Predicting Disease-related Associations by Heterogeneous Network Embedding

Yun Xiong
Shanghai Key Laboratory of Data Science
School of Computer Science, Fudan University
Shanghai, China, yunx@fudan.edu.cn

Lu Ruan
Shanghai Key Laboratory of Data Science
School of Computer Science, Fudan University
Shanghai, China, lruan15@fudan.edu.cn

Mengjie Guo
Shanghai Key Laboratory of Data Science
School of Computer Science, *Fudan University*
Shanghai, China, 17210240006@fudan.edu.cn

Chunlei Tang
Brigham and Women's Hospital
Harvard Medical School
Boston, MA, USA, ctang5@partners.org

Xiangnan Kong
Department of Computer Science,
Worcester Polytechnic Institute
Worcester, USA, xkong@wpi.edu

Yangyong Zhu
Shanghai Key Laboratory of Data Science, School of Computer Science
Fudan University, Shanghai, China, yyzhu@fudan.edu.cn

Wei Wang
Scalable Analytics Institute (ScAi), University of California
Los Angeles, USA, weiwang@cs.ucla.edu

***Abstract*—Elucidating biological mechanisms underlying complex diseases is an important goal in biomedical research. Recent advances in biological technology have enabled the generation of massive volume of data in genomics, transcriptomics, proteomics, epigenomics, metagenomics, metabolomics, nutriomics, etc., leading to the emergence of systems biology approach to investigating complex diseases. However, most of the data remain underutilized after their initial acquisition and analysis. There is a growing gap between the generation of the multifaceted data and our ability to integrate and analyze them.**

Inspired by the observation that many of the aforementioned data can be represented by networks, we propose a network-based model to encapsulate the rich information provided in each database and to connect across different databases. We integrate several public databases to construct a heterogeneous network in which nodes are entities such as genes, miRNAs, diseases, and edges represent known relationships between them. One fundamental challenge is how to perform meaningful analysis on such network, overcoming the intrinsic heterogeneity. We propose a network embedding method to learn a low-dimensional vector space that best preserves the known relationships between entities. Based on the learned vector representations, entities that are close to each other but currently do not have known direct connections, are likely to have an association and therefore are good candidates for future investigation.

In the experiments, we construct a heterogeneous network of genes, miRNAs and diseases using data from six public databases. To evaluate the performance of the proposed method, we predict disease-gene and disease-miRNA associations. Comparison of our novel method with several state-of-the-art methods clearly demonstrates the advantage of our method, as it is the only one that takes full advantage of the rich contextual information provided by the heterogeneous network. The encouraging results suggest that our method can provide help in identifying new hypotheses to guide future research.

***Index Terms*—Heterogeneous Network, Network Embedding, Disease-related Association Prediction**

I. Introduction

Recent technological advances have equipped scientists with the abilities to generate and analyze a massive volume of data to better elucidate biological mechanisms underlying complex diseases [1]. This has led to the development of many large databases to store and organize the accumulated data and knowledge, many of which were produced and maintained by large collaborative efforts. For example, the DisGeNET database [2] integrates human gene-disease associations from various expert-curated databases [3]–[6]. The miRNet database [7] collects data from eleven miRNA-disease datasets [8], [9]. A recent survey reviewed that there are more than 500 such databases that are actively maintained. Many of these databases also provide observed and/or derived knowledge of relationships between biological entities and diseases. For example, the Human Reference Protein Database (HPRD) [10] maintains the protein-protein interaction network; the MISIM database [11] has the miRNA similarity network; the MimMiner [12] provides a disease similarity network. Accurate modeling of complex biological interactions requires a systems approach to simultaneously consider these multifaceted data, including but not limited to genes [13], miRNAs [14], proteins [15], drugs [16], side-effects [17]. It may not only bring mechanical insights to deepen our understanding of complex diseases but also help us to identify new hypotheses to guide future research and exploration. Even though there has been significant progress made by several large consortia such as ENCODE, GTEx, we observe a growing gap between our abilities in generating data and our abilities in analyzing, integrating, and interpreting the data. Most current studies usually focus on data generated in a controlled environment by themselves or by collaborators from the same consortium, to ensure that data are generated under homogeneous settings and are directly comparable. Consequently, data generated from prior studies as well as the derived knowledge (maintained in public databases) remain largely underutilized. A fundamental challenge is the heterogeneity in data types, experimental technologies, and settings. In this paper, we propose a network-

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

based analytics model to overcome this challenge.

Our method was inspired by the observation that many of the aforementioned data can be represented by networks in which the nodes are entities such as genes, miRNAs, proteins, diseases, and the edges represent known relationships between these entities. As there are many different types of entities, the relationships may also be of different types (e.g. gene-disease association, protein-protein interactions). Both nodes and edges may have auxiliary attributes (or weights) further depicting the properties of corresponding entities and relationships. In order to make full use of the information provided by the constructed network, we propose to employ the network embedding method [18], [19] which has been successfully demonstrated its power in detecting and predicting relationships between individuals in a social network. Network embedding produces a low-dimensional vector space that best preserves the known relationships between entities. Each entity (e.g., gene, miRNA or disease) is represented by a vector and mapped to a point in the embedding space. The stronger the relationship(s) between two entities, the closer they are in the embedding space. Figure 1(a) shows a sub-network of one disease (i.e., prostate cancer), two genes (i.e., ATM, ZNF804A), two miRNAs (hsa-mir-21, hsa-mir-223), as well as their known connections (represented by solid edges) to other diseases, genes, and miRNAs. Figure 1(b) shows a two-dimensional projection of a small region surrounding prostate cancer in the network embedding space in which genes and miRNAs that are known to be associated with prostate cancer are distributed within close proximity to prostate cancer. The four dashed edges indicate the top two genes and two miRNAs predicted to be associated with prostate cancer by our method.

Learning network embedding on aforementioned heterogeneous networks faces several challenges. The nodes may represent entities of disparate nature. Edges between them may represent vastly different relationships, each may be with different weight or confidence. Traditional network embedding methods [18], [19] were designed for homogeneous networks and thus may not be able to incorporate this rich information. To tackle these challenges, we propose HeteWalk, which employs a meta path-controlled random walk to learn network embedding for a heterogeneous network. The meta paths are used to capture rich contextual information provided by the heterogeneous network. The random walk procedure on the network is controlled by both the meta paths and edge attributes/weights. In the learned embedding space, entities that are in close proximity but currently do not have direct connections are the ones of high potential of being associated and thus are good candidates for future research by scientists.

