The Journal of Open Source Software

GraphSL: An Open-Source Library for Graph Source Localization Approaches and Benchmark Datasets

Junxiang Wang[1] and **Liang Zhao**[1]

1 Emory University, United States

DOI: 10.21105/joss.06796

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @range-et
- @mbeyss
- @rkdan

Submitted: 05 April 2024
Published: 30 July 2024

Summary

We introduce GraphSL, a new library for studying the graph source localization problem. Graph diffusion and graph source localization are inverse problems in nature: graph diffusion predicts information diffusions from information sources, while graph source localization predicts information sources from information diffusions. GraphSL facilitates the exploration of various graph diffusion models for simulating information diffusions and enables the evaluation of cutting-edge source localization approaches on established benchmark datasets. The source code of GraphSL is made available at GitHub Repository. Bug reports and feedback can be directed to the GitHub issues page.

Statement of Need

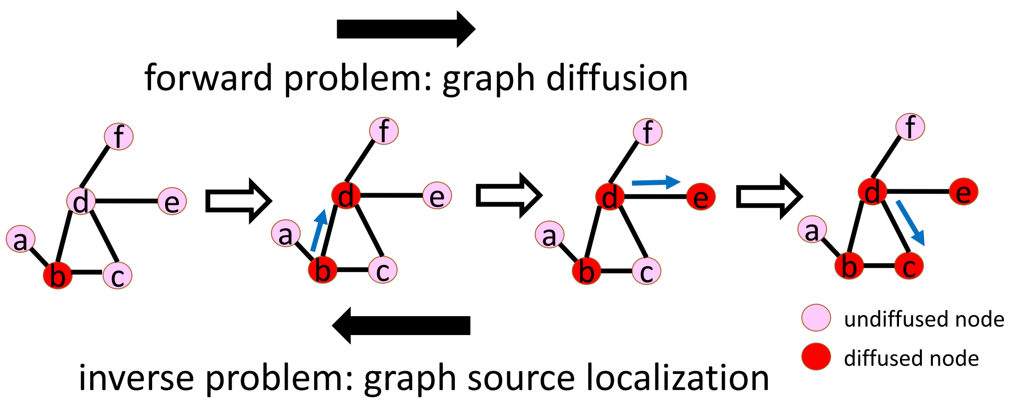

Figure 1: An example of graph source localization.

Graph diffusion is a fundamental task in graph learning, which aims to predict future information diffusions given information sources. Its inverse problem is graph source localization, which is an extremely important topic even though rarely explored: it focuses on the detection of information sources given their future information diffusions. As illustrated in Figure 1, graph diffusion seeks to predict the information diffusion $\{b, c, d, e\}$ from a source node b, whereas graph source localization aims to identify the source node b from the information diffusion $\{b, c, d, e\}$. Graph source localization spans a broad spectrum of promising research and real-world applications such as rumor detection (Gallotti et al., 2020), tracking of sources for computer viruses (Kephart & White, 1993), and failure detection in smart grids (Amin & Schewe, 2007). Please refer to the survey paper (Jiang et al., 2016) for more information.

Hence, the graph source localization problem demands attention and extensive investigations from machine learning researchers.

Due to its importance, some open-source tools have been developed to support research of the graph source localization problem. Two recent examples are cosasi (McCabe, 2022) and RPaSDT (Fraszczak, 2022). However, they do not support various simulations of information diffusion, and they also miss real-world benchmark datasets and state-of-the-art source localization approaches. To fill this gap, we propose a new library GraphSL: the first one to include real-world benchmark datasets and recent source localization methods to our knowledge, enabling researchers and practitioners to evaluate novel techniques against appropriate baselines easily. These methods do not require prior assumptions about the source (e.g. single source or multiple sources) and can handle graph source localization based on various diffusion simulation models such as Independent Cascade (IC) and Linear Threshold (LT) (Shakarian et al., 2015). Our GraphSL library is standardized: for instance, tests of all source inference methods return a Metric object, which provides five performance metrics (accuracy, precision, recall, F-score, and area under ROC curve) for performance evaluation.

Our GraphSL library targets both developers and practical users: they are free to add algorithms and datasets for personal needs by following the guidelines in the "Contact" section of README.md.

Methods and Benchmark Datasets

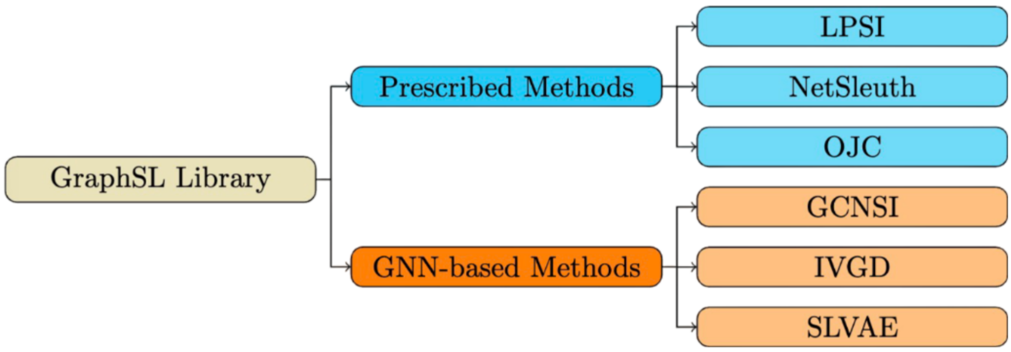

Figure 2: The hierarchical structure of the GraphSL library: in total six algorithms are implemented, which can be divided into two categories: prescribed methods that rely on hand-crafted rules and GNN-based methods which learn rules from graph data.

