An Ontology-Independent Representation Learning for Similar Disease Detection Based on Multi-layer Similarity Network

Ruiqi Qin, Lei Duan, Huiru Zheng, Jesse Li-Ling, Kaiwen Song and Yidan Zhang

Abstract—To identify similar diseases has significant implications for revealing the etiology and pathogenesis of diseases and further research in the domain of biomedicine. Currently most methods for the measurement of disease similarity utilize either associations of ontological disease concepts or functional interactions between disease-related genes. These methods are heavily dependent on the ontology, which are not always available, and the selection of datasets. Moreover, many methods suffer from a drawback that they only use a single metric to evaluate disease similarity from an individual data source, which may result in biased conclusions without consideration of other aspects. In this study, we proposed a novel ontology-independent framework, namely *RADAR*, for learning representations for diseases to deduce their similarities from an integrative perspective. By leveraging the associations between diseases and disease-related biomedical entities, a disease similarity network was built under various metrics. Then a multi-layer disease similarity network was constructed by integrating multiple disease similarity networks derived from multiple data sources, where the representation learning was derived to provide a comprehensive evaluation of disease similarities. The performance of *RADAR* was assessed by a benchmark disease set and 100 random disease sets. Experimental results demonstrated that *RADAR* can detect similar diseases effectively.

Index Terms—disease similarity, disease information network, representation learning, multi-layer similarity network

1 Introduction

Knowledge of how various diseases are related can facilitate deepening the understanding of their etiology and pathogenesis. Theoretically, diseases may be related from the following aspects:

- Phenotypically: this is exactly how it happens in clinics - for every new patient, doctors will try to give it a diagnosis based on his or her phenotypical similarity (in symptoms and signs) to a well-defined disease.
- Genetically: allelic disorders are genetic disorders that have different phenotype but are caused by different mutations in the same gene. I-cell disease (ML II) and pseudo-Hurler polydystrophy (ML III) are a pair of examples. They are both caused by mutations in *GNPTAB* gene, with I-cell being more severe and presenting earlier.
- At molecular level: some diseases, though with drastically different clinical feature and different causative genes, are related by sharing the same molecular pathway. For instance, phenylketonuria, goitrous cretinism, albinism, tyrosinosis, alkaptonuria are all caused by defects of phenylalanine metabolism.

With large amount of data generated from medical literature and molecular biology research, there is no doubt that many relationships between diseases are await to be discovered.

To detect disease similarity has made significant contribution to the discovery of relationships among many other biomedical data for further research. For example, the disease similarity has been used to infer the relationship among microRNAs [1], [2], to explore the relationship among long non-coding RNAs [3], [4], [5], [6], and for the prediction of therapeutic drugs for diseases [7], [8], [9].

Typically, there are two queries w.r.t. similar diseases:

- *Top-k query*: to search top-k most similar diseases with respect to a given disease. Such query can be that, for example, which ten diseases are most similar to Alzheimer's disease in a given disease set?
- *Similar pair query*: to discover the most similar disease pairs from a given disease set. For example, a similar pair query can be that: which pairwise diseases are most similar to each other in a given disease set?

Various aspects including pathogenesis and phenotypes can be exploited to compute the similarity of pairwise diseases. Current methods for the measurement of disease similarities may be classified into two categories:

- *Semantics-based*: the disease similarity is computed by measuring the similarity between the disease-associated ontological terms. Based on the information theory, the concept of *information theory* (IC) was

- *R. Qin is with the School of Computer Science, Sichuan University, Chengdu, China.*
- *L. Duan is with the School of Computer Science, Sichuan University, Chengdu 610041, China. E-mail: leiduan@scu.edu.cn.*
- *H. Zheng is with the School of Computing, Ulster University, Northern Ireland, United Kingdom.*
- *J. Li-Ling is with the State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, China.*
- *K. Song and Y. Zhang are with the School of Computer Science, Sichuan University, Chengdu, China.*

proposed and has been widely used for measuring the semantic similarity of ontological terms. For example, Resnik *et al* [10] used the "is_a" relationship between terms as the basis of similarity computation, and Lin *et al* [11] considered both the commonality and difference between terms based on IC. Wang *et al* [12] focused on the semantic similarity of Gene Ontology (GO) [13] terms, and first used two types of relationships as a hybrid measurement to compute such similarity. Ever since the emergence of Disease Ontology (DO) [14], which is the first standardized ontology for human diseases based on the disease terms collected from multiple sources such as Medical Subject Headings (MeSH) [15] and Online Mendelian Inheritance in Man (OMIM) [16], methods have been proposed to calculate disease similarity based on DO terms, such as the system *DOSim* [17], where ten representative semantic similarity measurements of ontological terms are implemented.

- *Semantics + Function-based*: Knowledge has been found that the genome is closely associated to diseases and can be used to reveal relationships between diseases in molecular level. Methods have been proposed to infer disease similarity by assessing disease-related genes, such as by the number of shared genes [18]. Recently, many researches utilize both DO and gene functional associations to enhance the measurement of disease similarity. For example, *SemFunSim* [8] obtained the disease similarity by two parts, one part measured by the weight value of disease-related genes in a weighted gene interaction network from *HumanNet* [19], the other part obtained from the semantic relationships between pairwise diseases based on DO terms. The similar idea was adopted by *InfDisSim* [20] and another method [21], which combined the disease functional similarity derived from [8] with the disease semantic similarity derived from [12].

However, it is worth noting that three limitations are commonly associated with the methods mentioned above: (1) all of them compute disease similarity by leveraging some quantitative information about diseases and disease-related biomedical entities, while the fact is that the precise numerical data describing their relationships are not always available; (2) many of these methods measure the disease similarity only by a single metric or only from a single data source, while the fact is that results tend to differ under different metrics or with different sources, leading to biased conclusion that lack of comprehensive assessment; (3) most of the above methods are limited in their strong dependence on the ontology (DO or GO) when computing the semantic similarity for diseases, without which such methods can no longer work. Though the ontology provides precise descriptions of disease concepts and their semantic relationships, not all biomedical entities have ontologies.

Through the analysis of these limitations, we derive the following observations.

- The computation of disease similarity should not strictly depend on quantitative information.
- More metrics and more data sources should be used for the measurement of disease similarity.
- Ontology-independent strategies should be developed to measure disease similarity for general usage.

To address the above mentioned challenges, we propose a novel approach, *RADAR* (short for representation learning across disease information networks), for similar disease detection. The characteristics of *RADAR* include: (1) it is capable of computing disease similarity without dependence on any ontology; (2) it computes disease similarity solely based on the associations between diseases and other disease-related biomedical entities; (3) it evaluates disease similarity based on multiple data sources under orthogonal similarity metrics (i.e. meta-path-based and neighborhood-based structural similarities); and (4) it flexibly supports the two typical queries on similar diseases.

The main contributions of this work are as follows:

- We propose *RADAR*, a general ontology-independent framework for learning latent representations for diseases that reflect their similarities from a perspective where multiple data sources are involved and considered. Such representations can be further applied to similar disease detection.
- We show how *RADAR* measures disease similarity under various similarity metrics, while solely based on the relationships between diseases and other related entities without reference to any numerical data.
- We evaluate *RADAR* on a benchmark disease set and multiple random disease sets to demonstrate its effectiveness in searching similar diseases and its insensitiveness to parameters.

