RADAR: Representation Learning across Disease Information Networks for Similar Disease Detection

Ruiqi Qin*, Lei Duan*§, Huiru Zheng†, Jesse Li-Ling‡, Kaiwen Song*, Xuan Lan*
*School of Computer Science, Sichuan University, Chengdu, China
†School of Computing, Ulster University, Northern Ireland, United Kingdom
‡State Laboratory of Biotherapy, Sichuan University, Chengdu, China
§Corresponding author. Email: leiduan@scu.edu.cn

***Abstract*—Discovering similar diseases can provide valuable clues for revealing their pathogenesis and predicting therapeutic drugs. Current methods for the measurement of disease similarity are mostly based on either semantic associations among diseases or functional associations between disease and related genes. In either case, quantitative data are required. However, such data are not always available. Moreover, many of these methods only use a single metric to evaluate disease similarity from an individual data source, which may lead to biased conclusions lacking consideration of other aspects. In this study, we propose a novel framework, namely *RADAR*, for learning representations for diseases to measure their similarities. *RADAR* calculates disease similarity by different measurements fully based on the associations between diseases and other disease-related data, and constructs a multi-layer similarity network by integrating multiple disease similarity networks derived from multiple data sources in order to provide a comprehensive evaluation of disease similarities. A benchmark set and 90 random sets were used to assess the performance of *RADAR*. Experimental results demonstrated that *RADAR* is effective for detecting similar diseases.**

***Index Terms*—disease similarity, disease information network, representation learning**

I. Introduction

Identifying similar diseases is meaningful in the domain of biomedicine. Apart from dedicating to the study of diseases, disease similarity can be exploited to discover relations among many other data, such as inferring the relationship among microRNAs [1], [2], computing the similarity among long non-coding RNAs [3], [4], [5], [6], [7], and predicting therapeutic drugs for diseases [8], [9], [10], [11].

In general, there are two typical queries about similar diseases:

- *Top-k query*: searching top-k most similar diseases with respect to a given disease.
- *Similar pair query*: discovering the most similar disease pairs from a given disease set.

Similarity of pairwise diseases can be measured by a variety of aspects including pathogenesis and phenotypes. Existing methods of measuring disease similarities can be classified into three categories:

- *Semantics-based*: the disease similarity is measured by the quantitative information of semantic associations among diseases. The Disease Ontology (DO) [12] is the first standardized ontology for human diseases based on the disease terms collected from multiple sources such as *MeSH* [13] and OMIM [14]. Thereafter, many methods have been proposed to calculate disease similarity based on the terms in DO [15], [16], [17]. Recently, a system called *DOSim* [18] was constructed to calculate disease similarity by implementing 10 representative semantic similarity measures including the three methods mentioned in [15], [16] and [17].
- *Function-based*: the disease similarity is calculated by exploiting the relationship among disease-related genes [19], [20], [21]. For example, the process-similarity based (PSB) method measures disease similarity by involving the associations between genes based on Gene Ontology [22] terms.
- *Semantics + Function*: some researches combine DO-based similarity with the gene functional associations to calculate disease similarity. For example, *SemFunSim* [9] divides disease similarity into two parts, one obtained from a weighted gene interaction network from *HumanNet* [23], the other obtained from the relationship between pairwise diseases from DO. The similar idea was adopted by another method called *InfDisSim* [24]. Recently, an online system that implements five advanced methods was established to calculate similarity between disease sets [25].

It is worth noting that three disadvantages exist in these methods mentioned above. First, all of them compute disease similarity by exploiting some quantitative information about diseases and other disease-related objects, while the fact is that the precise numerical data describing the relationship among objects are not always available. Next, all these methods compute the disease similarity by a single metric, while it is very likely that the results will differ under different metrics. Thus, it is necessary to measure disease similarity under various metrics. Last but not the least, many of these methods evaluate disease similarity only from a single data source, which can lead to biased results that lack of full consideration. Since two diseases which are regarded as similar based on

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

one data source do not necessarily share the same similarity based on another. It would be more comprehensive to evaluate the disease similarity from a perspective where multiple data sources can all be considered.