To evaluate the performance of the proposed method, we construct a heterogeneous network of genes, miRNAs, and diseases using data from six public databases and perform two disease-related prediction tasks, which are gene-disease and miRNA-disease association prediction. We compare our method with the state-of-the-art disease-related association prediction methods and network embedding methods. We find that our method outperforms all alternative methods. We also observe that including additional datasets will always improve the accuracy of the prediction by our method. We also find that many predictions made by our method are confirmed in the latest miRNet dataset [7], which further demonstrates the effectiveness of our methods in guiding biological experiments to identify novel disease-related associations.

TABLE I
DESCRIPTION OF THE CONSTRUCTED HETEROGENEOUS NETWORK

network		#links	weight	source
Gene (proteins) interaction network	G - G	39,240	1	HPRD [10]
microRNA similarity network	M - M	56,289	0 to 1	MISIM [11]
Disease phenotype similarity network	D - D	3,162,016	0 to 1	MimMiner [12]
Gene-Disease association network	G - D	19,714	0 to 1	DisGeNET [2]
Gene-miRNA interaction network	G - M	21,259	0.3 or 1	miRTarBase [20]
miRNA-Disease association network	M - D	878	1	Chen et al. [8] and miR2Disease [7]

The rest of the paper is organized as follows. Section II introduces the materials and methods proposed in this paper. Experiments and results are shown in Section III. Finally, the conclusion is in Section IV.

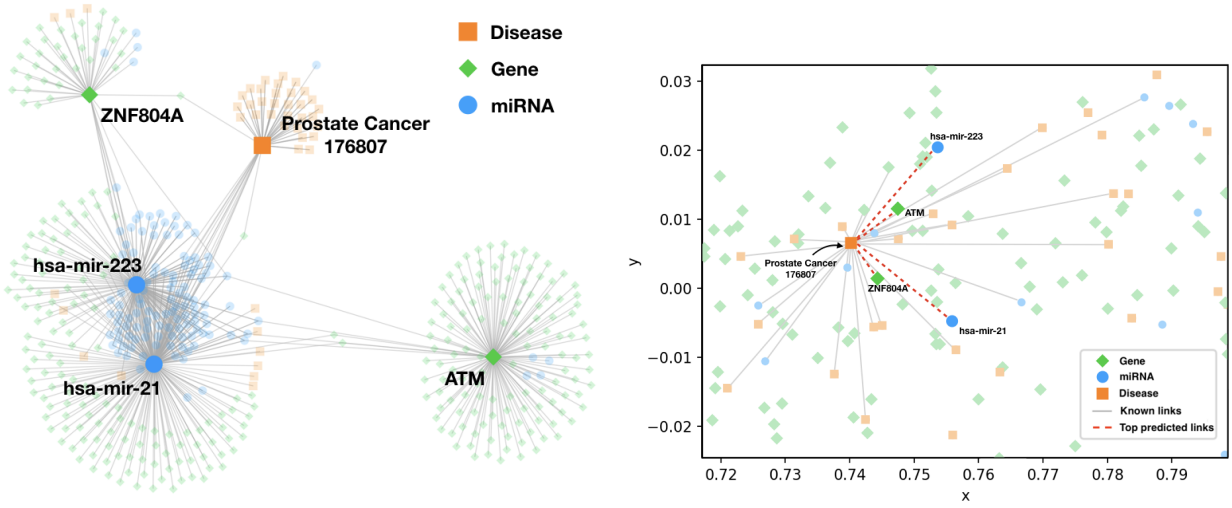
II. MATERIALS AND METHODS

A. Network Construction

Although there has been a large number of databases to store and organize the accumulated biological data, data generated from prior studies remains largely underutilized. One fundamental challenge is the heterogeneity in data types, experimental technologies and settings. In this section, we present how to construct a weighted heterogeneous network to integrate data from different sources. This serves as the foundation of all subsequent analysis.

1) Datasets: In this paper, we use data from six public databases to illustrate the concept and demonstrate the utility of our method. These six databases provide the similarity networks and association networks between three types of biological entities: genes, miRNAs and diseases.

- Gene (proteins) interaction network: 39,240 protein-protein interactions (PPI) are obtained from the Human Protein Reference Database (HPRD) [10] that was manually curated by expert biologists extracted from literature. For any two proteins that have direct interactions, their corresponding protein-coding genes are connected by an edge with weight 1.0 in the HPRD network.
- miRNA similarity network: The miRNA similarities are obtained from the MISIM database [11], which contains the pairwise functional similarity of 271 miRNAs. The similarity score of each link ranges from 0 to 1.
- Disease phenotype similarity network: The disease similarities for human were derived from the MimMiner [12], which uses a text-mining approach to classify human phenotypes from the Online Mendelian Inheritance in Man (OMIM) database [21]. Each link is associated with a score between 0 and 1 to indicate their similarity.
- Gene-Disease association network: This network is extracted from DisGeNET database [2], which integrates



(a) Sub-network Relation Visualization (b) Network Embedding Visualization

Fig. 1. An illustration of heterogeneous network embedding. In the left figure, we plot one disease i.e., prostate cancer), two genes (i.e., ATM, ZNF804A), two miRNAs (hsa-mir-21, hsa-mir-223), as well as their known connections (represented by solid edges) to other diseases, genes, and miRNAs. In the right, we plot their node vectors in the two dimensional projection of a small region surrounding prostate cancer after applying network embedding. Different types of instances are projected into the same vector space. The four dashed edges indicate the top two genes and two miRNAs which have no direct links but of high potential to be associated with prostate cancer predicted by our proposed method.

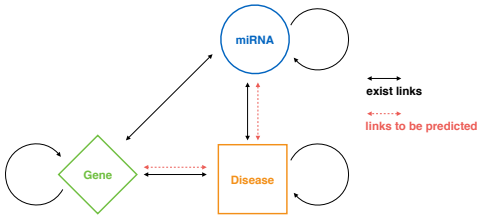


Fig. 2. Network Schema.

human gene-disease associations from various expert-curated databases. We use 19,714 entries whose disease phenotypes can be linked to OMIM terms. Each association has a score between 0 and 1 according to confidence.

