Biomedical association inference on pandemic knowledge graphs: A comparative study*

Mengjia Wu[1,*], Chao Yu[2], Jian Xu[2], Ying Ding[3] and Yi Zhang[1]

[1]Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, 15 Broadway, Ultimo, NSW, Australia

[2]School of Information Management, Sun Yat-sen University, Guangzhou, China

[3]School of Information, University of Texas, Austin, TX, USA

Abstract

Acquiring insights and understanding from historical pandemics is crucial for reducing the likelihood of their recurrence. The utilization of knowledge graphs stands as an essential tool for researchers, with knowledge inference emerging as a prominent task within these graphs to deduce previously unidentified connections between entities. This study endeavors to construct a knowledge graph centered on pandemic research and to evaluate the efficacy of various mainstream methodologies in the context of biomedical association inference. Our findings indicate that techniques for graph representation hold significant promise in executing these tasks and heterogeneous graph representation techniques demonstrate high predicting accuracy. Nonetheless, the advancement in this area of research necessitates more refined experimental designs and the adoption of more adaptive learning strategies.

Keywords

Biomedical knowledge graph, graph representation, knowledge inference

1. Introduction

Biomedical entity association inference is a long-term task for scientific researchers and industry practitioners to understand the relationships between biomedical entities and propose first-hand literature-based evidence for further investigations [1, 2]. Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS) and Coronavirus Disease 2019 (COVID-19), the three notorious pandemics in public health history, presented huge threats to human lives and social stability [3, 4]. Uncovering knowledge inference from the pandemic knowledge foundation encompassing tremendous coronavirus-related research articles published in human history may bring insights to uncover the evolutionary mechanisms of coronavirus for reducing public uncertainties towards and developing precautions for future infectious disease crises [5, 6]. However, the complexity, heterogeneity and intricate associations of biomedical entities present a challenge in exploring newly emerging knowledge.

Knowledge graphs, which are extensively used to depict intricate data relationships, serve as the foundation for analyzing and inferring associations [2, 6, 7, 8]. These graphs represent biomedical entities such as genes, diseases, chemicals, and drugs as nodes, with their relationships illustrated as either directed or undirected edges, sometimes accompanied by supplementary descriptive attributes. Leveraging network analysis techniques, various methods have been introduced to investigate patterns of association and predict previously unknown relationships.

In this study, we developed a knowledge graph from scholarly articles on SARS, MERS, and COVID-19, comprising 9,142 nodes and 81,707 connections. We conducted a validation test to assess how well various mainstream techniques could predict relationships within this graph. By masking 10% of the connections of each type, we applied five different methods to the masked graph to identify the hidden connections from an equal mix of randomly inserted non-existent connections. The findings revealed the diverse effectiveness of these methods in identifying the obscured connections, with HetGNN proven as the most effective. Nonetheless, the flexibility and applicability of different graph representation methods across varied contexts need enhancement. This research illustrates the application of multiple prominent methods in deducing associations in knowledge graphs and verifies the precision of these methods.

The following of this paper is organized as follows: We introduced the pandemic knowledge graphs and examined methods in the section Data and Method, followed by Experimental Settings and Results. We concluded the study and anticipated some future directions in the section of Discussion and Conclusions.

2. Data and Method

The integrative Biomedical Knowledge Hub (iBKH) is a knowledge graph dataset that curates the associations of 11 categories of biomedical entities from 17 publicly available data sources [9]. Using the iBKH as the global dataset, we searched scholarly articles across PubMed using search strategies from [3] and cross-matched the search results to iBKH. By extracting the nodes and edges relevant to papers in the search results, we constructed a pandemic-specific sub-graph of the iBKH dataset. The overall description of sub-graphs relevant to each pandemic is given in Table 1.

The pandemic graph is denoted as $G = (V, E)$, and

$$V = \{V_{dis}, V_{dg}, V_g\} \tag{1}$$

$$E = \{E^{dg}_{dg}, E^{dis}_{dg}, E^{g}_{dg}, E^{dis}_{dis}, E^{g}_{dis}, E^{g}_{g}\} \tag{2}$$

where V_{dis}, V_{dg}, and V_g respectively represent the node set of diseases, drugs and genes. $E^i_j(i, j \in \{dis, dg, g\})$ denotes the edge set of associations between nodes of types i and j. Entity association inference on this pandemic graph aims to predict emerging associations between nodes in V that have not yet appeared in E.