To address these three challenges above, we propose a novel method, *RADAR* (short for representation learning across disease information networks), for similar disease detection. The characteristics of *RADAR* include: (1) it flexibly supports the two typical queries on similar diseases (top-k query and similar pair query); (2) it computes disease similarity based on the associations between diseases and other types of disease-related data; and (3) it evaluates disease similarity based on multiple data sources under orthogonal (i.e. semantic and structural) similarity metrics.

The main contributions of this work are as follows:

- We propose *RADAR*, a general framework for learning latent representations for diseases that reflect their similarities from a perspective where multiple data sources are considered at the same time, and such representations can be further applied to detect similar diseases.
- We show how *RADAR* computes the similarity between pairwise diseases under various similarity metrics, while solely based on the relationship between diseases and other types of related objects without references to any accurate numerical data.
- We evaluate *RADAR* on a benchmark set and random sets to demonstrate its effectiveness in discovering similar diseases.

The rest of the paper is organized as follows. The related work is reviewed in Section II, and the design of *RADAR* is presented in details in Section III. The experiments based on real-world data is carried out in Section IV and Section V concludes the paper.

II. Related Work

Many studies such as [26] [27] [28] have shown that by analyzing the similarity from a view where multiple data sources are all under consideration, the better performances can be achieved in discovering similar objects.

To compute the similarity between pairwise diseases is actually the process of building a similarity network. Fusing multiple similarity networks on the basis of samples was recently proposed in *SNF* [26]. It first builds several sample-similarity networks for 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. This may lead to lost of original information.

Another approach integrates various omics data based on a multiplex network to identify cancer subtypes [27]. Similar to *SNF*, it first constructs a patient-wise similarity network for each given dataset and then it uses a coupling strength to link each node with its counterparts in different network slices to build the multiplex networks.

Recently, a new method called *FNSemSim* [28] was proposed to improve calculating disease similarity. *FNSemSim* fuses two gene functional networks into one network and applies a random walk with restart to compute disease similarity. *FNSemSim* refers to the weights of pairwise genes given in the existing functional gene networks *FunCoup* and *HumanNet* as the connection weights in the fused functional association network.

The representation learning technique has been widely applied to a wide range of applications including image analysis, speech recognition and so on. It is prominent in capturing the essential semantics of objects by presenting them as dense vectors in low-dimensional space. Among miscellaneous representation learning models, Skip-Gram [29] has been proved to be very efficient in learning embeddings for textual data such as words and sentences.

Network representation learning was first proposed by *DeepWalk* [30], which 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 [31]. 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.

III. The Design of *RADAR*

To answer the two typical queries on similar diseases, the key point of *RADAR* is constructing the *disease similarity network*, which is an undirected graph describing the similarities among diseases. To begin with, 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 object types and R denotes the set of all relations. 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 [32] 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