- Gene-miRNA interaction network: The miRTarBase database [20] provides gene-miRNA interactions collected by manually surveying pertinent literature related to functional studies of miRNAs. The collected interactions are validated experimentally by reporter assay, western blot, microarray or next-generation sequencing experiments. When constructing the network, the weights of 7,269 interactions supported by strong experimental evidences (Reporter assay or Western blot) are set to 1, while the weights of 13,990 interactions supported by weak experimental evidences (Microarray or pSILAC) are set to 0.3.
- miRNA-Disease association network: We integrate two datasets to generate this network. The first one contains 242 miRNA-disease phenotype associations provided by Chen et al. [8]. The second one is extracted from miRNet datasets [7], which contain a large collection of verified miRNA-disease associations integrated from miR2Disease [22], HMDD [23] and Phenomir [24]. We extract the records whose disease names can be mapped to their corresponding OMIM ids to get 666 miRNA-disease associations. After removing duplicated records, there are 878 miRNA-disease associations within 267 miRNAs and 59 diseases in total. All the weights are set to 1.0 since the associations have been verified at a high confidence level.

2) Weighted Heterogeneous Network: In the aforementioned networks, genes are represented by their gene symbols in HPRD [10], miRNAs are represented by their names and disease phenotypes are represented by their corresponding OMIM ids [21]. We construct a heterogeneous network by connecting all six networks via shared nodes. The description of the constructed heterogeneous network is summarized in Table I. The schema of the constructed heterogeneous network is illustrated in Figure 2. It contains three types of nodes, where squares are diseases, circles are miRNAs, and rhombuses are genes. The solid lines indicate the links observed from the datasets, while the red dashed lines are the associations we want to predict, including gene-disease associations and miRNA-disease associations.

The constructed heterogeneous network not only contains different types of entities, but also provides relationships with different weights or confidence levels. Note that the weights on different types of links are not directly comparable since they are derived from different datasets. For example, if the weight between *prostate cancer* (disease) and *ATM* (gene) is higher than *prostate cancer* and *hsa-mir-21* (miRNA), it does not necessarily indicate *ATM* is more related to *prostate cancer* than *hsa-mir-21*. Therefore, given a weighted heterogeneous network, our next step is to generate a feature space in which similarities and associations between entities of heterogeneous types become quantitatively measurable and predictable.

B. HeteWalk

Our method applies network embedding method to learn a low dimensional vector representation for each node in the network, which preserves the known relationships. A key intuition of our approach is that genes (or miRNAs, diseases)

that are of close proximity in the heterogeneous network are more likely to be related. For example, a gene that plays a significant role in one disease may likely to play a similar role in a similar disease. This insight enables us to predict unknown disease-related associations based on the known relationships in the network.

1) Learning the Network Embedding: Given a network, network embedding aims to learn a low dimensional vector representation of nodes which preserves contextual similarities of the nodes. The idea originated from the concept of word embedding in natural language processing, that learns a feature space in which words surrounded by similar contexts tend to carry similar semantics and have similar feature representations [25]. Recently, several network embedding methods [18], [19] demonstrated superb performance in tasks such as node classification and link prediction. In order to learn a good network embedding, we need to maximize the co-occurrence probability of a node and its related nodes (i.e., those connected with direct links) in the network. Given a node v_i and the set of related nodes $N(v_i)$, we want to maximize the probability that $N(v_i)$ appear in close proximity from node v_i. If we assume that the probability of observing each node is independent of observing any other node, our goal is to maximize the objective function:

$$\prod_{v_i \in \mathcal{V}} \Pr(N(v_i)|v_i) = \prod_{v_i \in \mathcal{V}} \prod_{v_j \in N(v_i)} \Pr(v_j|v_i) \tag{1}$$

The conditional probability is defined by the softmax function:

$$\Pr(v_j|v_i) = \frac{e^{\vec{x_i} \cdot \vec{c_j}}}{\sum_{k \in V} e^{\vec{x_i} \cdot \vec{c_k}}} \tag{2}$$

where V is the set of nodes in the network. $\vec{x_i}$ is the vector representation of node v_i and $\vec{c_j}$ is the auxiliary context vector of node v_j.

Both vectors are latent d-dimensional vectors to be learned. Most of the existing network embedding models, however, assume homogeneous networks in which all nodes are of the same type. How to compute meaningful network embedding on heterogeneous networks remains an open problem. In our setting, a disease node may be linked to other diseases, genes, miRNAs. To better capture the contextual properties of a node in such a network, it may be beneficial to go beyond related nodes connected with direct links. For example, if a gene and a disease can be connected by a path of multiple links such as $Gene \xrightarrow{similar\ with} Gene \xrightarrow{associated\ with} Disease$ or $Gene \xrightarrow{associated\ with} miRNA \xrightarrow{associated\ with} Disease$, they may be related as well. Next, we show how such paths may be utilized in learning network embedding for a heterogeneous network.

2) Meta Path Controlled Random Walks: A *meta path* $\mathcal{P}$ is used to describe a category of relationships between two types of nodes, which can be denoted in the form of $\mathcal{A}_1 \rightarrow \mathcal{A}_2 \rightarrow \cdots \rightarrow \mathcal{A}_m$, where $\mathcal{A}_i$ represents an object type (e.g., gene, miRNA or disease) [26]. Two nodes in a heterogeneous network may have multiple relationships that can be categorized by several meta paths.

For example, the meta path $Gene \xrightarrow{assoc} Disease$ describes a directly linked gene-disease association; the meta path $Gene \xrightarrow{sim} Gene \xrightarrow{assoc} Disease$ describes a relationship that a gene is similar to another gene has known to be associated with the disease; the meta path $Gene \xrightarrow{assoc} miRNA \xrightarrow{assoc} Disease$ describes that a gene and a disease are both related to a given miRNA.