The rest of the paper is organized as follows. We review the related work in Section 2. The framework and algorithm of *RADAR* is detailed in Section 3, followed by experiments and results discussed in Section 4. We conclude the paper in Section 5.

2 Related Work

Many recently proposed methods related to the disease similarity computation have used multiple data sources. For example, *SemFunSim* [8] integrated five disease-related gene databases to get the disease-related gene sets for calculating the disease functional similarity. The same idea was later adopted by another two methods *InfDisSim* [20] and *FNSemSim* [22] to collect disease-gene associations, while *FNSemSim* further fuses two gene functional networks *FunCoup* and *HumanNet* to improve the calculation of disease similarity. Despite the fact that all of these methods have used multiple data sources, such multi-source data are not fully utilized because they are simply used for data collection at the beginning rather than for the measurement of disease similarities. One significant problem with such methods is their poor scalability, as all the previous computation would have to be done again if a new data source were added.

In the recent year, many studies in the field of biomedicine such as [23] and [24] have shown that by integrative analysis of multiple data types, better performances can be achieved in discovering similar objects.

The computation of disease similarity is actually the process of constructing a similarity network. Inspired by the theoretical multiview learning framework, a method called *SNF* [23] fuses multiple similarity networks on the basis of samples. *SNF* first builds several sample-similarity networks based on different data types and then fuses all these networks into a single similarity network, which represents the full spectrum of the underlying data. It is distinctive in iteratively updating each similarity network with the information from the others, and at last, all similarity networks are fused into one. During the process of fusing networks, the weak similarities disappear while the strong similarities are kept. Though this may lead to lost of original information, *SNF* manages to utilize the similarity information of all the similarity networks and thus is reasonable and powerful.

Based on network science, Mucha *et al.* studied the community structure of arbitrary multi-slice networks, which are the combinations of individual networks coupled through links [25]. Inspired by this, a multiplex network-based method integrates various omics data to identify cancer subtypes [24]. Similar to *SNF*, it first constructs a patient-wise similarity network for each type of data and then uses a coupling strength to link each node in a network slice with its counterparts in the other network slices to build the multiplex network, where the analysis is done. This method outperforms the methods that only use a single data type.

In recent years, the representation learning technique has been applied to a wide range of applications. It is capable of capturing the essential semantics of objects and presents them as dense vectors in low-dimensional space (known as *embeddings*), providing convenience for further analysis. Among various representation learning models, Skip-Gram [26] has been proven to be fairly effective and efficient in learning embeddings for textual data such as words and sentences.

Network representation learning was first proposed by *DeepWalk* [27]. *DeepWalk* considers that nodes with closer locations in the network are likely to have similar contexts. Thus, *DeepWalk* generates sequences for nodes by carrying out random walks on the network and then uses the Skip-Gram model to learning embeddings from such sequences. Later, an improved method called *node2vec* was proposed to learn features for nodes that maximize the probability of preserving the network neighborhoods of nodes [28]. It uses a second order biased random walk to generate contexts for nodes to capture the homophily as well as structural equivalence. This method is more flexible compared with the previous method for generating contexts. However, all these methods are designed for the homogenous networks and cannot be directly applied to heterogenous networks.

In Table 1, we compare our method with several typical methods which are for the measurement of disease similarity. As presented, our method is characterized in three aspects: using multiple sources, using multiple metrics and ontology-independent.

We tackled the problem of similar disease detection in [29], a preliminary paper of this study. Compared to that work, in this paper, we present a more complete analysis of the related work, provide a more detailed description of the key steps in our method and perform more extensive empirical evaluations to demonstrate the effectiveness of our method.

TABLE 1
Comparison of current methods for the measurement of disease similarity

Method	Multiple Sources	Multiple Metrics	Ontology Independent
Resnik's [10]	×	×	×
Lin's [11]	×	×	×
Wang's [12]	×	×	×
SemFunSim [8]	✓	✓	×
RADRA	✓	✓	✓

3 The Proposed *RADAR* Approach

To address the problem of similar disease detection, the key step of *RADAR* is the construction of the *disease similarity network*, an undirected graph expressing the similarities among diseases. At the very beginning, a *disease information network* will be built from each data source, which is a typical heterogeneous information network defined as:

Definition 1 (Disease Information Network). A disease information network (DIN) is a graph $G = (V, E)$ with an object mapping function $\phi : V \rightarrow A$ and a link mapping function $\psi : E \rightarrow R$, where A refers to the set of disease-related biomedical object types and R denotes the set of relations between objects. Each object $v \in V$ belongs to an object type $\phi(v) \in A$, and each link $e \in E$ belongs to a relation $\psi(e) \in R$.

Due to the space limitation, please refer to [30] for the details of the process of building a disease information network from a given data source, considering this is not the focus in our study.

Definition 2 (Disease Similarity Network). A disease similarity network (DSN) is an undirected graph $S = (\mathcal{D}, \mathcal{E})$ composed of a set of nodes and a set of edges, where each node $d \in \mathcal{D}$ corresponds to a disease and each edge $e \in \mathcal{E}$ refers to the similarity between two diseases that it connects.

In the case of multiple data sources, multiple DSNs will be constructed. Each disease node will then be connected to itself in all the other DSNs by *RADAR*. In other words, a multi-layer DSN may be constructed, from which the similarity between diseases can be derived.

The main steps of the *RADAR* framework are illustrated in Figure 1.

Step 1 **Single-layer DSN Construction**: For a disease information network, the meta-path based and neighborhood-based structural similarities between every disease pair are calculated by two similarity metrics. After the combination of the similarities, one united disease similarity network is constructed. (Section 3.1)

Step 2 **Multi-layer DSN Construction**: Associate all the disease similarity networks obtained in the previous step into a multi-layer disease similarity network