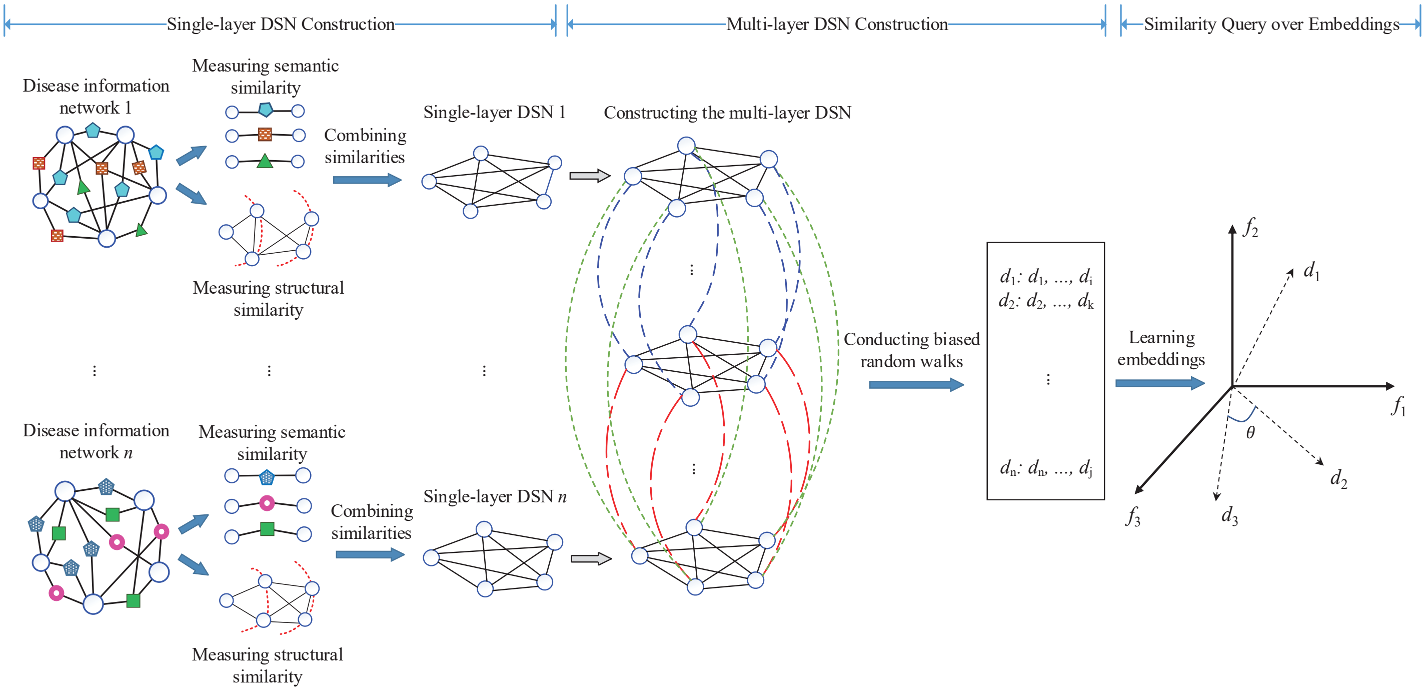
Single-layer DSN Construction
Multi-layer DSN Construction
Similarity Query over Embeddings
Disease information network 1
Measuring semantic similarity
Combining similarities
Single-layer DSN 1
Constructing the multi-layer DSN
Measuring structural similarity
Disease information network n
Measuring semantic similarity
Combining similarities
Single-layer DSN n
Measuring structural similarity
Conducting biased random walks
d_1: d_1, ..., d_i
d_2: d_2, ..., d_k
d_n: d_n, ..., d_j
Learning embeddings
f_2
d_1
f_1
θ
d_2
f_3
d_3

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

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, *RADAR* connects the same disease nodes existing in different DSN. That is, a multi-layer DSN is constructed, based on which the similar query can be performed.

Figure 1 illustrates the main steps of the *RADAR* framework.

Step 1 **Single-layer DSN Construction**: For a disease information network, calculate the semantic similarity and structural similarity between every disease pair by two similarity measurements and combine them to get one disease similarity network. (Section III-A)

Step 2 **Multi-layer DSN Construction**: Associate all the disease similarity networks obtained in the previous step into a multi-layer disease similarity network and conduct the biased random walks on it to generate a context for every disease. (Section III-B)

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

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

A. Single-layer Disease Similarity Network Construction

In a disease information network, two diseases can be connected through different paths. The *meta path* [32] is defined in as follows:

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

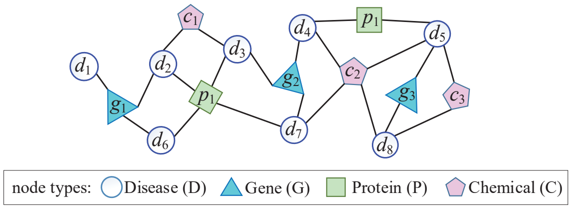

Fig. 2. An example of disease information network

Specifically, 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: Figure 2 shows an example of disease information network. There are in total four types of objects $\{D, G, P, C\}$. Multiple disease meta paths can be found such as "*D-G-D*" (short for "*Disease-Gene-Disease*"), indicating two diseases caused by the same gene, with disease path instances like "$d_1 - g_1 - d_2$" and "$d_3 - g_2 - d_4$". And the disease meta path "*D-C-D*" (short for "*Disease-Chemical-Disease*") indicates two diseases that can be both treated with the same chemical, with disease path instances such as "$d_7 - c_2 - d_5$" and "$d_8 - c_3 - d_5$".