Meta paths are an effective means to characterize indirect relationships between certain types of entities. The number of distinct meta paths grows exponentially with the number of entity and relationship types as well as the meta path length, providing a rich language describing contextual properties of the network. To further take consideration of the link weight, for each meta path, we employ a *meta path controlled random walk* to explore the related nodes. A meta path defines the type of node to be visited at each step, while the link weights determine the probability of which each node is selected. Starting at node v_i, given a meta path $\mathcal{P} = \mathcal{A}_1 \rightarrow \mathcal{A}_2 \rightarrow \cdots \rightarrow \mathcal{A}_m$, the random walk will only visit a node of type $\mathcal{A}_k$ at the kth step. If multiple nodes of type $\mathcal{A}_k$ exist, we select a node randomly with a probability proportional to its link weight. The higher the link weight, the higher the probability to be selected. For the current node v_i with type $\mathcal{A}_k$, the transition probability to another node v_j is defined as:

$$\Pr(v_j|v_i;\mathcal{P}) = \begin{cases} 0 & (v_i, v_j) \notin E \\ 0 & (v_i, v_j) \in E, \phi(v_j) \neq \mathcal{A}_{k+1} \\ \frac{w_{ij}}{\sum_{\phi(v_k)=\mathcal{A}_{k+1}} w_{ik}} & (v_i, v_j) \in E, \phi(v_j) = \mathcal{A}_{k+1} \end{cases} \tag{3}$$

where $\phi(v_i)$ indicates the node type and w_{ij} is the link weight between v_i and v_j. Given a meta path, the random walk procedure generates a node sequence starting from each node.

In order to generate a sufficient number of node sequences, the random walk procedure will be repeated starting from each node. The number of executions of each meta path determines the number of related nodes explored by such meta path in the generated node sequences. In other words, we can adjust the importance of each meta path by controlling the number of node sequences generated from it, which is a tunable parameter in our method.

After we get a set of node sequences, we now learn the node vector representations. As described in Eq(2), the goal is to maximize the co-occurrence probability of each node and its related nodes. In other words, for nodes appearing in the same node sequence, we will update their node embeddings according to Eq(2). However, the number of node pairs in all node sequences is very large, so the evaluation of Eq(2) can be prohibitively expensive. Inspired by its success in word embedding methods, we apply negative sampling [27] and derive the following approximation term:

$$\log \Pr(v_j|v_i) = \log \sigma(\vec{x_i} \cdot \vec{x_j}) + \sum_{n=1}^{K} \mathbb{E}_{v_n \sim NEG(v_j)} \log \sigma(-\vec{x_i} \cdot \vec{x_n}) \tag{4}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ and $NEG(v_j)$ is the distribution from which a negative node v_j is selected. K is the size of negative samples.

For each node pair (v_i, v_j) appearing in the same node sequence, we randomly select K negative node pairs (v_i, v_N), where $v_N \neq v_j$ and $\phi(v_N) = \phi(v_j)$.

The model is updated by maximizing the score of positive sample (v_i, v_j) and minimizing the scores of all negative samples (v_i, v_N). For example, if we have a node sequence $(Gene1, Gene2, Disease1)$ generated by the meta path $Gene \xrightarrow{sim} Gene \xrightarrow{assoc} Disease$, there are 3 positive node pairs $(Gene_1, Gene_2)$, $(Gene_1, Disease_1)$, $(Gene_2, Disease_2)$. Take $(Gene_1, Disease_1)$ as an example. We then randomly choose K nodes whose types are disease, denoted as $Disease_{N_1}, \cdots, Disease_{N_K}$, where $Disease_{N_i} \neq Disease_1$. The positive sample $(Gene_1, Disease_1)$ and K negative samples $(Gene_1, Disease_{N_i})$ are fed into the model together and their corresponding node vectors are updated according to Eq(4) by Stochastic Gradient Descent (SGD) [28].

3) Predict Disease-related Associations: After we learn the network embedding, all types of nodes (genes, miRNAs and diseases) in the heterogeneous network are projected into the same vector space. We can now use the cosine distance to measure their relationships. For disease-related association prediction, if a disease vector and a gene/miRNA vector that do not have direct connection in the network are close to each other in the learned embedding space, they have high potential to be associated and thus are a good candidate for future research by scientists.

III. Results and Discussion

A. Comparison with other methods

To demonstrate the effectiveness of HeteWalk, we compare our method with several state-of-the-art methods. These methods are divided into two groups. Methods in the first group, including CATAPULT [5], HSMP and HSSVM [6], [9], are designed to predict Disease-Gene associations [5], [6] or Disease-miRNA associations [9]. All these methods are designed for detecting a specific type of associations (i.e., gene-disease or miRNA-disease). We adapt them to run on our gene-disease-miRNA heterogeneous network. CATAPULT uses a biased support vector machine in which the features are derived from paths of different lengths. Both HSMP and HSSVM use the HeteSim score [29] to measure the relevance between nodes. The HeteSim score of a given path between two nodes measures the reachability of the two nodes along that path. HSMP combines HeteSim scores of different paths with a constant that dampens the contributions from longer paths, while HSSVM method uses a supervised machine learning method to combine HeteSim scores.

The second type of method we compare to is DeepWalk [18] which is a popular network embedding method defined on a homogeneous network. Different from our method, DeepWalk ignores the heterogeneous information and uses a vanilla random walk procedure, although the network we constructed is heterogeneous.

Our method, HeteWalk, utilizes meta path-controlled random walks for the heterogeneous network embedding. The learned node vectors are used to predict entities (e.g., genes, miRNAs) that are of high potential associated with diseases.

TABLE II
META PATHS BETWEEN GENE-DISEASE AND MIRNA-DISEASE.

	With 2 types of nodes	With 3 types of nodes
gene-disease	Gene $\xrightarrow{assoc}$ Disease	Gene $\xrightarrow{assoc}$ miRNA $\xrightarrow{assoc}$ Disease
	Gene $\xrightarrow{sim}$ Gene $\xrightarrow{assoc}$ Disease	Gene $\xrightarrow{sim}$ Gene $\xrightarrow{assoc}$ miRNA $\xrightarrow{assoc}$ Disease
	Gene $\xrightarrow{assoc}$ Disease $\xrightarrow{sim}$ Disease	Gene $\xrightarrow{assoc}$ miRNA $\xrightarrow{sim}$ miRNA $\xrightarrow{assoc}$ Disease
	Gene $\xrightarrow{assoc}$ Disease $\xrightarrow{assoc}$ Gene $\xrightarrow{assoc}$ Disease	Gene $\xrightarrow{assoc}$ miRNA $\xrightarrow{assoc}$ Disease $\xrightarrow{sim}$ Disease
miRNA-disease	miRNA $\xrightarrow{assoc}$ Disease	miRNA $\xrightarrow{assoc}$ Gene $\xrightarrow{assoc}$ Disease
	miRNA $\xrightarrow{sim}$ miRNA $\xrightarrow{assoc}$ Disease	miRNA $\xrightarrow{assoc}$ Gene $\xrightarrow{sim}$ Gene $\xrightarrow{assoc}$ Disease
	miRNA $\xrightarrow{assoc}$ Disease $\xrightarrow{sim}$ Disease	miRNA $\xrightarrow{sim}$ miRNA $\xrightarrow{assoc}$ Gene $\xrightarrow{assoc}$ Disease
	miRNA $\xrightarrow{assoc}$ Disease $\xrightarrow{assoc}$ miRNA $\xrightarrow{assoc}$ Disease	miRNA $\xrightarrow{assoc}$ Gene $\xrightarrow{assoc}$ Disease $\xrightarrow{sim}$ Disease