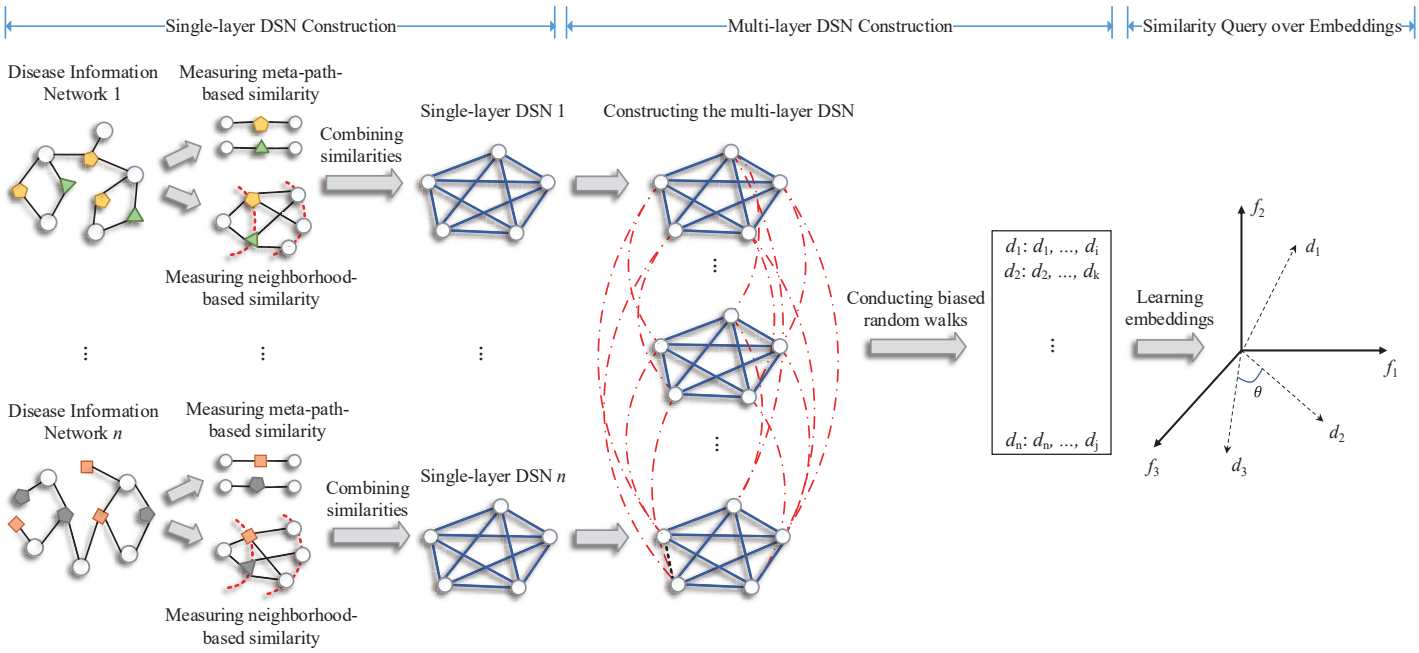


Fig. 1. The similar disease detection framework of *RADAR*

and then the biased random walk is conducted on it to generate a context for every disease. (Section 3.2)

Step 3 **Similarity query over Embeddings**: Apply the Skip-Gram model to learn the latent representation for each disease from its context. (Section 3.3)

Next, we introduce each step of *RADAR* in details.

3.1 Single-layer Disease Similarity Network Construction

In a disease information network, two diseases can be connected through different paths. The *meta path* [30] is a special path that is defined as:

Definition 3 (Meta Path). A meta path $\mathcal{P}$ is a path defined on the information network and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$, where $R = R_1 \circ R_2 \circ \dots \circ R_l$ is a composite relation between object type A_1 and A_{l+1}, where $\circ$ denotes the composition operator on relations.

In particular, a meta path $\mathcal{P}$ is a *disease meta path* if the two end nodes of $\mathcal{P}$ are two diseases belonging to $\mathcal{D}$.

Definition 4 (Disease Path Instance Set). Given a disease meta path $\mathcal{P}$ in a DIN, the disease path instance set, denoted by $Ins(\mathcal{P}_{d \to d'})$, is a set of paths which go from d to d' following $\mathcal{P}$, where $d, d' \in \mathcal{D}$.

Example 1. An example of disease information network is illustrated in Figure 2. In total there are four object types $\{D, G, P, W\}$ and multiple disease meta paths can be found. For example, the disease meta path "*D-W-D*" (short for "*Disease-Pathway-Disease*") indicates two diseases sharing the same molecular pathway, with disease path instances such as "$d_1 - w_1 - d_2$" and "$d_4 - w_2 - d_5$". And "*D-G-D*" (short for "*Disease-Gene-Disease*"), indicates that two diseases are triggered by the same gene, with disease path instances like "$d_2 - g_2 - d_3$" and "$d_6 - g_3 - d_7$".

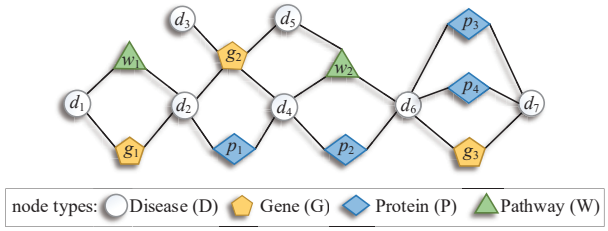


Fig. 2. An example of disease information network

Observation 1. In a disease information network, two diseases are considered to be more similar in two aspects: (1) they are connected via more disease meta paths; and (2) they are connected via a shorter disease meta path.

Example 2. In Figure 2, d_7 is more similar to d_6 compared with d_5 to d_6. This is because d_6 and d_7 share three paths, i.e., $\{d_6 - p_3 - d_7, d_6 - p_4 - d_7, d_6 - g_3 - d_7\}$, while d_6 and d_5 only share one path $\{d_6 - w_2 - d_5\}$. Besides, d_6 and d_1 are unlikely to be similar compared with d_6 to d_4. This is because the disease meta path via d_6 and d_1 is much longer than that via d_6 and d_4, which indicates a loose relationship between d_6 and d_1.

By leveraging the structure of a disease information network, *RADAR* constructs its corresponding disease similarity network under two orthogonal similarity measurements, i.e., the meta path-based and the neighborhood-based structural similarity metrics.

3.1.1 Measuring Meta Path-based Structural Similarity

In a heterogenous network, the meta path, which is a special path through a certain number of nodes, is usually used to imply the subtle relationship between two end nodes. A meta path-based similarity measure called *PathSim* [30]

was proposed to find similar peer nodes in a network based on a symmetric meta path, and has received fairly good effects. Similarly, *RADAR* searches similar diseases in a disease information network by a predefined symmetric disease meta path that indicates the relationship between two diseases.

Formally, for two diseases $d_1 \in \mathcal{D}$ and $d_2 \in \mathcal{D}$, their meta path-based structural similarity is defined as:

$$StrSim_{path}(d_1, d_2) = \frac{2 \times |Ins(\mathcal{P}_{d_1 \to d_2})|}{|Ins(\mathcal{P}_{d_1 \to d_1})| + |Ins(\mathcal{P}_{d_2 \to d_2})|} \quad (1)$$

where $|Ins(\mathcal{P}_{d_1 \to d_2})|$ is the number of distinct paths following the disease meta path $\mathcal{P}_{d_1 \to d_2}$.

According to the self-maximum property of *Path-Sim* [30], we have $0.0 \leq StrSim_{path}(d_1, d_2) \leq 1.0$, where $StrSim_{path}(d_1, d_2) = 0.0$ if d_1 and d_2 are not connected, and $StrSim_{path}(d_1, d_2) = 1.0$ if d_1 and d_2 share exactly every related object involved in the disease meta path.

By traversing the whole network, *RADAR* computes the meta path-based structural similarity for every disease pair based on the predefined disease meta path according to Equation 1. In case of more than two types of objects in the DIN (i.e., $|A| > 2$), domain knowledge is needed for making decision on defining the disease meta path(s) to be used. If over two disease meta paths are used, multiple results will be produced accordingly. In default, each disease meta path takes equal importance, while different weights can be assigned to different parts in terms of their contributions to the expression of disease similarity with the help of domain knowledge. The final meta path-based structural similarity is the weighted summation of each part.