Observation 1: In a disease information network, two diseases are more similar if (1) they share more relationships with other objects, and (2) they are connected via a short disease meta path.

Example 2: In the disease information network shown in Figure 2, d_8 is more similar to d_5 compared with d_7 to d_5, since d_8 and d_5 have three common paths $\{d_8 - c_2 - d_5, d_8 - c_3 - d_5, d_8 - g_3 - d_5\}$ while d_7 and d_5 only share one path $\{d_7 - c_2 - d_5\}$. Besides, d_8 and d_1 are unlikely to be similar compared with d_2 to d_1. This is because the disease meta path connecting d_1 and d_8 is too long, which indicates a loose relationship between them.

RADAR starts with the construction of a disease similarity network from a single disease information network under two orthogonal similarity measurements, i.e., semantic similarity and the structural similarity.

1) Measuring Semantic Similarity: In a heterogenous network, the meta path, which captures subtle semantics of the nodes that pass through [32], is usually used to imply the relationship between two nodes. A meta path-based similarity measure called *PathSim* [32] was proposed to find similar peer nodes in a network and has received fairly good effects. Analogously, *RADAR* searches similar diseases in a disease information network by a certain disease meta path that indicates the semantic relationship between two diseases.

The semantic similarity between disease $d_1 \in \mathcal{D}$ and disease $d_2 \in \mathcal{D}$ is defined as:

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

where $|Ins(\mathcal{P}_{d_1 \to d_2})|$ is the number of path instances from d_1 to d_2 under the given disease meta path, $|Ins(\mathcal{P}_{d_1 \to d_1})|$ is that from d_1 to d_1, and $|Ins(\mathcal{P}_{d_2 \to d_2})|$ is that from d_2 to d_2. Clearly, $0.0 \leq SemSim(d_1, d_2) \leq 1.0$.

By traversing the whole network, *RADAR* computes the semantic similarity for every disease pair based on the defined disease meta path according to Equation 1.

2) Measuring Structural Similarity: It may be noted that the disease path-based similarity measurement introduced above can only capture the relationship between two nodes on each end of the path and may fail to discover more potential similar nodes due to the constraint of the disease path.

Example 3: In Figure 2, d_2 and d_3 will not be considered to be similar under the disease path "*D-G-D*", even if they both share one related chemical and one protein.

An approach was given in [33] to calculate the similarity among nodes solely based on their structural identities in a network, which succeeds in discovering more similar node pairs. Thus, *RADAR* further assess the similarity between two diseases based on the structural identities of the disease network.

In a disease information network, the structure of a node d can be described by that of its vicinity, which is composed of a set of nodes belonging to the same object type as d. We call this set of nodes as the *ϵ-Neighbor Set* defined as:

Definition 5 (ϵ-Neighbor Set): In a disease information network, we denote $\ell_\epsilon(d)$ 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_2, $\ell_1(d_2) = \{d_1, d_3, d_6, d_7\}$ and $\ell_2(d_2) = \{d_4, d_5, d_8\}$. That is, d_2 has four 1-hop neighbors and three 2-hop neighbors.

Observation 2: In a disease information network, nearer neighbors make more contribution to describe the structural identity of a node.

Inspired by the idea of *Katz* centrality [34], a decaying weight factor is introduced to penalize the distant neighbors of a node.

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

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

where $StrDis_0(d_1, d_2) = 0$, and $\mathcal{T}(DS(\ell_\epsilon(d_1)), DS(\ell_\epsilon(d_2)))$ measures the *distance* between two ordered degree sequences $DS(\ell_\epsilon(d_1))$ and $DS(\ell_\epsilon(d_2))$. In *RADAR*, the Dynamic Time Warping (DTW) [35] method is adopted to calculate the distance between two sequences, as DTW has been proved to be very effective in handling numerical sequences by using some optimal alignment strategies to ensure the distance of two sequences is minimal. Specifically, 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:

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

where λ is a parameter preventing $dis(\cdot)$ being too large. (We set $\lambda = 0.5$ as in [33].)