B. Experiment settings

In the experiments, we evaluate the performance of predicting two types of association: gene-disease association and miRNA-disease association. We use the meta paths to generate node sequences. We set the vector dimension to be 128, the number of walks for each node and each meta path to be 10, and the size of negative samples to be 5. We demonstrate in Section III-E that the performance is not sensitive to the setting of vector dimensionality and number of walks per node. Except for DeepWalk that can only run on homogeneous networks, all other methods use the same set of meta paths. Considering meta path construction, we first extract all non-redundant meta paths for correlations of target entities type separately. Then we combine two or more to construct redundant meta paths. We only extract short meta paths with limited path length, because long meta paths are not quite useful in capturing the linkage structure [26]. The meta paths we extract are shown in Table II.

C. Effectiveness measurement

We partition the known disease-related associations randomly into 10 sets of equal size. In each experiment, a subset of them are used for training and the remaining ones are used for testing. We vary the training ratio from 50% to 90%. For each training ratio, the experiments are repeated 10 times and the average Area under Receiver Operating Characteristic curve (AUROC) score is reported. The results for each method with different training ratio are listed in Table III (miRNA-disease association prediction) and Table IV (gene-disease association prediction).

We observe that HeteWalk consistently achieves the best performance in both disease-related association prediction tasks for all training ratios. In the miRNA-disease association prediction, HeteWalk is able to achieve an almost perfect AUROC score 0.969 with 90% training data. In the gene-disease prediction, since we have a relatively large number of candidate associations, the best score of HeteWalk is 0.798.

HeteWalk outperforms other heterogeneous network-based disease prediction methods, including CATAPULT, HSMP, and HSSVM. Although these methods explore the heterogeneous network by the same meta paths as HeteWalk, they all extract simple features on path connectivity between two nodes. In contrary, HeteWalk optimizes on preserving known relationships between nodes by maximizing the co-occurrence probability of all node pairs in a node sequence generated by a meta path.

DeepWalk shows poor performance mainly because it's designed for homogeneous networks. When selecting the next

TABLE III
AUROC SCORE OF MIRNA-DISEASE ASSOCIATION PREDICTION

Method/Train ratio	50%	60%	70%	80%	90%
CATAPULT	0.811	0.833	0.843	0.867	0.877
HSMP	0.833	0.864	0.878	0.899	0.869
HSSVM	0.841	0.877	0.902	0.922	0.932
DeepWalk	0.498	0.511	0.534	0.611	0.677
HeteWalk	**0.937**	**0.951**	**0.953**	**0.946**	**0.969**

TABLE IV
AUROC SCORE OF GENE-DISEASE ASSOCIATION PREDICTION

Method/Train ratio	50%	60%	70%	80%	90%
CATAPULT	0.611	0.619	0.622	0.659	0.685
HSMP	0.621	0.625	0.679	0.708	0.747
HSSVM	0.609	0.653	0.693	0.734	0.779
DeepWalk	0.454	0.461	0.481	0.433	0.477
HeteWalk	**0.638**	**0.674**	**0.723**	**0.759**	**0.798**

node at each step of a random walk, it treats nodes and links of different types equally. Therefore, the embedding space fails to preserve the relationships between certain types of entities.

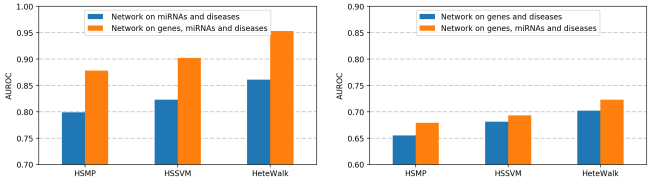
D. Power of heterogeneity

In this section, we investigate the ability of each method in handling heterogeneity and demonstrate the benefit of integrating multiple data sources. We generate another two sets of heterogeneous networks that contain only two types of nodes. In the gene-disease association prediction task, we only combine G-G, G-D and D-D networks listed in Table I. In the miRNA-disease association prediction task, we only use D-D, M-M, and D-M networks.

In this experiment, we use 3-fold cross-validation: we split the known disease-related associations into three groups of equal size, and each time use two groups for training and one for testing. The average score for each method is shown in Figure 3. We can observe a clear advantage of combining all networks together into a heterogeneous network, especially in the miRNA-disease association prediction tasks. That is because the known relations between miRNAs and diseases are sparse which alone may be insufficient to warrant reliable prediction. The gene-related datasets help to establish some indirect connections between miRNAs and diseases, which may be captured by the meta paths on the heterogeneous network. This suggests that integrating multifaceted data will deepen our understanding of complex diseases and further improve the prediction. In fact, even though we demonstrate the utility of HeteWalk on six databases, HeteWalk has the ability to integrate any number of data sources as long as they may be represented as a heterogeneous network. There is no limit on the number of entity types or relationship types.

E. Parameter sensitivity

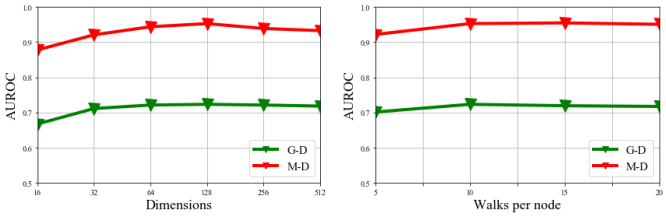
Following the same experimental setting in the 3-fold cross validation, we study the parameter sensitivity of HeteWalk measured by AUROC. Figure 4 shows the performance with



(a) miRNA-Disease Association Prediction (b) Gene-Disease Association Prediction

Fig. 3. AUROC results on different networks.

different embedding dimensions and different number of walks per node. From Figure 4(a), we observe that the best performance is achieved around 128 dimensions. Similarly, we find that the performance is almost stable when the number of execution is more than 10 walk per node. Both figures suggest that 128 embedding dimensions and 10 walks per node are optimal settings for network embedding in our study.