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024)

*Corresponding author.

mengjia.wu@uts.edu.au (M. Wu); yuch25@mail3.sysu.edu.cn (C. Yu); issxj@mail.sysu.edu.cn (J. Xu); ying.ding@ischool.utexas.edu (Y. Ding); yi.zhang@uts.edu.au (Y. Zhang)

0000-0003-3956-7808 (M. Wu); 0000-0003-4886-4708 (J. Xu); 0000-0003-2567-2009 (Y. Ding); 0000-0002-7731-0301 (Y. Zhang)

Table 1
The basic information of pandemic knowledge graphs

	SARS	MERS	COVID-19	Pandemic graph
#Paper	9,991	1,494	281,569	293,054
Drug	439	46	1,429	1,507
Disease	522	94	1,841	1,814
Gene	1,939	345	5,435	5,821
Drug-drug	145	13	1,626	1,678
Drug-disease	951	59	9,085	9,381
Drug-Gene	710	49	3,709	4,135
Disease-disease	148	26	928	939
Disease-gene	6,256	575	54,503	57,236
Gene-gene	2,199	347	7,461	8,338

There have been substantial efforts in the development of association inference methodologies. In this study, we selected the following representative methods to experiment:

- Random Walk with Restart (RWR) is a commonly used method for inferring relationships within graphs, particularly in the biomedical field. It models a random walking process that begins at node a and calculates the likelihood of reaching node b as a measure of relevance between nodes a and b. To avoid the walk from becoming trapped in local areas, it introduces a restart probability p, which allows the walk to restart from node a at each step, thereby ensuring broader exploration of the graph.
- Resource allocation (RA) [10]: RA is a link prediction algorithm that conceptualizes the graph as a transportation network, viewing edges as channels for resource diffusion. Under this model, the likelihood of forming a link between any two nodes is approximated by the total resources these nodes are expected to receive through their shared neighbors. This approach leverages the idea that the more resources two nodes can exchange via their common connections, the higher the probability they will establish a direct link.
- Node2Vec [11]: Node2Vec is a scalable graph representation technique that utilizes random walks to learn low-dimensional vector representations of nodes within a graph. It operates by optimizing an objective that aims to preserve neighborhood relationships, ensuring that nodes with similar network neighborhoods are close to each other in the vector space.
- Heterogeneous graph neural networks (HetGNN) [12]: HetGNN is a graph representation technique designed to work with heterogeneous graphs, characterized by their inclusion of various types of nodes, each possessing diverse content attributes such as text and images. It introduces a novel two-step information aggregation process aimed at effectively learning from the information presented by neighboring nodes, both of the same and different types. This process allows HetGNN to capture the complex structural and content heterogeneity of the graph, enabling the model to generate more accurate and meaningful representations of each node.
- Heterogeneous graph neural network with co-contrastive learning (HeCo) [13]: HeCo is a self-supervised learning technique designed for heterogeneous graph representation, which utilizes contrastive learning to derive node representations.

3. Experiment settings

The setup for the experiment is detailed in Figure 1. The objective was to assess the efficacy of various algorithms in predicting associations between biomedical entities. To this end, a validation experiment was structured in the following manner: From each category of edges, denoted as E^i_j where i, j belong to the set dis, dg, g (representing disease, drug-gene, and gene respectively), 10% of the edges were randomly selected and removed. The resulting graph, with these edges removed, was labeled as $G_m = (V, E_m)$. The edges that were removed are represented by $rE = rEj^i | i, j \in dis, dg, g$, and these were considered the 'true' associations for the purposes of this experiment. In addition to this, an equivalent number of node pairs, which were not connected by edges in the original graph G, were randomly chosen. These pairs are denoted by $nE = nEj^i | i, j \in dis, dg, g, nEj^i \cap Ej^i = \emptyset$, and they were defined as the negative sample set for this study. This methodical approach enabled a balanced evaluation, comparing the algorithms' abilities to correctly infer both existing and non-existing associations, thereby providing a comprehensive understanding of their performance in the context of biomedical entity association inference.