For any disease node $d \in \mathcal{D}$, as the hop count ϵ increasing, the according ring 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 rings of neighbors of d into consideration when describing the structural identity of d. The decaying weight factor α as well as the hop count ϵ will be evaluated in the experiment to test their impacts on *RADAR*.

The natural exponential function is used to restrict the value of similarity in the range between 0.0 and 1.0 and the final structural similarity between d_1 and d_2 is

$$StrucSim(d_1, d_2) = e^{-StrDis_\epsilon(d_1, d_2)} \quad (4)$$

For any disease pair, *RADAR* first gets their degree sequences at each ring of neighborhoods starting from themselves to their ϵ-hop neighborhoods. Then *RADAR* computes their structural similarity according to Equation 4.

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 DSN. While the arithmetic mean of the two similarities is used to get the final disease similarity (ranging between 0.0 and 1.0), any other merging methods can be adopted by *RADAR* to combine the similarities obtained under any other metrics besides $SemSim$ and $StrucSim$.

B. 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 due to the characteristics of the disease information networks differ. To best keep the original information about every DSN, *RADAR* integrates all the DSNs into a multi-layer DSN. Specifically, each disease node locating in one DSN is associated with its counterpart in another by an edge with the weight set to 1.

Compared with *SNF* [26], the main advantage of *RADAR* is that the multi-layer DSN is constructed without lost of any information about the original DSN, while *SNF* fuses multiple similarity networks into a single one, only keeping the strong similarities but losing the weak ones.

Next, *RADAR* generates a context for each disease node by conducting random walks, particularly the biased random walks over the multi-layer DSN. Before each step, a random number will be generated, based on which the walker then decides whether to walk in the current network or to change to another one. If the walker stays in the current layer, it prefers walking towards a node that is more similar to the current node, i.e., it will walk along an edge among those with a bigger weight. If the walker changes the layer, 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 its similar nodes.

In summary, *RADAR* first starts from a random layer at a random disease node. Then the biased random walks are conducted for every disease node and its context will be produced accordingly in the end. By walking in the multi-layer DSN, the context generated is able to capture the similarity relationships for a disease node from multiple perspectives.

C. Similarity Query over Embeddings

The Skip-Gram model is adopted by *RADAR* to learn embeddings for all disease nodes based on the contexts generated by the biased random walks on the multi-layer DSN. In such way, the embeddings of nodes have successfully captured the similarities derived from multiple disease information networks. Note that besides Skip-Gram, any other representation learning models can be used in *RADAR* to learn embeddings for diseases.