(a) AUROC vs. dimensions (b) AUROC vs. number of walks

Fig. 4. Parameter sensitivity

F. Top-ranked disease-related associations for selected diseases

In order to study which genes or miRNAs may play an important role in a certain disease, we show the top candidates predicted by HeteWalk for four disease phenotypes (In Table V), For each disease, the left column contains the top-ranked genes while top-ranked miRNAs are in the right. The numbers indicate their original ranking before known associations are removed in the results. The results are ranking according to the cosine distances between each disease and the candidate genes/miRNAs and the known associations are not shown.

Although the diseases may already have a lot of associated genes and miRNAs in the training set, the known associations are not always ranked high on the list. For example, insulin resistance (125853) has 33 known related genes, while only 5 of them are among the top-10 genes for the disease. This is because their link weights are relatively low in the network, showing a weak association with insulin resistance. Genes that have a complex relationship with insulin resistance can be captured by several meta paths and their distances in the embedding space are likely to be closer than some known associations.

In addition to insulin resistance, the results of other diseases also contain a large number of unknown associations with genes or miRNAs. Such results may help guide biological experiments to identify novel disease-related associations.

G. Validation and comparision of the top-ranked miRNA-disease associations prediction

To verify the effectiveness of our methods, we manually check the predicted miRNA-disease associations. We use the miRNet dataset [7] to verify the predicted disease-related miRNAs. The dataset contains a large collection of verified miRNA-disease associations from miR2Disease [22], HMDD [23] and Phenomir [24]. Since each disease is represented by disease name instead of OMIM id, we only incorporate part of the records (666 of 19,342) when constructing the heterogeneous network. We use the remaining ones to verify the top-ranked miRNA-disease associations predicted by HeteWalk.

TABLE V
TOP 10 UNKNOWN DISEASE-RELATED ASSOCIATIONS PREDICTED BY HETEWALK.

Leukemia OMIM: 601626				Alzheimer disease OMIM: 104300				Insulin resistance OMIM: 125853				Prostate cancer OMIM: 176807			
	Gene		*miRNA*		*Gene*		*miRNA*		*Gene*		*miRNA*		*Gene*		**miRNA**
2	TNF	3	hsa-mir-21	2	GRN	1	hsa-mir-223	1	BCR	1	hsa-mir-659	1	ATM	1	hsa-mir-223
4	APOE	4	hsa-mir-17	8	CHMP2B	2	hsa-mir-659	2	ABL1	2	hsa-mir-21	2	ZNF804A	2	hsa-mir-21
5	ATM	7	hsa-mir-146a	10	TNF	3	hsa-let-7c	4	ARID3B	3	hsa-mir-223	3	BEND2	4	hsa-mir-144
6	PRRX1	8	hsa-mir-510	12	CEBPA	4	hsa-mir-21	8	MAST1	4	hsa-let-7c	4	TBP	5	hsa-mir-331
7	CD81	10	hsa-mir-20b	13	ATM	5	hsa-mir-15a	9	CEBPA	5	hsa-mir-16-1	5	PLTP	6	hsa-mir-17
8	USP8	11	hsa-mir-331	15	PPARG	6	hsa-mir-16-1	11	CDH8	6	hsa-mir-15a	6	ELP5	8	hsa-mir-510
9	PPARG	12	hsa-mir-155	16	BCR	7	hsa-mir-17	12	ZNF609	7	hsa-mir-17	7	KLHL35	10	hsa-mir-143
10	IL1B	13	hsa-mir-143	17	ABL1	8	hsa-mir-155	13	TBP	8	hsa-mir-155	8	ENTPD6	11	hsa-mir-20b
11	SH2B1	14	hsa-mir-539	18	USP8	9	hsa-mir-510	14	IL1RAPL1	9	hsa-mir-146a	9	RBP2	12	hsa-mir-425
12	IL6	15	hsa-mir-192	19	HNF1B	11	hsa-let-7a-1	15	ENTPD6	10	hsa-mir-510	10	U2AF2	14	hsa-let-7a-1

In the experiment, we use the whole dataset in Table I to construct the heterogeneous network and run the algorithm to learn the embedding for each node. We select three miRNAs which have the largest number of verified records in the miRNet dataset and then report the top 10 diseases predicted to have associations to each miRNA in Table VI. We are able to confirm many of these predictions. There are 8, 7 and 6 confirmed associations for hsa-mir-21, hsa-let-7a-1, and hsa-mir-125b-1 respectively. This result further validates the effectiveness of our methods.

The first column in Table VI shows the ranking of each predicted disease among all associated diseases. The second column shows the diseases with predicted associations and their OMIM id. The third shows whether the predicted associations can be found in miRNet. For these three miRNAs (hsa-mir-21, hsa-let-7a-1, hsa-mir-125b-1), the number of their known disease associations in the training set is 7, 11 and 4 respectively. From the ranking results, we observe that not all known associations are ranked high. The reasons are two folds. (1) Some of the known associations have relatively low weights, indicating a weak relationship with the disease. (2) Although some diseases and miRNAs currently do not have direct connections in the training data, they are well connected with each other by multiple meta paths in the heterogeneous network. Such candidate diseases therefore are more related to the miRNAs than those diseases connected with low link weights, and they have a high probability to be predicted by HeteWalk.

We also compare the top 10 predicted disease phenotypes of 3 miRNAs returned by alternative methods. The results predicted by CATAPULT, HSMP and HSSVM are listed in Table VII, VIII and IX. In these tables, records verified by miRNet are in bold. The numbers indicate their original rankings. We omit known associations in the lists. We compare them with the results returned by HeteWalk in Table VI.