3.1.2 Measuring Neighborhood-based Structural Similarity

It is worth noting that due to the constraint of the defined disease meta path, the disease meta path-based similarity measurement can only capture the relationship between two end nodes of the path and may thus fail to discover more potential similar nodes.

Example 3. In Figure 2, if the disease meta path is defined as "*D-W-D*", then d_2 and d_4 will not be considered to be similar even if they share one related gene and one related protein.

From Example 3, it is clear that the disease similarity depends on the disease meta path used, which requires domain knowledge when defined, and improper selection of disease meta path may lead to biased results.

A recent approach given in [31] provides a new in-sight into solving the problem of finding similar pairs in a homogeneous network. The similarity between pairwise nodes is calculated solely based on their structural identities of neighborhoods in a network, and thus more similar node pairs are able to be discovered. As *struc2vec* [31] can only handle homogeneous data, *RADAR* further adapted this idea into the disease information network, to measure the disease similarity based on their structural identities of neighborhoods, as a supplement of the computation of disease similarity.

For any two nodes in a disease information network, we use *hop* to denote the least number of moves that will be made from one node to the other. We refer the neighborhood of a node d to a set of nodes directly connected to d or having an indirect connection to d, and we call this set of nodes as the *ϵ-Neighbor Set* defined as:

Definition 5 (ϵ-Neighbor Set). In a disease information network, we use $\ell_\epsilon(d)$ to denote the set of nodes which are ϵ hop(s) ($\epsilon \geq 1$) from d, where $d \in \mathcal{D}$.

Example 4. In Figure 2, for d_4, it has four 1-hop neighbors, i.e., $\ell_1(d_4) = \{g_2, p_1, p_2, w_2\}$, and four 2-hop neighbors, i.e., $\ell_2(d_4) = \{d_2, d_3, d_5, d_6\}$. For d_5, $\ell_1(d_5) = \{g_2, w_2\}$ and $\ell_2(d_5) = \{d_2, d_3, d_4, d_6\}$, which means d_5 has two 1-hop neighbors and four 2-hop neighbors.

Due to the heterogeneity of the disease information network and the characteristics of biomedical entities, we derive the following observation:

Observation 2. For any node in a disease information network, its nearer neighbors contribute more to describe its structural identity than its farther neighbors.

Inspired by the idea of *Katz* centrality [32], a decaying weight factor α in the range between 0 and 1 is introduced to penalize the contributions of distant neighbors of a node in the DIN.

In a disease information network, the number of edges incident to a node $v \in V$ is called the *degree* of v. We denote $DS(\ell_\epsilon(d))$ the degree sequence of each node in $\ell_\epsilon(d)$ sorted in the ascending order for accelerating the computation. Let α be the decaying weight factor that determines the importance of neighborhoods of nodes at different hops. Given a disease information network containing a set of diseases $\mathcal{D}$, the neighborhood-based structural distance between two disease nodes $d_1, d_2 \in \mathcal{D}$ is defined as:

$$\begin{aligned} StrDis_\epsilon(d_1, d_2) &= StrDis_{\epsilon-1}(d_1, d_2) + \\ \alpha^\epsilon &\times \mathcal{T}(DS(\ell_\epsilon(d_1)), DS(\ell_\epsilon(d_2))) \end{aligned} \quad (2)$$

where $\mathcal{T}(DS(\ell_\epsilon(d_1)), DS(\ell_\epsilon(d_2)))$ measures the *distance* between two ordered degree sequences $DS(\ell_\epsilon(d_1))$ and $DS(\ell_\epsilon(d_2))$, and $StrDis_0(d_1, d_2) = 0$. Since the ordered degree sequences are composed of numerical elements, they can be regarded as time series. Therefore, the Dynamic Time Warping (DTW) [33] method is adopted to calculate the approximate distance between two sequences, as DTW has been verified to be very effective in handling time series by using some optimal element alignment strategies to ensure the distance of two sequences is minimal. For two ordered degree sequences DS_1 and DS_2, the distance between the i-th element in DS_1 (denoted by $DS_1[i]$) and the j-th element in DS_2 (denoted by $DS_2[j]$) is defined as:

$$dis(DS_1[i], DS_2[j]) = \frac{max(DS_1[i], DS_2[j]) + \eta}{min(DS_1[i], DS_2[j]) + \eta} - 1 \quad (3)$$

where η is a parameter preventing $dis(\cdot)$ being too large. (We set $\eta = 0.5$ as in [31].)

In this way, by applying Equation 3, the distances between every pair of matched elements are obtained, and $\mathcal{T}(\cdot)$ further sums the distances to get the final distance between two degree sequences.

For any disease node $d \in \mathcal{D}$, as the hop count ϵ increasing, the according hop of its neighborhood takes less importance with regards to d, since α gives more penalty to the further neighborhood. In such sense, it would be

meaningless to go too far from d. Therefore, *RADAR* only takes the first several hops (we set $\epsilon = 2$) of neighbors of d into consideration when describing the structural identity of d. The decaying weight factor α will be evaluated in the experiment to test its impact on *RADAR*.

We use the natural exponential function to restrict the value of similarity in the range between 0.0 and 1.0, and the final neighborhood-based structural similarity between diseases d_1 and d_2 is

$$StrSim_{nei}(d_1, d_2) = e^{-StrDis_{\epsilon}(d_1,d_2)} \tag{4}$$

For every disease in a pair, *RADAR* traverses each hop of its neighborhood starting from itself to its ϵ-hop neighborhood and computes the neighborhood-based structural similarity between them according to Equation 4. Similar to the computation of meta-path-based disease similarity, if more than two types of objects exist in the DIN, a set of neighborhood-based structural similarities will be produced based on each type of object, and thus multiple sets of similarities will be produced. The final similarity is obtained by the weighted summation of each part.

Though both targeted at a certain disease pair based on the disease information network, the main difference between the two similarity metrics that we have applied lies in that: $StrSim_{path}$ requires some predefined knowledge (i.e., disease meta path), and only focuses on the characteristics of the two diseases and counts the number of their path instances, while $StrSim_{nei}$ needs no professional knowledge and considers a wider local area of the network structure (i.e., neighborhood) of each disease compared.

3.1.3 Similarity Combination

After measuring the disease similarity under two orthogonal measurements on a disease information network, two sets of disease similarities have been obtained. Now *RADAR* merges these similarities together to build a united disease similarity network.

A straightforward way is to use the linear combination, with the weight for each part predefined. Formally, the combined similarity for a DIN is computed as:

$$Sim = w_1 \times StrSim_{path} + w_2 \times StrSim_{nei} \tag{5}$$

where $w_1, w_2 \in [0, 1]$ are the weights that adjust the contribution of each similarity and $w_1 + w_2 = 1$. Here, equal importance is given to each part, i.e., $w_1 = w_2 = 0.5$. In case of more than two types of metrics, each part will be assigned a weight with the sum of all weights being 1.0 if linear combination is adopted for similarity fusion. Any other merging method can be adopted to combine the similarities obtained under any other metrics besides $StrSim_{path}$ and $StrSim_{nei}$.