Finally, the disease similarity can be calculated by applying a favorable distance measurement, such as the cosine, to their embeddings. 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
$\mathcal{G} \leftarrow \emptyset$
for $N \in \mathcal{N}$ do
    compute semantic similarity and structural similarity for
    every pair of diseases in $N$
    $G \leftarrow$ the single-layer DSN constructed from $N$
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{G\}$
end for
Connect the same disease nodes in different layers in $\mathcal{G}$
Generate contexts $Con$ by conducting the biased random
walks on $\mathcal{G}$
Learn embeddings $\mathcal{M}$ for all nodes from $Con$
$\mathcal{R} \leftarrow$ Perform similar disease query over $\mathcal{M}$
return $\mathcal{R}$
```

TABLE I
CHARACTERISTICS OF THE DISEASE INFORMATION NETWORKS

DIN	Type of Nodes	# of Nodes	# of Edges
Dis-Gene	disease	2818	117191
	genes	9658	
Dis-PTC	disease	916	11134
	chemicals	3516	

IV. EXPERIMENTS AND RESULTS DISCUSSION

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

A. Datasets

We applied *RADAR* to the datasets provided by [9] to search similar diseases. One dataset contains associations between diseases and genes, based on which a disease information network called "Dis-Gene" was built. The other one contains associations between diseases and potential therapeutic chemicals (PTC), based on which a disease information network called "Dis-PTC" was built. Table I lists the characteristics of these two disease information networks.

B. Effectiveness

We adopted the benchmark set given by [9] as the positive samples, which contains disease pairs that have been confirmed to be similar. Besides, 90 random sets were generated and adopted as the negative samples, which were regarded as dissimilar disease pairs. We applied *RADAR* on this benchmark set as well as the random sets to test its effectiveness of finding similar diseases. Throughout this experiment, the running parameters were set as $\alpha = 0.5$ and $\epsilon = 2$ in default.

To the best of our knowledge, *RADAR* is the first work that measures disease similarity from multiple data sources under the combination of semantic and structural similarity measurements. We verified the necessities of (1) measuring

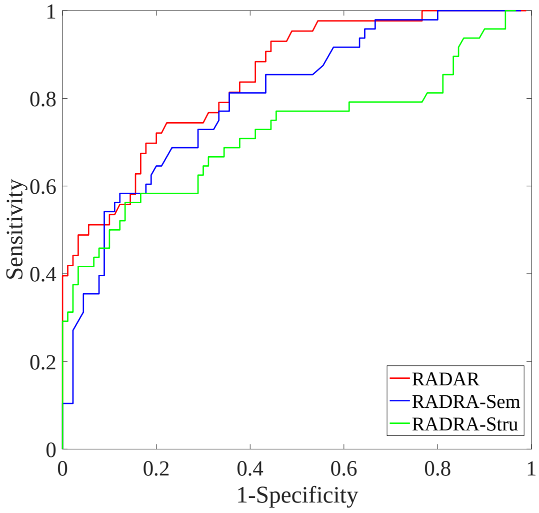

Fig. 3. The performance analysis of combining similarities under the semantic similarity measurement and the structural similarity measurement.

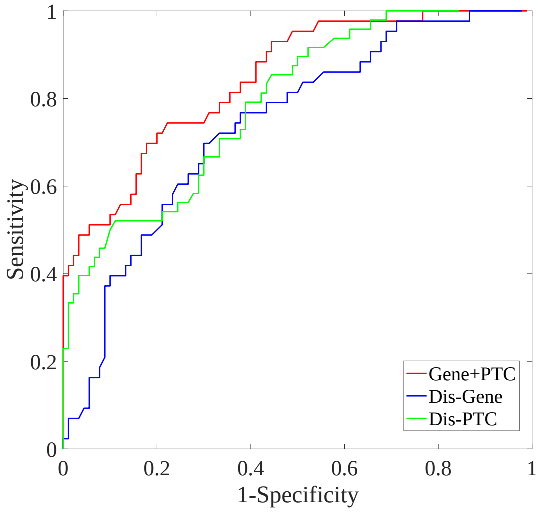

Fig. 4. The performance analysis of integrating multiple similarity networks.

disease similarity under two orthogonal metrics, (2) measuring disease similarity across different disease information networks, respectively.

First, we compared *RADAR* with its two variations. Specifically, we implemented two versions of *RADAR*. One only computed the semantic similarity (Equation 1), which was called *"RADRA-Sem"*. The other one only computed the structural similarity (Equation 4), which was called *"RADRA-Stru"*.

Figure 3 illustrates the Receiver Operating Characteristic (ROC) curves drawn for *RADAR*, *"RADRA-Sem"* and *"RADRA-Stru"*, respectively. Clearly, *RADAR* achieved the best performance with an Area under the ROC Curve (AUC) of 0.8465, while *"RADRA-Sem"* and *"RADRA-Stru"* performed relatively poorly with the AUCs of 0.8005 and 0.7206, respectively. Nevertheless, all of them did much better than the random performance. This demonstrated that combining the similarities computed under two types of similarity metrics is more effective than only using one of them for detecting similar diseases.

Second, we applied *RADAR* on the *"Dis-Gene"* DIN, the *"Dis-PTC"* DIN, and both of the DINs, respectively, to evaluate the necessity of integrating different similarity networks obtained from different datasets.