The predictions returned by CATAPULT (Table VII) contain a very high degree of overlap between the 3 miRNAs. We can see that Male germ cell tumor (273300) appears in all predicted candidate diseases of the 3 miRNAs with a high rank (top 3). Male germ cell tumor (273300), Nonmedullary Thyroid cancer 1(188550) and Enterocolitis(226150) also appear in all 3 lists. This is because CATAPULT has a bias towards the nodes with larger node degrees, and may therefore overlook important connections only specific to an individual miRNA.

The predictions returned by HSMP (Table VIII) and HSSVM (Table IX) have lower degree of overlap in the top-ranked predictions compared to CATAPULT. The associations verified by miRNet are in bold in each table, where we can find the number of verified predictions are 5, 5, 5 and 5, 6, 1 respectively, fewer than that of HeteWalk, which are 8, 7, 6.

IV. CONCLUSION

In this paper, we propose a novel method, HeteWalk, for disease-related association prediction. The method first constructs a heterogeneous network from multiple data sources and then learns a low-dimensional feature vector for each node in the network. The learned network embedding well captures the contextual properties of the network to allow for accurate prediction of disease-related associations. In the experiment, we evaluate our method on gene-disease association prediction and miRNA-disease association prediction tasks. We demonstrate the advantages of our method over alternative methods across multiple measurements. Although we use only three types of biological entities (i.e., genes, miRNAs, and diseases) from six databases to construct the network, HeteWalk is capable of incorporating any number of heterogeneous data sources to achieve better prediction power.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China Projects No.91546105, No.U1636207, and the Shanghai Science and Technology Development Fund No.16JC1400801, No.17511105502, No.17511101702.

REFERENCES

[1] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature genetics*, vol. 33, no. 3s, p. 228, 2003.

[2] J. Piero, lex Bravo, N. Queraltrosinach *et al.*, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, no. Database issue, pp. D833–D839, 2017.

[3] X. Wu, R. Jiang, M. Q. Zhang *et al.*, "Network-based global inference of human disease genes," *Molecular systems biology*, vol. 4, no. 1, p. 189, 2008.

[4] O. Vanunu, O. Magger, E. Ruppin *et al.*, "Associating genes and protein complexes with disease via network propagation," *PLoS computational biology*, vol. 6, no. 1, p. e1000641, 2010.

[5] U. M. Singh-Blom, N. Natarajan, A. Tewari *et al.*, "Prediction and validation of gene-disease associations using methods inspired by social network analyses," *PloS one*, vol. 8, no. 5, p. e58977, 2013.

[6] X. Zeng, Y. Liao, Y. Liu *et al.*, "Prediction and validation of disease genes using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 3, pp. 687–695, 2017.

[7] Y. Fan, K. Siklenka, S. K. Arora *et al.*, "miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis," *Nucleic Acids Research*, vol. 44, no. W1, pp. W135–W141, 2016.

TABLE VI

TOP 10 DISEASES WITH UNKNOWN ASSOCIATIONS PREDICTED BY HETEWALK.

hsa-mir-21			hsa-let-7a-1			hsa-mir-125b-1		
Rank	*Disease*	*Verified*	*Rank*	*Disease*	*Verified*	*Rank*	*Disease*	*Verified*
3	188550 Nonmedullary Thyroid cancer 1	miR2Disease	2	155255 Medulloblastoma	PhenomiR	1	137800 Glioma susceptibility 1	miR2Disease
5	608232 Chronic myeloid leukemia	PhenomiR	4	176807 Prostate cancer	PhenomiR, HMDD, miR2Disease	2	266600 Inflammatory bowel disease 1	
6	266600 Inflammatory bowel disease 1	HMDD	6	256700 Neuroblastoma	PhenomiR	4	188550 Nonmedullary Thyroid cancer 1	HMDD
8	607464 Thyroid carcinoma		7	608232 Chronic myeloid leukemia	PhenomiR	5	273300 Male germ cell tumor	
9	273300 Male germ cell tumor		9	151430 B-cell lymphoma 2	PhenomiR	6	608232 Chronic myeloid leukemia	PhenomiR
10	151430 B-cell lymphoma 2	PhenomiR	10	150699 Uterine leiomyoma		7	155601 Cutaneous malignant melanoma	HMDD
11	155601 Cutaneous malignant melanoma	PhenomiR	12	600634 Pituitary adenoma	miR2Disease	9	145500 Hypertension	
12	145500 Hypertension	HMDD	15	236000 Hodgkin lymphoma	PhenomiR, HMDD, miR2Disease	10	181500 Schizophrenia	
13	256700 Neuroblastoma	HMDD	16	607464 Thyroid carcinoma		11	151430 B-cell lymphoma 2	PhenomiR
14	176807 Prostate cancer	PhenomiR, HMDD, miR2Disease	18	226150 Enterocolitis		13	260350 Pancreatic cancer	PhenomiR, HMDD, miR2Disease

TABLE VII

TOP 10 DISEASES PREDICTED BY CATAPULT TO HAVE ASSOCIATION WITH A GIVEN MIRNA.

hsa-mir-21		hsa-let-7a-1		hsa-mir-125b-1	
4	**151430 B-cell lymphoma 2**	7	**151430 B-cell lymphoma 2**	3	**260350 Pancreatic cancer**
7	273300 Male germ cell tumor	9	**608232 Chronic myeloid leukemia**	4	**137800 Glioma susceptibility 1**
9	155601 Cutaneous malignant melanooma	10	273300 Male germ cell tumor	6	273300 Male germ cell tumor
11	**266600 Inflammatory bowel disease 1**	13	188550 Nonmedullary Thyroid cancer 1	7	**151430 B-cell lymphoma 2**
13	**608232 Chronic myeloid leukemia**	14	137800 Glioma susceptibility 1	9	155601 Cutaneous malignant melanooma
14	**188550 Nonmedullary Thyroid cancer 1**	15	226150 Enterocolitis	10	114500 Colorectal cancer
15	226150 Enterocolitis	17	**600634 Pituitary adenoma**	11	226150 Enterocolitis
16	181500 Schizophrenia	19	605027 Non-Hodgkin Lymphoma	12	236000 Hodgkin lymphoma
17	131440 Myeloproliferative disorder with eosinophilia	20	266600 Inflammatory bowel disease 1	13	**188550 Nonmedullary Thyroid cancer 1**
18	605027 Non-Hodgkin Lymphoma	21	268210 Rhabdomyosarcoma	14	266600 Inflammatory bowel disease 1