3.2 Multi-layer Disease Similarity Network Construction

Though calculated by the same measurements, the DSNs obtained from multiple disease information networks are different from each other because the characteristics of the disease information networks differ. In order to best keep the original information about every DSN, all DSNs are integrated into a multi-layer DSN by associating each disease node located in one DSN with its counterpart in another by an unweighted edge. The edges are unweighted because they are simply used to indicate the same nodes in different networks.

RADAR has an advantage over *SNF* [23] in terms of retaining original information. *RADAR* constructs the multi-layer DSN without lost of any information about each similarity network, while *SNF* fuses multiple similarity networks into a single one, only keeping the strong similarities but losing the weak ones.

Over the multi-layer DSN, *RADAR* then conducts the random walks, particularly the biased random walks to generate a node sequence for each disease node, which can be regarded as its context. In an individual network, at each step, the walker randomly selects the next node with a probability proportional to the weights of the current node's links. Since the weight of each edge in an DSN is represented by the similarity score of the disease pair it connects, the random walking for a disease node in a DSN is actually a process of choosing a set of nodes for it, with each similar to its previous one and the whole node sequence similar to the starting node. Obviously, even for the same node, the edges incident to it in different DSN may have quite different weights. Therefore, it is necessary to switch layers of networks during the random walks for greater weights.

Taking this into consideration, the layer of DSN where node d has the maximal link weight among all will be selected, which is denoted by $layer_{max}(d)$, and the random walk for d will only be conducted at $layer_{max}(d)$. In each step of the process of generating a node sequence for node d, a random step will only be made if the current layer is $layer_{max}(d)$, while in the other case, the walker will switch to $layer_{max}(d)$ and no step will be made in this turn. In this way, the sequence of a node generated by the walker will be composed of a series of similar nodes across the multi-layer DSN.

In summary, *RADAR* will first start from a random layer at a random disease node. Then the biased random walk with the length of δ is conducted for every disease node and its context will be produced accordingly in the end. By walking in the multi-layer DSN, the generated context is able to capture the similarity relationships for a disease node from multiple perspectives. For the sake of sufficiency of node sequences, the random walking is repeated several times for each node.

3.3 Similarity Query over Embeddings

We adopt the Skip-Gram model to learn embeddings (with dimension λ) for all disease nodes based on the generated contexts. As introduced in Section 2, the embeddings of nodes can successfully capture the similarities obtained from multiple disease information networks. Though the Skip-Gram model is adopted by *RADAR*, any other representation learning models can be used as an alternative to learn embeddings for diseases.

Since each disease is now represented by a vector, the disease similarity can be easily calculated by applying a favorable distance measurement, such as the cosine and Pearson correlation coefficient. The framework of *RADAR* is summarized in Algorithm 1.

Algorithm 1 RADAR ($\mathcal{N}$)

Input: $\mathcal{N}$: the set of disease information networks
Output: $\mathcal{R}$: the results of similar disease query
1: $\mathcal{G} \leftarrow \emptyset$
2: **for** $N \in \mathcal{N}$ **do**
3: compute $StrSim_{path}$ and $StrSim_{nei}$ for every pair of diseases in N
4: $G \leftarrow$ the single-layer DSN constructed from N
5: $\mathcal{G} \leftarrow \mathcal{G} \cup \{G\}$
6: **end for**
7: Connect the same disease nodes in different layers in $\mathcal{G}$
8: Produce the set of layers $\mathcal{L}$ where the maximum link weight of each node is obtained
9: Generate contexts Con by conducting the biased random walks on $\mathcal{G}$ based on $\mathcal{L}$
10: Learn embeddings $\mathcal{M}$ for all nodes from Con
11: $\mathcal{R} \leftarrow$ Perform similar disease query over $\mathcal{M}$
12: **return** $\mathcal{R}$

TABLE 2
Characteristics of the Disease Information Networks

DIN	Type of Nodes	# of Nodes	# of Edges
Dis-C	Disease Chemical	1626 4127	405567
Dis-P	Disease Pathway	1626 2313	222047

4 Experiments, Results and Discussion

In this section, we evaluated the ability and performance of *RADAR* in measuring disease similarity with capturing the structural identities of diseases in multiple disease information networks.

4.1 Datasets

We used the data from Comparative Toxicogenomics Database[1] (CTD) [34], which is a public database mainly providing information about environmentally influenced diseases and their relationships with chemicals, genes and some other biomedical entities. We adopted two datasets including three types of biomedical entities and their corresponding associations to search similar diseases. One dataset contains associations between diseases and chemicals, based on which a disease information network called "Dis-C" was built. The other one contains associations between diseases and pathways, based on which a disease information network called "Dis-P" was built. The detailed characteristics of these two disease information networks are presented in Table 2.

4.2 Effectiveness

We extracted the benchmark disease set including 40 diseases between 56 disease pairs from the set given by [8] as the positive samples, which contained disease pairs that have been confirmed to be similar by Suthram *et al* [35] and Pakhomov *et al* [36]. A random disease set with 200 disease pairs was generated by randomly selecting 200 disease pairs from the whole disease set excluding the diseases included in the benchmark set. Such random disease pairs were regarded as the negative samples, i.e., dissimilar disease pairs, because it is unlikely that two diseases picked up randomly happen to be so similar. In each experiment, we applied *RADAR* on the benchmark disease set as well as a random disease set to test its effectiveness of finding similar diseases. Throughout each experiment, several running parameters were set in default as $\alpha = 0.5$, $\epsilon = 2$, $\delta = 160$ and $\lambda = 128$. The experiment was iterated for 100 times in order to reduce occasionality. The cosine measurement was adopted to measure the distance of vectors generated in Section 3.3.

To the best of our knowledge, *RADAR* is the first work that measures disease similarity from multiple data sources under two DIN-based structural similarity measurements. Thus, we first verified the necessities of measuring disease similarity (1) under two orthogonal metrics, (2) across multiple disease information networks, respectively.

First, we compared *RADAR* with its two variations. Specifically, we implemented two versions of *RADAR*. One only computed the meta path-based similarity (Equation 1), which was called "Sim_{path}". The other one only computed the neighborhood-based structural similarity (Equation 4), which was called "Sim_{nei}". Since only one set of disease similarity will be produced under a single similarity metric, there is no need for similarity combination as introduced in Section 3.1.3.

Figure 3 illustrates the Receiver Operating Characteristic (ROC) curves drawn for *RADAR*, "Sim_{path}" and "Sim_{nei}", respectively, based on the benchmark disease set and random disease sets. It is clear that, *RADAR* achieved the best performance with an Area Under the ROC Curve (AUC) of 0.8698, followed by "Sim_{path}" and "Sim_{nei}", with the AUCs of 0.8479 and 0.7807, respectively. Nevertheless, all of them did much better than the random performance. The result demonstrates that combining similarities obtained under two types of metrics makes it more effective than only using a single metric for similar disease detection.

