




## Data and text mining

# Metapaths: similarity search in heterogeneous knowledge graphs via meta-paths

Ayush Noori <sup>1,2</sup>, Michelle M. Li <sup>2,\*</sup>, Amelia L. M. Tan<sup>2,\*</sup>, Marinka Zitnik <sup>2,\*</sup>

<sup>1</sup>Harvard College, Cambridge, MA 02138, United States

<sup>2</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States

\*Corresponding author. Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. E-mail: marinka@hms.harvard.edu (M.Z.)

<sup>†</sup>These authors contributed equally to this work.

Associate Editor: Zhiyong Lu

### Abstract

**Summary:** Heterogeneous knowledge graphs (KGs) have enabled the modeling of complex systems, from genetic interaction graphs and protein-protein interaction networks to networks representing drugs, diseases, proteins, and side effects. Analytical methods for KGs rely on quantifying similarities between entities, such as nodes, in the graph. However, such methods must consider the diversity of node and edge types contained within the KG via, for example, defined sequences of entity types known as meta-paths. We present *metapaths*, the first R software package to implement meta-paths and perform meta-path-based similarity search in heterogeneous KGs. The *metapaths* package offers various built-in similarity metrics for node pair comparison by querying KGs represented as either edge or adjacency lists, as well as auxiliary aggregation methods to measure set-level relationships. Indeed, evaluation of these methods on an open-source biomedical KG recovered meaningful drug and disease-associated relationships, including those in Alzheimer's disease. The *metapaths* framework facilitates the scalable and flexible modeling of network similarities in KGs with applications across KG learning.

**Availability and implementation:** The *metapaths* R package is available via GitHub at <https://github.com/ayushnoori/metapaths> and is released under MPL 2.0 (Zenodo DOI: [10.5281/zenodo.7047209](https://doi.org/10.5281/zenodo.7047209)). Package documentation and usage examples are available at <https://www.ayushnoori.com/metapaths>.

## 1 Introduction

Relational data across biological systems—such as the cellular interactome, single cell similarity graphs, gene co-expression networks, and patient interaction networks—can be represented by graph architectures. A simple graph  $G = (\mathcal{V}, \mathcal{E})$  is defined by a set of nodes  $\mathcal{V}$  and edges  $\mathcal{E}$  of a single type. However, real-world networks are often comprised of diverse data modalities; thus, they are poorly modeled by homogeneously typed networks. For instance, a homogeneous network is insufficient for modeling the complexities of drug mechanisms and indications. A graph with many types of nodes—such as drugs, diseases, and proteins—that are connected by different relation types—such as “is indicated for,” “is therapeutic target of,” or “physically interacts with”—is necessary. Interconnected objects from various data sources that are represented as a single multigraph with heterogeneous knowledge-informed node and edge types are known as *knowledge graphs* (KGs) (Hogan et al. 2022). Formally, if  $\mathcal{A}$  is the set of node types with mapping function  $\varphi: \mathcal{V} \rightarrow \mathcal{A}$ , the edges  $\mathcal{E}$  of a KG can be represented as a set of tuples  $(u, v)$ , where nodes  $u, v \in \mathcal{V}$  are connected by an edge, and each belongs to a specific node type  $\varphi(u)$  and  $\varphi(v) \in \mathcal{A}$ .

Relationships between entities in KGs are modeled by network similarities quantified using sequences of nodes—or linkage paths—in the network. However, meaningful similarity search methods on KGs must account for the diverse types

in these walks. For example, consider a KG of the biological interactome with the following node types:  $\mathcal{A} = \{\text{disease } (D), \text{drug } (R), \text{protein } (P), \text{protein function } (F), \text{side effect } (S)\}$ . Classic random walk-based similarity metrics would not differentiate between the following paths of length three: *RD**P* (i.e. drug, disease, protein) and *R**S**P* (i.e. drug, side effect, protein), even though the former considers disease-mediated associations while the latter considers mechanisms of side effects (Fig. 1).