Figure 4 shows the ROC curves drawn for *"Dis-Gene + Dis-PTC"*, *"Dis-Gene"* and *"Dis-PTC"*, respectively. Clearly, *"Dis-Gene + Dis-PTC"* achieved the best result, followed by *"Dis-Gene"* and *"Dis-PTC"* with the AUC of 0.7898 and 0.7380. This indicates that integrating multiple similarity networks is more effective than only using a single similarity network for detecting similar diseases.

From the results above, we can see that by adopting the combination of various similarity measurements and integrating multiple similarity networks together, a better performance can be achieved for the detection of similar diseases.

Next, we evaluated the performances of *RADAR* in searching top-k most similar diseases for a given query and discovering the most similar disease pairs in a given disease set, respectively.

Several diseases were randomly selected from the benchmark set as the queries, and a result list comprising the top-5 most similar diseases to each of the queries was generated by *RADAR*. The results were recorded in Table II. Take Alzheimer's Disease for example, *RADAR* discovered that hypertension was most similar to it in the given disease set. A good amount of studies can be found to prove their close relationship. Next, we searched in the whole disease set to find the most similar disease pairs. The top 10 most similar disease pairs found by *RADAR* in the given disease set were recorded in Table III. These two sets of results above verified that *RADAR* performed well in fulfilling the tasks of top-k query and similar pair query.

C. Running Parameter Analysis

1) α: We first evaluated the impact of assigning the decaying weight factor with different values when computing the structural similarity as introduced in Section III-A2. The ROC curves were presented in Figure 5. As expected, despite *RADAR* did the best when $\alpha = 0.3$, there were just very trivial differences among all results, which implies that *RADAR* is insensitive to α.

2) ϵ: We evaluated the hop count ϵ when measuring the structural similarity as described in Section III-A2. The ROC curves were shown in Figure 6. As presented, *RADAR* had the best performance when $\epsilon = 5$. This is probably because that a bigger vicinity is considered when ϵ is under this value compared with the others, and a bigger vicinity of a node covers a wider area of its neighbors, which can provide more information about the structural identities for the node. However, the difference among these results was very trivial,

TABLE II
TOP-k SIMILAR DISEASES FOR THE GIVEN QUERIES

Query	Top-5 results	Score
Alzheimer's Disease	*hypertension*	0.7223
	obesity	0.7119
	melanoma	0.6996
	cerebrovascular accident	0.6989
	ovarian cancer	0.6927
Diabetes Mellitus	*pre-eclampsia*	0.7579
	coronary heart disease	0.7462
	brain ischemia	0.7414
	leukemia	0.7285
	cerebrovascular accident	0.7162
Myocardial Infarction	*congestive heart failure*	0.7080
	melanoma	0.6833
	atherosclerosis	0.6744
	kidney disease	0.6631
	brain ischemia	0.6602
Rheumatoid Arthritis	*coronary arteriosclerosis*	0.6357
	juvenile rheumatoid arthritis	0.6312
	squamous cell carcinoma	0.6257
	atherosclerosis	0.6186
	diabetes mellitus	0.6168
Epilepsy	*vascular disease*	0.6504
	hypersensitivity reaction disease	0.6466
	liver disease	0.6390
	hypertrophic cardiomyopathy	0.6311
	glucose intolerance	0.6157

TABLE III
THE MOST SIMILAR DISEASE PAIRS

Similar Disease Pair	Score
(melanoma, atherosclerosis)	0.8029
(neurodermatitis, functional colonic disease)	0.7908
(acrodermatitis enteropathica, acrodermatitis)	0.7861
(retrograde amnesia, alcoholic psychosis)	0.7829