Second, we applied *RADAR* only on the "*Dis-C*" DIN (called "*DIN-C*"), only on the "*Dis-P*" DIN (called "*DIN-P*"), and across the two DINs, respectively, to evaluate the necessity of integrating multiple disease similarity networks obtained from different data sources.

Figure 4 shows the ROC curves drawn for *RADAR*, "*DIN-C*" and "*DIN-P*", respectively. *RADAR* performed the best, while "*DIN-C*" and "*DIN-P*" received relatively lower AUCs of 0.8429 and 0.8284, respectively. The result verifies the effectiveness of integrating multiple similarity networks.

We can see from the above results that by combining various similarity measurements and integrating multiple similarity networks, better performance can be achieved for similar disease detection.

Next, several diseases were randomly selected from the disease set as the query diseases, and a list comprising the top-5 most similar diseases to each query was generated by *RADAR*. The results were recorded in Table 3. Take the disease *Anemia* for example, *RADAR* has discovered that *Melanoma* was most similar or related to it within the given disease set. Many studies on these two diseases have revealed their close relationship, such as it was found that

1. http://ctdbase.org/downloads/

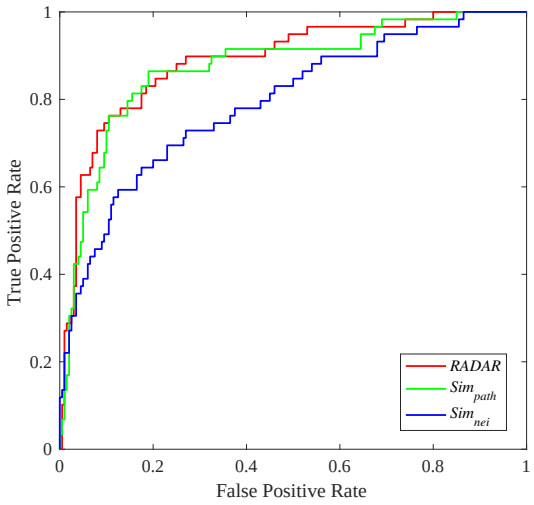


Fig. 3. Performance analysis of combining similarities under the meta path-based structural similarity measurement and the neighborhood-based structural similarity measurement compared with using only single similarity metric.

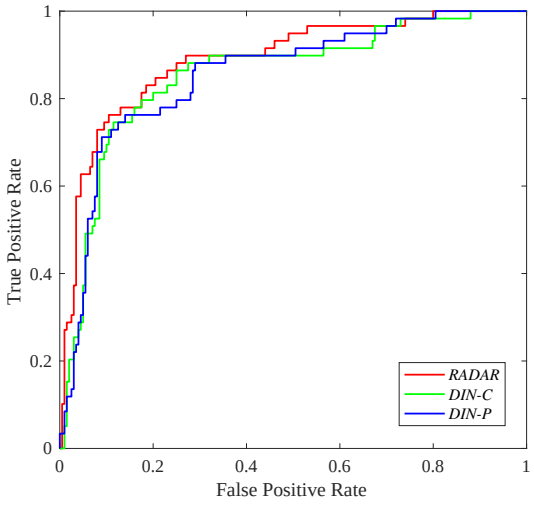


Fig. 4. Performance analysis of integrating multiple similarity networks compared with using single similarity network.

the autoimmune hemolytic anemia is induced by anti-PD-1 therapy in metastatic melanoma [37].

4.3 Case Study

Three diseases *Myotonia Congenita*, *Myotonic Dystrophy*, and *Myotonic Disorders* were selected as the targets and analysed for further evaluation of the effectiveness of our method. Besides, two non-related diseases *Vitiligo* and *Kidney Calculi* were selected as the contrasts. Table 4 presents the similarity score of each disease pair measured by our method.

It is known that clinically *Myotonia Congenita*, *Myotonic Dystrophy*, and *Myotonic Disorders* all belong to certain neuromuscular disorders. In the curated medical vocabulary resource *MeSH* [15], both the terms *Myotonia Congenita* and *Myotonic Dystrophy* are found to be the children of the term

TABLE 3
Top-k similar diseases for the given queries

Query	Top-5 Results	Score
Anemia	Melanoma	0.6424
	Spinocerebellar Ataxia	0.6332
	Shock, Hemorrhagic	0.6255
	Brain Injuries	0.6217
	Pulmonary Emphysema	0.6175
Hypertension	Trigeminal Neuralgia	0.6429
	Hyperplasia	0.6373
	Spinocerebellar Ataxia	0.6152
	Mammary Neoplasms	0.6047
	Inflammation	0.6047
Obesity	Stomach Neoplasms	0.6122
	Atopic Dermatitis	0.6116
	Pulmonary Fibrosis	0.6109
	Liver Cirrhosis	0.5960
	Neoplasm Metastasis	0.5953
Arthritis, Experimental	Spinocerebellar Ataxia	0.6020
	Acute Coronary Syndrome	0.5904
	Hyperplasia	0.5841
	Small Cell Lung Carcinoma	0.5799
	Leukemia, Promyelocytic, Acute	0.5742
Coronary Disease	Myocardial Ischemia	0.5620
	Carcinoma, Non-Small-Cell Lung	0.5561
	Hepatitis, Chronic	0.5547
	Hyperinsulinism	0.5536
	Glioma	0.5510

Myotonic Disorders, and the definition of *Myotonic Disorders* presents its close relationships with *Myotonia Congenita* and *Myotonic Dystrophy*. As a contrast, *Vitiligo* is a type of skin disorder and *Kidney Calculi* is a disease originated in human kidneys. Both of them have not been found to have any associations with the above three targeted diseases. As shown in Table 4, the high similarity scores between targeted diseases and low similarity scores between the contrast disease pairs demonstrate that our method is effective in detecting similar diseases.

4.4 Parameter Sensitivity

Several parameters were tested to evaluate their impacts on *RADAR* measured by AUC score, including α, δ, and λ, which refer to the decaying weight factor, the length of random walk per node, and the embedding dimension, respectively. In this part, all the experiments were performed across all datasets (*"DIN-C"* and *"DIN-P"*) and two similarity metrics (meta-path-based similarity and neighborhood-based similarity) were both adopted. Their performances are presented in Figure 5.

As shown in Figure 5(a), *RADAR* did the best when $\alpha = 0.3$, while the other performances did not vary a lot from each other. It is observed from Figure 5(b) that the overall performance became stable when δ was over 80 and the best performance was achieved when δ is around 120. Similarly, despite the best performance was achieved when λ was around 32 as shown in Figure 5(c), there were just very trivial differences among all results.

The above results suggest that the proposed method is not critically sensitive to α, δ, and λ in general.

