C. Li, Q. M. Yang, B. M. Pang, T. Chen, Q. Cheng, and J. Liu, "A mixed strategy of higher-order structure for link prediction problem on bipartite graphs," *Mathematics*, vol. 9, no. 24, pp. 3195–3207, 2021.

[6] L. Y. Lv and T. Zhou, "Link prediction in complex networks: a survey," *Physica A*, vol. 390, pp. 1150–1170, 2011.

[7] Z. S. Kuang, J. Zhang, X. L. Shao, and B. Chang, "Fault diagnosis methods based on support vector machine and cosine similarity," *Hans Journal of Data Mining*, vol. 10, no. 2, pp. 136–142, 2020.

[8] R. Mahapatra, S. Samanta, M. Pal, and Q. Xin, "RSM index: a new way of link prediction in social networks," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 2, pp. 2137–2151, 2019.

[9] G. Salton and M. J. Mcgill, *Introduction to modem information retrieval*, Mcgraw-Hill BookCo, New York, 1983.

[10] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Network*, vol. 25, no. 3, pp. 211–230, 2003.

[11] E. Haihong, J. J. Tong, S. Meina, and S. Junde, "QoS prediction algorithm used in location-aware hybrid web service," *The Journal of China Universities of Posts and Telecommunications*, vol. 22, no. 1, pp. 42–49, 2015.

[12] X. Cai, J. Shu, and L. Liu, "Study on similarity indices for link prediction in opportunistic networks," *Advances in Mechanical Engineering*, vol. 10, no. 10, Article ID 168781401880319, 2018.

[13] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *The Journal of Mathematical Sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[14] J. K. Ochab and Z. Burda, "Maximal entropy random walk in community detection," *European Physical Journal Special Topics*, vol. 216, no. 1, pp. 73–81, 2013.

[15] E. Nasiri, K. Berahmand, and Y. L. Li, "A new link prediction in multiplex networks using topologically biased random walks," *Chaos, Solitons and Fractals*, vol. 151, article 111230, 2021.

[16] W. P. Liu and L. Y. Lv, "Link prediction based on local random walk," *Europhysics Letters*, vol. 89, article 58007, 2010.

[17] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[18] S. Grewenig, S. Zimmer, and J. Weickert, "Rotationally invariant similarity measures for nonlocal image denoising," *Journal of Visual Communication and Image Representation*, vol. 22, no. 2, pp. 117–130, 2011.

[19] M. Khajehnejad, "SimNet: Similarity-based network embeddings with mean commute time," *Plos One*, vol. 14, article 0221172, 2019.

[20] F. Aziz, H. Gul, I. Uddin, and G. V. Gkoutos, "Path-based extensions of local link prediction methods for complex networks," *Scientific Reports*, vol. 10, article 19848, 2020.

[21] Q. Zhang, M. Z. Li, and Y. Deng, "Measure the structure similarity of nodes in complex networks based on relative entropy," *Physica A*, vol. 491, pp. 749–763, 2018.

[22] L. J. Li, L. Wang, H. S. Luo, and X. Chen, "Towards effective link prediction: a hybrid similarity model," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 3, pp. 4013–4026, 2021.

[23] W. C. Jiang and Y. H. Wang, "Node similarity measure in directed weighted complex network based on *n* ode nearest neighbor local network relative weighted entropy," *IEEE Access*, vol. 8, pp. 32432–32441, 2020.

[24] W. P. Zheng, S. Q. Liu, and J. F. Mu, "A random walk similarity-measure model based on relative entropy," *Journal of Nanjing University*, vol. 55, pp. 984–999, 2019.

[25] W. Tao, S. Y. Duan, and W. Jiang, "Node similarity measuring in complex networks with relative entropy," *Communications in Nonlinear Science and Numerical Simulation*, vol. 78, pp. 104867–104869, 2019.

[26] S. H. Liu and X. Z. Chen, "Random walk-based similarity measure method for patterns in complex object," *Open Physics*, vol. 15, no. 1, pp. 154–159, 2017.

[27] S. Najari, M. Salehi, V. Ranjbar, and M. Jalili, "Link prediction in multiplex networks based on interlayer similarity," *Physica A: Statistical Mechanics and its Applications*, vol. 536, article 120978, 2019.

[28] J. Chen and I. Safro, "A measure of the local connectivity between graph vertices," *Procedia Computer Science*, vol. 4, pp. 196–205, 2011.

[29] C. M. Yu, X. L. Zhao, L. An, and X. Lin, "Similarity-based link prediction in social networks: a path and node combined approach," *Journal of Information Science*, vol. 43, no. 5, pp. 683–695, 2017.

[30] K. Steinhaeuser and N. V. Chawla, "Identifying and evaluating community structure in complex networks," *Pattern Recognition Letters*, vol. 31, no. 5, pp. 413–421, 2010.

[31] K. Berahmand, A. Bouyer, and M. Vasighi, "Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 1021–1033, 2018.

[32] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[33] "SC-TS Biological Networks Data," http://networkrepository.com/bio-SC-TS.php.

[34] Y. J. Liu, S. H. Liu, and W. H. Xu, "Link prediction method based on topology stability of effective path," *Application Research of Computers*, vol. 39, pp. 90–95, 2021.

[35] D. Zhong and I. Defee, "Performance of similarity measures based on histograms of local image feature vectors," *Pattern Recognition Letters*, vol. 28, pp. 2003–2010, 2010.