(urinary bladder cancer, squamous cell carcinoma)	0.7805
(neurodermatitis, alcoholic psychosis)	0.7777
(brain ischemia, coronary heart disease)	0.7645
(leukemia, squamous cell carcinoma)	0.7644
(ovarian cancer, squamous cell carcinoma)	0.7628
(cerebrovascular accident, squamous cell carcinoma)	0.7626

which means the value of ϵ has not exerted so big influence on measuring structural similarity.

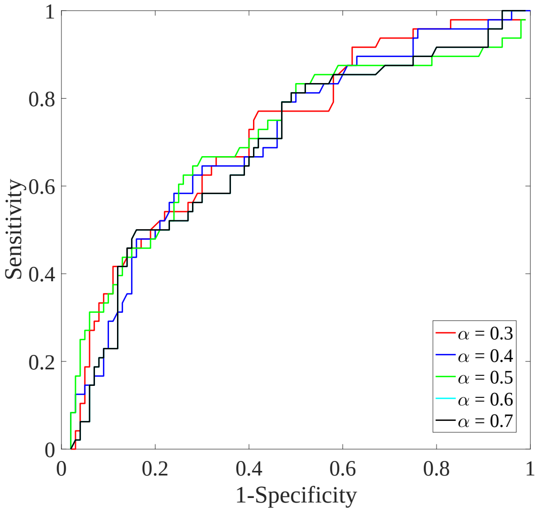

Fig. 5. Analysis of the parameter α

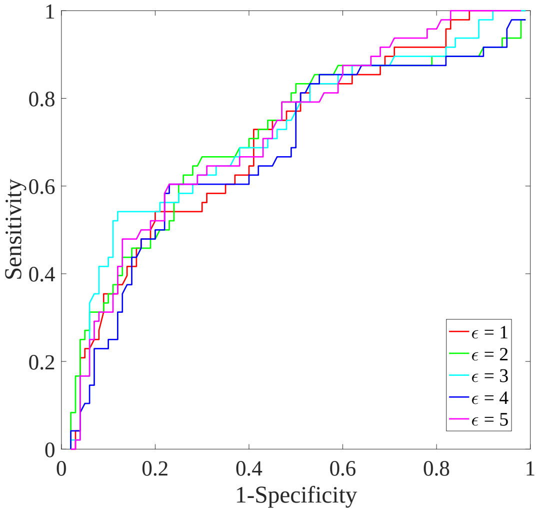

Fig. 6. Analysis of the parameter ϵ

V. CONCLUSION

Similar disease detection is an important issue in the field of biomedicine. Most of the existing methods search similar diseases based on numerical data, but this requirement can not always be met. Besides, many of them evaluate disease similarity only under a single metric and only from a single data source.

We propose *RADAR*, a general framework for learning representations for diseases that capture their semantics and structural identities from a more comprehensive perspective. Such representations can be used to detect similar diseases. *RADAR* is unique in computing disease similarity under various metrics, solely based on the associations between diseases and other types of biomedical objects without referring to any numerical information.

The performance of *RADAR* was evaluated based on a benchmark set as well as 90 random sets. The high AUC (0.8465) indicates *RADAR* is effective in discovering similar diseases. Besides, *RADAR* provides a novel way to discover the relationship between disease pairs by maximizing the exploitation of associations among multiple disease-related data. This may facilitate relevant studies and can be further improved to achieve more accurate results.

For the future work, we will focus on the following tasks to improve *RADAR*. The scalability of *RADAR* can be improved so that it can be applied to large-scale datasets smoothly. When combining similarities obtained under various similarity metrics, an improved merging method can be designed to better balance the importance of those metrics. When building the multi-layer similarity network, advanced 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. Improvement may also be made to allow real-time update of the multi-layer network when a new data source is added.

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, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Scientific Reports*, vol. 5, p. 13186, 2015.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] L. Cheng, J. Sun, W. Xu, L. Dong, Y. Hu, and M. Zhou, "OAHG: an integrated resource for annotating human genes with multi-level ontologies," *Scientific Reports*, vol. 6, p. 34820, 2016.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

[20] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, A. J. Butte, and Y. Ofran, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," *PLoS Comput Biol*, vol. 6, no. 2, 2010.

[21] S. Mathur and D. Dinakarpandian, "Finding disease similarity based on implicit semantic similarity," *Journal of Biomedical Informatics*, vol. 45, no. 2, pp. 363–371, 2012.

[22] 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.

[23] 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.

[24] 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.

[25] Y. Hu, L. Zhao, Z. Liu, H. Ju, H. Shi, P. Xu, Y. Wang, and L. Cheng, "DisSetSim: An online system for calculating similarity between disease sets," in *Proc. of the 2016 IEEE Int'l Conf. on Bioinformatics and Biomedicine*, 2016, pp. 1641–1652.

[26] 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, article.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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