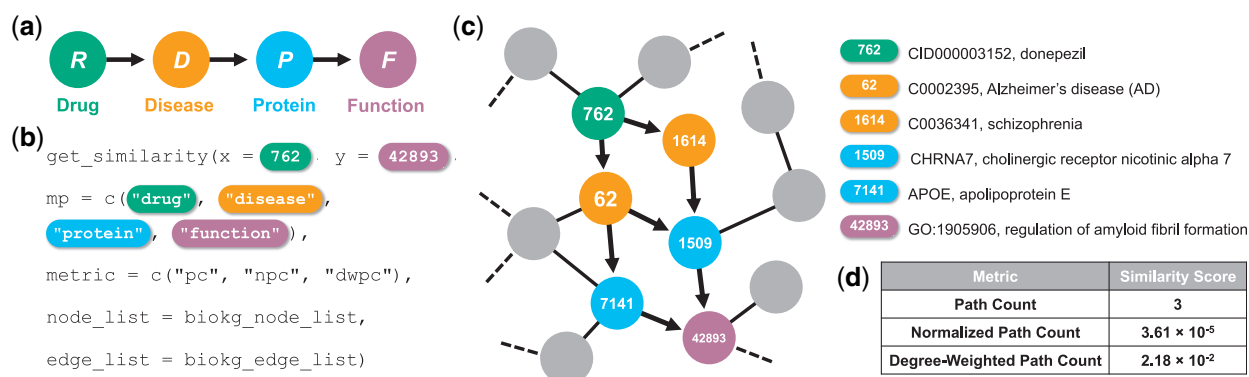
To distinguish between such paths of identical lengths, we leverage meta-paths, a general graph-theoretic approach for flexible similarity search in large networks. Meta-paths are sequences of node types which define a walk from an origin node to a destination node (Sun et al. 2011b). Note that edge types may also be specified in the walk; here, we do not consider such meta-paths. Fundamentally, the similarity between an origin node and a destination node is measured by the number of meta-paths that exist between them. While meta-paths are frequently used in biomedical network analysis (e.g. Fu et al. 2016; Himmelstein et al. 2017; Zhang et al. 2020), there is currently no package available in R that offers a wide range of support for meta-paths.

Informative meta-paths in KGs are often engineered by hand based on domain knowledge or expertise (e.g. the meta-path *DRS* is clinically meaningful, since it describes associations between a disease and the side effects of its

Received: October 8, 2022. Revised: March 10, 2023. Accepted: April 29, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1.** Evaluation of the metapaths package for similarity search in the ogbl-biokg biomedical KG. (a) We query using the *RDPF* (i.e. drug–disease–protein–function) meta-path. (b) The function call used to calculate meta-path-based similarity scores is shown. (c) The meta-path traversal function identifies three paths following the specified meta-path that connect donepezil—a drug used to treat Alzheimer’s disease (AD)—with the regulation of amyloid fibril formation pathway, which is implicated in AD (Supplementary data). (d) The computed similarity scores using Path Count, Normalized Path Count, and Degree-Weighted Path Count metrics are shown

treatments, whereas the meta-path *PSF* would not be). Alternatively, optimal meta-paths can be discovered in an unsupervised fashion by feature selection metrics [e.g. maximal spanning tree (Wang et al. 2018), Laplacian score, or ranking based on meta-path frequency or uniqueness (Zhu et al. 2018)] after enumerating meta-paths from a sampled subset of nodes  $\mathcal{V}_M \subset \mathcal{V}$ , multi-hop reasoning with reinforcement learning (Wan et al. 2020), or graph attention applied on vector embedded meta-path instance sets (Li et al. 2021), among other approaches.

Once informative meta-paths for a given KG have been identified, these meta-paths define the semantics of the relationships between nodes in the KG, thereby enabling heterogeneous graph convolutional (Zhang et al. 2019) and graph attention (Wang et al. 2019) networks for downstream machine learning analyses such as link prediction (Himmelstein et al. 2017), node classification (Wang et al. 2021), and subgraph prediction (Alsentzer et al. 2020); additional downstream applications are discussed in the Supplementary data. Although various algorithms exist to model meta-path-based node similarities in a KG, a unifying framework is lacking to compute and compare these similarity scores. Here, we introduce metapaths which, to the best of our knowledge, is the first software package in the R ecosystem to implement meta-paths. The metapaths package enables the computation of meta-path-based similarity search in heterogeneous KGs.

## 2 Implementation and evaluation

The primitives of the metapaths package identify the neighbors of a specified node with a given type by querying either an edge list or, for efficiency, an adjacency list precomputed from an edge list. The meta-path traversal function accepts an origin node, a destination node, and a specified meta-path; then, via the neighbor identification functions, it starts at the origin node and recursively expounds the sequence of node types until the destination node is reached. The resulting paths are used to compute meta-path-based similarity scores using various available similarity metrics, including the natively supported Path Count, Normalized Path Count, Degree-Weighted Path Count, and PathSim (Supplementary data) (Sun et al. 2011a,b; Himmelstein and Baranzini 2015). Users may also use the framework provided by the metapaths package to define and test custom similarity metrics of their

choosing or evaluate the similarity between two sets of nodes via auxiliary aggregation functions.