4.5 Comparison with Other Methods

RADAR was further compared with four typical methods (i.e., Resnik's method [10], Lin's method [11], Wang's

method [12], and *SemFunSim* [8]) as introduced in Section 1 for searching similar diseases. The similarity scores between 3525 diseases calculated by these methods were downloaded from the system *DincRNA*[2] [38]. Considering the identifiers of diseases in *DincRNA* are different from that in the datasets used by this paper, disease identifiers were first mapped and 877 shared diseases were then obtained. The mapping was also done on the benchmark disease set and 40 disease pairs were extracted. Next, 100 random disease sets were generated by the same way as introduced in Section 4.2. Based on the extracted benchmark disease set and the random disease sets, the aforementioned five methods were compared with on measuring similar diseases.

Table 5 presents the distribution of 100 iterations of AUC scores for all of the methods. From Table 5, it is observed that the performance of the first three methods was very close, all much lower than the performance of the latter ones. The possible reason for their poor performance may be that Resnik's method and Lin's method measure the semantic similarity of terms in certain general taxonomies rather than biomedical ontologies, and Wang's method is mainly designed for the measurement of GO terms. It remains unclear whether these three methods are suitable for measuring disease similarity. *SemFunSim* did the best among all. We attribute this to the support of the richness in DO semantics and gene functional associations to the interpretation of disease similarity. Though *SemFunSim* performed better than *RADAR*, we still managed to achieve a relatively good effect and the results in Figure 5 verified the stableness of our method.

TABLE 4
Similarity scores of selected diseases computed by our method

Group	Disease Pair	Score
Target	(Myotonic Disorders, Myotonia Congenita)	0.6452
	(Myotonic Disorders, Myotonic Dystrophy)	0.5826
	(Myotonia Congenita, Myotonic Dystrophy)	0.4561
Constrast 1	(Myotonia Congenita, Vitiligo)	0.0491
	(Myotonic Disorders, Vitiligo)	0.0209
	(Myotonic Dystrophy, Vitiligo)	0.0895
Constrast 2	(Myotonia Congenita, Kidney Calculi)	0.0820
	(Myotonic Disorders, Kidney Calculi)	0.0302
	(Myotonic Dystrophy, Kidney Calculi)	0.0105

TABLE 5
The distribution of 100 iterations of AUC scores for each method

Method	Max AUC	Min AUC	Average AUC
Resnik's [10]	0.6925	0.6148	0.6524
Lin's [11]	0.7351	0.6771	0.7048
Wang's [12]	0.7283	0.6439	0.6873
SemFunSim [8]	0.9565	0.8964	0.9289
RADAR	0.9020	0.8367	0.8736

5 Conclusion

Similar disease detection has significant implications in the field of biomedicine. Most of the current methods search similar diseases based on numerical data and are ontology dependent, while these requirements can not always be met. Besides, many of them evaluate disease similarity only under a single metric and only from a single data source, which lacks full consideration of multiple aspects.

We propose *RADAR*, a general ontology-independent and network-based framework for learning representations for diseases that capture their structural identities from a comprehensive perspective. Such representations were used to detect similar diseases. *RADAR* computes disease similarity under various metrics, and it is novel in discovering the relationship between disease pairs by maximizing the exploitation of associations among multiple disease-related data, without referring to any numerical information or ontologies. This may facilitate relevant studies and can be further improved to attain more accurate results.

The performance of *RADAR* was evaluated based on a benchmark disease set as well as 100 random disease sets. The high AUC suggested that *RADAR* is effective in discovering similar diseases and the sensitivity test showed that *RADAR* is generally insensitive to the selection of parameters.

In this paper, though we focus on the problem of computing disease similarity without reference to ontologies, it is also interesting to consider using the ontology to improve the performance of similar disease detection, which can be studied in the future work. As for future work, we also intend to focus on the following tasks. First, *RADAR* will be applied to other applications to further test its performance. The scalability of *RADAR* should be improved so that it can be smoothly applied to large-scale datasets. When combining similarities obtained under various similarity metrics, improved merging methods can be designed to better balance the importance of each metric and to better utilize current information. When building the multi-layer similarity network, techniques such as object recognition and object matching may be adopted to allow diseases from various data sources with different representations to match with each other. Strategies of random walks could be further improved. Improvement may also be made to allow real-time update of the multi-layer network when a new data source is added.

2. http://bio-annotation.cn:18080/DincRNAClient/

Acknowledgments

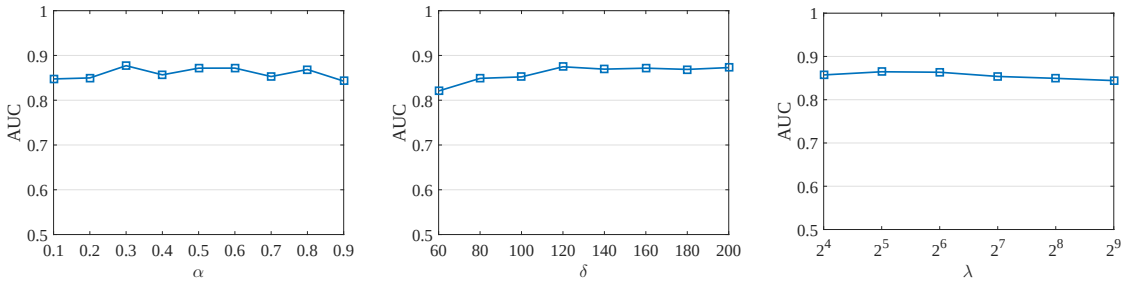
This work was supported in part by National Natural Science Foundation of China (61572332, 81473446), Fundamental Research Funds for the Central Universities (2016SCU04A22), China Postdoctoral Science Foundation (2016T90850), and COST ACTION Open Multiscale Systems Medicine (CA15120).

References

[1] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[2] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, pp. 1070–1074, 2014.

[3] X. Chen, C. C. Yan, C. Luo, W. Ji, Y. Zhang, and Q. Dai, "Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity," *Scientific Reports*, vol. 5, p. 11338, 2015.



(a) AUC w.r.t. α (b) AUC w.r.t. δ (c) AUC w.r.t. λ

Fig. 5. Performance analysis of parameters α, σ, and λ

[4] J. Sun, H. Shi, Z. Wang, C. Zhang, L. Liu, L. Wang, W. He, D. Hao, S. Liu, and M. Zhou, "Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network," *Mol. BioSyst*, vol. 10, no. 8, pp. 2074–2081, 2014.

[5] M. Zhou, X. Wang, J. Li, D. Hao, Z. Wang, H. Shi, H. Lu, H. Zhou, and J. Sun, "Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network," *Mol. BioSyst*, vol. 11, no. 3, pp. 760–769, 2015.

[6] L. Cheng, H. Shi, Z. Wang, Y. Hu, H. Yang, Z. Chen, J. Sun, and M. Zhou, "Intnetlncsim: an integrative network analysis method to infer human lncRNA functional similarity," *Oncotarget*, vol. 7, pp. 47 864–47 874, 2016.

[7] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Mol Syst Biol*, vol. 7, no. 1, p. 496, 2011.

[8] L. Cheng, J. Li, P. Ju, J. Peng, and Y. Wang, "SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association," *PLoS One*, vol. 9, no. 6, pp. 1–11, 2014.