Subsequently, each candidate algorithm was applied to the modified graph G_m to ascertain the likelihood of edge formation between every pair of nodes within both rE and nE. In the cases of the Random Walk with Restart (RWR) and resource allocation algorithms, this procedure involved computing the random walk probability and the resource allocation score, respectively, for each node pair. Conversely, for the three graph representation techniques, the process entailed converting every node in the set V into embedding vectors. The representation for edges was then determined through an average pooling strategy, which involves aggregating the features of node embeddings to form a single representation for each edge.

Following the generation of these probabilities or representations, the combined dataset of rE and nE was divided, with 80% allocated for training and the remaining 20% for testing. This division was employed to train a logistic regression classifier, the purpose of which was to predict the likelihood of edge formation between node pairs in the test set. The predictions made by the logistic regression model were then used to calculate the Area Under the Curve (AUC) metric for each method. By focusing exclusively on the test

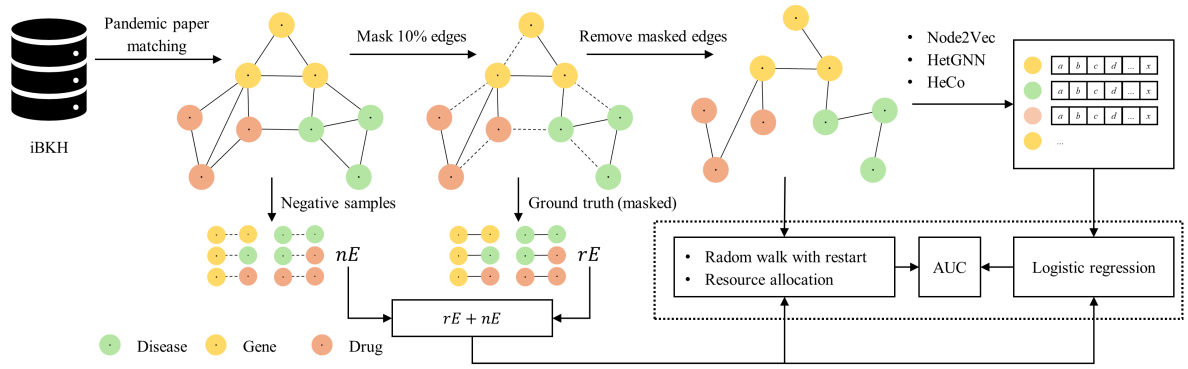

Figure 1: The overall experiment design

Table 2
Performance comparison of selected algorithms

Method	RWR	RA	Node2Vec	HeCo	**HetGNN**
E_{dg}^{dg}	0.5827	0.5830	0.7257	-	**0.9566**
E_{dg}^{dis}	0.7081	0.7651	0.8079	-	**0.8315**
E_{dg}^{g}	0.8298	0.8741	0.9250	0.9120	**0.9584**
E_{dis}^{dis}	0.7585	0.7893	0.7086	-	**0.8495**
E_{dis}^{g}	0.5327	0.5410	0.7802	0.7990	**0.8001**
E_{g}^{g}	0.7561	0.8110	0.8327	0.8530	**0.9050**

data, which comprised 20% of the total dataset, a standardized evaluation criterion was established. This approach allowed for a fair comparison of the five candidate methods, with the AUC metric serving as a measure of each method's ability to accurately classify node pairs as either connected or not connected, based on the generated classification probabilities.

4. Results

Table 2 presents the AUC scores for the five candidate methods. It is noted that HeCo needs a metapath definition to function, and a gene-based metapath was chosen for this purpose. Consequently, HeCo's evaluation was limited to gene-related associations. It was found that HetGNN outperformed others in recovering the removed links.Compared to RWR and RA, the three graph representation methods demonstrated better accuracy in identifying connections. Yet, their advantage is not definitive because they utilize a supervised learning approach, requiring both positive and negative samples to train a classifier, whereas RWR and RA can be applied directly to the existing graph structure without any pre-existing knowledge of it.

From the perspective of edge types, the analysis of gene-drug and drug-drug connections showed superior outcomes. Importantly, both RWR and RA displayed similar levels of effectiveness as graph representation techniques in the task of deducing disease-disease associations. This suggests that inferring disease similarities might be distinct from other tasks, meriting additional investigation.