To validate the metapaths package, we demonstrate that the package similarity functions report higher connectivity in the ogbl-biokg (BioKG) biomedical KG (Hu et al. 2020) between Alzheimer’s disease (AD)-related drugs (e.g. donepezil, memantine, and galantamine) and AD-related pathways (Supplementary Table S1); this increased connectivity also holds at both the node pair and node set level (Fig. 1 and Supplementary data). Finally, by testing metapaths functions on randomly sampled BioKG subgraphs of increasing size, we demonstrate that the performance of the metapaths package scales well with input size (Supplementary Fig. S1).

## 3 Conclusion

The metapaths R software package facilitates the scalable and flexible modeling of network similarities in KGs. Relationships between individual nodes in a KG can be quantified using built-in or user-defined similarity metrics; such metrics can also be applied to model set-level relationships via aggregation methods. Evaluation on AD-related pathways in BioKG recovers meaningful drug and disease-associated relationships as quantified by high similarity scores. The applications of such similarity search in KGs extend across KG learning.

## Supplementary data

Supplementary data is available at *Bioinformatics* online.

## Code and data availability

The metapaths R package is available via GitHub at <https://github.com/ayushnoori/metapaths> and is released under MPL 2.0 (Zenodo DOI: 10.5281/zenodo.7047209). Package documentation and usage examples are available at <https://www.ayushnoori.com/metapaths>. The ogbl-biokg and ogbn-arxiv datasets are publicly available from the Open Graph Benchmark at <https://ogb.stanford.edu>.

## Conflict of Interest

None declared.

## Funding

A.N. gratefully acknowledges the support of the SPARK Fellowship from the Center for Public Service and Engaged Scholarship at Harvard College. M.M.L. and M.Z. gratefully acknowledge the support of National Institutes of Health No. R01HD108794, Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001, and awards from Harvard Data Science Initiative, Amazon Research, Google Research Scholar Program, Bayer Early Excellence in Science, AstraZeneca Research, and Roche Alliance with Distinguished Scientists. M.M.L. is supported by T32HG002295 from the National Human Genome Research Institute and a National Science Foundation Graduate Research Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

## References

- Alsentzer E, Finlayson S, Li M *et al.* Subgraph neural networks. In: Larochelle H, Ranzato M, Hadsell R. *et al.* (eds.) *Advances in Neural Information Processing Systems*. New York: Curran Associates, Inc., 2020, 8017–29.
- Fu G, Ding Y, Seal A *et al.* Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinformatics* 2016;17:160.
- Himmelstein DS, Baranzini SE. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS Comput Biol* 2015;11:e1004259.
- Himmelstein DS, Lizee A, Hessler C *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 2017;6:e26726.
- Hogan A, Blomqvist E, Cochez M *et al.* Knowledge graphs. *ACM Comput Surv* 2022;54:1–37.
- Hu W, Fey M, Zitnik M *et al.* Open graph benchmark: datasets for machine learning on graphs. In: Larochelle H, Ranzato M, Hadsell R. *et al.* (eds.) *Advances in Neural Information Processing Systems*. New York: Curran Associates, Inc., 2020, 22118–33.
- Li Y, Jin Y, Song G *et al.* GraphMSE: efficient meta-path selection in semantically aligned feature space for graph neural networks. *AAAI* 2021;35:4206–14.
- Sun Y, Barber R, Gupta M *et al.* Co-author relationship prediction in heterogeneous bibliographic networks. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011a, 121–8.
- Sun Y, Han J, Yan X *et al.* PathSim: meta path-based top-K similarity search in heterogeneous information networks. *Proc VLDB Endow* 2011b;4:992–1003.
- Wan G, Du B, Pan S *et al.* Reinforcement learning based meta-path discovery in large-scale heterogeneous information networks. *AAAI* 2020;34:6094–101.
- Wang X, Ji H, Shi C *et al.* Heterogeneous graph attention network. In: *The World Wide Web Conference, WWW '19*. New York, NY, USA: Association for Computing Machinery, 2019, 2022–32.
- Wang S, Pisco AO, McGeever A *et al.* Leveraging the cell ontology to classify unseen cell types. *Nat Commun* 2021;12:5556.
- Wang C, Song Y, Li H *et al.* Unsupervised meta-path selection for text similarity measure based on heterogeneous information networks. *Data Min Knowl Disc* 2018;32:1735–67.
- Zhang C, Song D, Huang C *et al.* Heterogeneous graph neural network. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*. New York, NY, USA: Association for Computing Machinery, 2019, 793–803.
- Zhang L, Liu B, Li Z *et al.* Predicting MiRNA-disease associations by multiple meta-paths fusion graph embedding model. *BMC Bioinformatics* 2020;21:470.
- Zhu Z, Cheng R, Do L *et al.* Evaluating top-k meta path queries on large heterogeneous information networks. In: *2018 IEEE International Conference on Data Mining (ICDM)*, 2018, 1470–5.