TABLE VIII

TOP 10 DISEASES PREDICTED BY HSMP TO HAVE ASSOCIATION WITH A GIVEN MIRNA.

hsa-mir-21		hsa-let-7a-1		hsa-mir-125b-1	
3	155601 Cutaneous malignant melanooma	5	**608232 Chronic myeloid leukemia**	3	266600 Inflammatory bowel disease 1
4	**608232 Chronic myeloid leukemia**	8	**151430 B-cell lymphoma 2**	5	**137800 Glioma susceptibility 1**
5	**151430 B-cell lymphoma 2**	9	**600634 Pituitary adenoma**	6	273300 Male germ cell tumor
6	151400 Leukemia	11	181500 Schizophrenia	7	**188550 Nonmedullary Thyroid cancer 1**
8	**188550 Nonmedullary Thyroid cancer 1**	12	131440 Myeloproliferative disorder with eosinophilia	9	**260350 Pancreatic cancer**
9	**145500 Hypertension**	14	**155255 Medulloblastoma**	10	181500 Schizophrenia
11	137580 Tourette syndrome	16	236000 Hodgkin lymphoma	11	**151430 B-cell lymphoma 2**
14	273300 Male germ cell tumor	17	**176807 Prostate cancer**	12	**608232 Chronic myeloid leukemia**
15	**256700 Neuroblastoma**	18	268210 Rhabdomyosarcoma	13	158350 Cowden syndrome 1
16	131440 Myeloproliferative disorder with eosinophilia	19	192600 Cardiomyopathy	14	600634 Pituitary adenoma

TABLE IX

TOP 10 DISEASES PREDICTED BY HSSVM TO HAVE ASSOCIATION WITH A GIVEN MIRNA.

hsa-mir-21		hsa-let-7a-1		hsa-mir-125b-1	
3	**608232 Chronic myeloid leukemia**	6	**600634 Pituitary adenoma**	4	114500 Colorectal cancer
4	155601 Cutaneous malignant melanooma	8	**608232 Chronic myeloid leukemia**	5	266600 Inflammatory bowel disease 1
5	**145500 Hypertension**	9	**155255 Medulloblastoma**	6	145500 Hypertension
7	**151430 B-cell lymphoma 2**	11	131440 Myeloproliferative disorder with eosinophilia	7	601626 Acute myeloid leukemia
8	**266600 Inflammatory bowel disease 1**	13	608232 Chronic myeloid leukemia	9	226150 Enterocolitis
10	**188550 Nonmedullary Thyroid cancer 1**	14	268210 Rhabdomyosarcoma	10	**137800 Glioma susceptibility 1**
12	601665 Obesity	15	**151430 B-cell lymphoma 2**	11	268210 Rhabdomyosarcoma
13	273300 Male germ cell tumor	16	150699 Uterine leiomyoma	12	273300 Male germ cell tumor
14	607464 Thyroid carcinoma	18	**176807 Prostate cancer**	13	600634 Pituitary adenoma
15	247640 Lymphoblastic leukemia	19	**256700 Neuroblastoma**	14	266600 Inflammatory bowel disease 1

[8] H. Chen and Z. Zhang, "Similarity-based methods for potential human microRNA-disease association prediction," *BMC Medical Genomics*, vol. 6, no. 1, p. 12, 2013.

[9] X. Zeng, X. Zhang, Y. Liao *et al.*, "Prediction and validation of association between microRNAs and diseases by multipath methods." *Biochimica Et Biophysica Acta*, vol. 1860, no. 11, pp. 2735–2739, 2016.

[10] T. Keshava Prasad, R. Goel, K. Kandasamy *et al.*, "Human protein reference database-2009 update," *Nucleic Acids Research*, vol. 37, no. suppl_1, pp. D767–D772, 2008.

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

[12] M. A. Van Driel, J. Bruggeman, G. Vriend *et al.*, "A text-mining analysis of the human phenome," *European Journal of Human Genetics: EJHG*, vol. 14, no. 5, p. 535, 2006.

[13] Q. Zou, J. Li, C. Wang *et al.*, "Approaches for recognizing disease genes based on network." *Biomed Research International*, vol. 2014, no. 5013, p. 416323, 2014.

[14] Q. Zou, J. Li, L. Song *et al.*, "Similarity computation strategies in the microrna-disease network: a survey." *Briefings in Functional Genomics*, vol. 15, no. 1, p. 55, 2016.

[15] D. A. Peter, M. C. Grondin, J. Robin *et al.*, "The comparative toxicogenomics database: update 2013," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. 1067–72, 2011.

[16] W. Wang, S. Yang, X. Zhang *et al.*, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics*, vol. 30, no. 20, pp. 2923–30, 2014.

[17] M. Campillos, M. Kuhn, A.-C. Gavin *et al.*, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

[18] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, 2014.

[19] J. Tang, M. Qu, M. Wang *et al.*, "LINE: Large-scale information network embedding," *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067–1077, 2015.

[20] C.-H. Chou, N.-W. Chang, S. Shrestha *et al.*, "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database," *Nucleic Acids Research*, vol. 44, no. D1, pp. D239–D247, 2015.

[21] V. McKusick, "Mendelian inheritance in man: a catalog of human genes and genetic disorders," 1998.

[22] Q. Jiang, Y. Wang, Y. Hao *et al.*, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. suppl_1, pp. D98–D104, 2008.

[23] M. Lu, Q. Zhang, M. Deng *et al.*, "An analysis of human microRNA and disease associations," *PloS one*, vol. 3, no. 10, p. e3420, 2008.

[24] A. Ruepp, A. Kowarsch, D. Schmidl *et al.*, "PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes," *Genome Biology*, vol. 11, no. 1, p. R6, 2010.

[25] T. Mikolov, I. Sutskever, K. Chen *et al.*, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[26] Y. Sun, J. Han, X. Yan *et al.*, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *VLDB*, vol. 4, no. 11, pp. 992–1003, 2011.

[27] T. Mikolov, K. Chen, G. Corrado *et al.*, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] L. Bottou, "Large-scale machine learning with stochastic gradient descent," *Proceedings of 19th International Conference on Computational Statistics*, pp. 177–186, 2010.

[29] C. Shi, X. Kong, Y. Huang *et al.*, "HeteSim: A general framework for relevance measure in heterogeneous networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2479–2492, 2014.