Among the graph representation strategies, two methods tailored for heterogeneous networks achieved superior AUC scores over Node2Vec. This superiority results from their training mechanisms being specifically designed for heterogeneous networks, as seen in this research and commonly in biomedical entity graphs. These methods incorporate the significance of node types into the computation, employing either type-specific or metapath-based aggregation strategies for information. While this heterogeneity-focused approach is beneficial, it limits the model's applicability and increases the cost of adaptation. Changes in the heterogeneous graph's structure necessitate adjustments to HetGNN's data inputs and HeCo's metapaths, along with significant methodological revisions. Additionally, HeCo's performance is influenced by the setting of a positive sample threshold and the definition of metapaths, which vary per case and affects the outcome significantly. Node2Vec, in contrast, offers a more generalized solution applicable to a wide range of graph types.

In conclusion, while heterogeneous graph representation methods hold promise for deducing relationships within pandemic knowledge graphs, enhancing their flexibility and general applicability remains a challenge.

5. Discussion and Conclusions

This study explores the performance of different methods of association inference and provides insights into the potential of graph representation methods. Despite some existing entity-relationship summarization tools like PubTator 3 [14], graph representation methods still hold the potential to infer more accurate biomedical associations but need improvement on adaptability and generalisability. Future work will modify the inference framework and perform real-world association inference on the built pandemic graph.

We anticipated the following future directions aligning with some limitations of the current study: 1) This study offered some preliminary understandings on selected baselines of graph representation learning in inferring the pandemic knowledge graph, but further customized redevelopment based on the unique features of the pandemic knowledge graph to enhance its performance might be beneficial. 2) Investigating the scientific community of a pandemic and its collaborative patterns will bring insights to analyze the societal context of a pandemic crisis and provide evidence-based decision support in terms of science policy, public health, and public administration.

Acknowledgments

This work was supported by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, in conjunction with the National Science Foundation (NSF) of the United States, under CSIRO-NSF #2303037.

References

[1] S. Henry, B. T. McInnes, Literature based discovery: Models, methods, and trends, Journal of Biomedical Informatics 74 (2017) 20–32.

[2] M. Wu, Y. Zhang, G. Zhang, J. Lu, Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study, Technological Forecasting and Social Change 164 (2021) 120513.

[3] M. Haghani, M. C. Bliemer, Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across sars, mers and 2019-ncov literature, Scientometrics 125 (2020) 2695–2726.

[4] Y. Zhang, X. Cai, C. V. Fry, M. Wu, C. S. Wagner, Topic evolution, disruption and resilience in early covid-19 research, Scientometrics 126 (2021) 4225–4253.

[5] A. L. Porter, Y. Zhang, Y. Huang, M. Wu, Tracking and mining the covid-19 research literature, Frontiers in Research Metrics and Analytics 5 (2020) 594060.

[6] M. Wu, Y. Zhang, M. Markley, C. Cassidy, N. Newman, A. Porter, Covid-19 knowledge deconstruction and retrieval: An intelligent bibliometric solution, Scientometrics (2023) 1–31.

[7] M. Wu, Y. Zhang, M. Grosser, S. Tipper, D. Venter, H. Lin, J. Lu, Profiling covid-19 genetic research: A data-driven study utilizing intelligent bibliometrics, Frontiers in Research Metrics and Analytics 6 (2021) 683212.

[8] K. Guo, M. Wu, Z. Soo, Y. Yang, Y. Zhang, Q. Zhang, H. Lin, M. Grosser, D. Venter, G. Zhang, et al., Artificial intelligence-driven biomedical genomics, Knowledge-Based Systems (2023) 110937.

[9] C. Su, Y. Hou, M. Zhou, S. Rajendran, J. R. Maasch, Z. Abedi, H. Zhang, Z. Bai, A. Cuturrufo, W. Guo, et al., Biomedical discovery through the integrative biomedical knowledge hub (ibkh), Iscience 26 (2023).

[10] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, The European Physical Journal B 71 (2009) 623–630.

[11] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.

[12] C. Zhang, D. Song, C. Huang, A. Swami, N. V. Chawla, Heterogeneous graph neural network, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 793–803.

[13] X. Wang, N. Liu, H. Han, C. Shi, Self-supervised heterogeneous graph neural network with co-contrastive learning, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1726–1736.

[14] C.-H. Wei, A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, Z. Lu, Pubtator 3.0: An ai-powered literature resource for unlocking biomedical knowledge, arXiv preprint arXiv:2401.11048 (2024).