[9] L. Cheng, J. Yue, Z. Wang, H. Shi, J. Sun, Y. Haixiu, Z. Shuo, Y. Hu, and M. Zhou, "DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs," *Scientific Reports*, vol. 6, p. 30024, 2016.

[10] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. of the 14th Int'l Joint Conf. on Artificial Intelligence*, 1995, pp. 448–453.

[11] D. Lin, "An information-theoretic definition of similarity," in *Proc. of the 15th Int'l Conf. on Machine Learning*, 1998, pp. 296–304.

[12] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[13] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[14] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. E. Parkinson, and L. M. Schriml, "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic Acids Research*, vol. 43, no. Database-Issue, pp. 1071–1078, 2015.

[15] H. J. Lowe and G. Barnett, "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches," *JAMA*, vol. 271, no. 14, pp. 1103–1108, 1994.

[16] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res*, vol. 33, pp. D514–D517, 2005.

[17] J. Li, B. Gong, X. Chen, T. Liu, C. Wu, F. Zhang, C. Li, X. Li, S. Rao, and X. Li, "DOSim: an R package for similarity between diseases based on Disease Ontology," *BMC Bioinformatics*, vol. 12, p. 266, 2011.

[18] S. Mathur and D. Dinakarpandian, "Automated ontological gene annotation for computing disease similarity," *Summit on Translat Bioinforma*, vol. 2010, pp. 12–16, 2010.

[19] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome Res*, vol. 21, no. 7, pp. 1109–1121, 2011.

[20] Y. Hu, M. Zhou, H. Shi, H. Ju, Q. Jiang, and L. Cheng, "Measuring disease similarity and predicting disease-related ncRNAs by a novel method," *BMC Med Genomics*, vol. 10, p. 71, 2017.

[21] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 4, pp. 905–915, July 2017.

[22] Y. Wang, L. Juan, Y. Chu, R. Wang, T. Zang, and Y. Wang, "FNSemSim: an improved disease similarity method based on network fusion," in *Proc. of the 2017 IEEE Int'l Conf. on Bioinformatics and Biomedicine*, 2017, pp. 630–633.

[23] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, pp. 333 EP –, Jan 2014.

[24] H. Wang, H. Zheng, J. Wang, C. Wang, and F. Wu, "Integrating omics data with a multiplex network-based approach for the identification of cancer subtypes," *IEEE Trans. on NanoBioscience*, vol. 15, no. 4, pp. 335–342, 2016.

[25] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[27] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. of the 20th ACM Int'l Conf. on Knowl. Discovery and Data Mining*, 2014, pp. 701–710.

[28] A. Grover and J. Leskovec, "Node2Vec: Scalable feature learning for networks," in *Proc. of the 22nd ACM Int'l Conf. on Knowl. Discovery and Data Mining*, 2016, pp. 855–864.

[29] R. Qin, L. Duan, H. Zheng, J. Li-Ling, K. Song, and X. Lan, "RADAR: representation learning across disease information networks for similar disease detection," in *Proc. of the 2018 IEEE Int'l Conf. on Bioinformatics and Biomedicine*, 2018, pp. 482–487.

[30] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: meta path-based top-k similarity search in heterogeneous information networks," *PVLDB*, vol. 4, no. 11, pp. 992–1003, 2011.

[31] L. F. R. Ribeiro, P. H. P. Saverese, and D. R. Figueiredo, "struc2vec: Learning node representations from structural identity," in *Proc. of the 23rd ACM Int'l Conf. on Knowl. Discovery and Data Mining*, 2017, pp. 385–394.

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

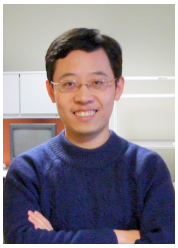
[33] S. Salvador and P. Chan, "FastDTW: toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, pp. 561–580, 2007.

[34] C. J. Grondin, D. Sciaky, J. Wiegers, R. J. Johnson, T. C. Wiegers, A. P. Davis, C. J. Mattingly, B. L. King, and R. McMorran, "The Comparative Toxicogenomics Database: update 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D972–D978, 09 2016.

[35] S. Suthram, J. T Dudley, A. P Chiang, R. Chen, T. Hastie, and A. Butte, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent

drug targets," *PLoS computational biology*, vol. 6, p. e1000662, 02 2010.
[36] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. B. Melton, "Semantic similarity and relatedness between clinical terms: An experimental study," *AMIA Annu Symp Proc*, vol. 2010, pp. 572–576, Nov 2010.
[37] B. Kong, K. P. Micklethwaite, S. Swaminathan, R. F. Kefford, and M. Carlino, "Autoimmune hemolytic anemia induced by anti-pd-1 therapy in metastatic melanoma," *Melanoma Research*, vol. 26, p. 1, 01 2016.
[38] J. Sun, L. Cheng, M. Zhou, Y. Hu, and Q. Jiang, "DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 01 2018.

Ruiqi Qin received her B.Sc. degree in Information Management and Information System from Beijing Forestry University in 2017. She is currently working towards her M.Sc. degree in the School of Computer Science at Sichuan University. Her research interests include bioinformatics and data mining.



Lei Duan received his B.Sc. and Ph.D. degrees both in Computer Science from Sichuan University in 2003 and 2008, respectively. He was a visiting Ph.D. student in the Department of Computer Science and Engineering at Wright State University from 2007 to 2008, and was a visiting scholar in the School of Computing Science at Simon Fraser University from 2012 to 2013. He is currently a Professor in the School of Computer Science at Sichuan University. His research interests include data mining, knowledge management, evolutionary computation, bioinformatics and health-informatics.

Huiru Zheng IEEE Senior Member, is a Professor of Computer Science with School of Computing at Ulster University, UK; and a Fellow of the UK Higher Education Academy. She was awarded a Ph.D. in Bioinformatics in 2003 and a Postgraduate Certificate in Teaching in Higher Education in 2005 from Ulster University. Prof. Zheng is an active researcher in bioinformatics and healthcare informatics. Within her broad interests in data mining, data integration, machine learning and healthcare decision support, Prof. Zheng has a particular research interest and expertise in integrative data analytics in the field of systems biology, and intelligent data analysis and assistive technology to support healthcare and independent living. She has published over 250 peer reviewed scientific research papers.

Jesse Li-Ling received his M.D. degree from West China University of Medical Sciences in 1993 and Ph.D. degree from University of Newcastle upon Tyne, U.K. in 2000. He conducted his postdoctoral research at Tsinghua University from 2001 to 2003. In 2007, he was promoted to full professor. Prof. Li-Ling is currently working at Sichuan University (State Key Laboratory of Biotherapy) and his main research interests include medical genetics, bioinformatics and Traditional Chinese Medicine.

Kaiwen Song received his B.Sc. degree in Computer Science from Sichuan University in 2017. He is currently working towards his M.Sc. degree in the School of Computer Science at Sichuan University. His research interest is data mining, especially network representation learning.

Yidan Zhang received her B.Sc. degrees in Biological Sciences and Software Engineering from Sichuan University in 2018. She is currently working towards her M.Sc. degree in the School of Computer Science at Sichuan University. Her research interests include bioinformatics and biomedicine.