The structure of our GraphSL library is depicted in Figure 2. Existing methods can be categorized into two groups: Prescribed methods and Graph Neural Networks (GNN)-based methods.

Prescribed methods rely on hand-crafted rules and heuristics. For instance, LPSI assumes that nodes surrounded by larger proportions of infected nodes are more likely to be source nodes (Z. Wang et al., 2017). NetSleuth employs the Minimum Description Length principle to identify the optimal set of source nodes and virus propagation ripple (Prakash et al., 2012). OJC identifies a set of nodes (Jordan cover) that cover all observed infected nodes with the minimum radius (Zhu et al., 2017).

GNN-based methods learn rules from graph data in an end-to-end manner by capturing graph topology and neighboring information. For example, GCNSI utilizes LPSI to enhance input and then applies Graph Convolutional Networks (GCN) for source identification (Dong et al., 2019); IVGD introduces a graph residual scenario to make existing graph diffusion models invertible, and it devises a new set of validity-aware layers to project inferred sources to feasible regions (J. Wang et al., 2022). SLVAE uses forward diffusion estimation and deep generative models to approximate source distribution, leveraging prior knowledge for generalization under arbitrary diffusion patterns (Ling et al., 2022).

Table 1: Six benchmark graph datasets: their numbers of nodes and edges.

Dataset	#Node	#Edge
Karate (Lusseau et al., 2003)	34	78
Dolphins (Lusseau et al., 2003)	62	159
Jazz (Gleiser & Danon, 2003)	198	2,742
Network Science (Newman, 2006)	1,589	2,742
Cora-ML (McCallum et al., 2000)	2,810	7,981
Power Grid (Watts & Strogatz, 1998)	4,941	6,594

Aside from methods, we also release six benchmark graph datasets to facilitate the research of graph source localization, whose statistics are shown in Table 1. Information sources and diffusions can be generated by the function diffusion_generation.

Availability and Documentation

GraphSL is available under the MIT License. The library may be cloned from the GitHub repository, or can be installed by pip: pip install GraphSL. Documentation is provided via Read the Docs, including a quickstart introducing major functionality and a detailed API reference. Extensive unit testing is employed throughout the library.

References

Amin, M., & Schewe, P. F. (2007). Preventing blackouts. *Scientific American*, *296*(5), 60–67. https://doi.org/10.1049/pe:20030305

Dong, M., Zheng, B., Quoc Viet Hung, N., Su, H., & Li, G. (2019). Multiple rumor source detection with graph convolutional networks. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 569–578. https://doi.org/10.1145/3357384.3357994

Fraszczak, D. (2022). RPaSDT—rumor propagation and source detection Toolkit. *SoftwareX*, *17*, 100988. https://doi.org/10.1016/j.softx.2022.100988

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the risks of "infodemics" in response to COVID-19 epidemics. *Nature Human Behaviour*, *4*(12), 1285–1293. https://doi.org/10.1038/s41562-020-00994-6

Gleiser, P. M., & Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems*, *6*(04), 565–573. https://doi.org/10.1142/S0219525903001067

Jiang, J., Wen, S., Yu, S., Xiang, Y., & Zhou, W. (2016). Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, *19*(1), 465–481. https://doi.org/10.1109/comst.2016.2615098

Kephart, J. O., & White, S. R. (1993). Measuring and modeling computer virus prevalence. *Proceedings 1993 IEEE Computer Society Symposium on Research in Security and Privacy*, 2–15. https://doi.org/10.1109/risp.1993.287647

Ling, C., Jiang, J., Wang, J., & Liang, Z. (2022). Source localization of graph diffusion via variational autoencoders for graph inverse problems. *Proceedings of the 28th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, 1010–1020. https://doi.org/10.1145/3534678.3539288

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., & Dawson, S. M. (2003). The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, *54*, 396–405. https://doi.org/10.1007/s00265-003-0651-y

McCabe, L. H. (2022). Cosasi: Graph diffusion source inference in Python. *Journal of Open Source Software*, *7*(80), 4894. https://doi.org/10.21105/joss.04894

McCallum, A. K., Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval*, *3*, 127–163. https://doi.org/10.1023/A:1009953814988

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, *74*(3), 036104. https://doi.org/10.1103/physreve.74.036104

Prakash, B. A., Vreeken, J., & Faloutsos, C. (2012). Spotting culprits in epidemics: How many and which ones? *2012 IEEE 12th International Conference on Data Mining*, 11–20. https://doi.org/10.1109/icdm.2012.136

Shakarian, P., Bhatnagar, A., Aleali, A., Shaabani, E., Guo, R., Shakarian, P., Bhatnagar, A., Aleali, A., Shaabani, E., & Guo, R. (2015). The independent cascade and linear threshold models. *Diffusion in Social Networks*, 35–48. https://doi.org/10.1007/978-3-319-23105-1_4

Wang, J., Jiang, J., & Zhao, L. (2022). An invertible graph diffusion neural network for source localization. *Proceedings of the 31th International World Wide Web Conference (WWW 2022)*. https://doi.org/10.1145/3485447.3512155

Wang, Z., Wang, C., Pei, J., & Ye, X. (2017). Multiple source detection without knowing the underlying propagation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*. https://doi.org/10.1609/aaai.v31i1.10477

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Zhu, K., Chen, Z., & Ying, L. (2017). Catch'em all: Locating multiple diffusion sources in networks with partial observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*. https://doi.org/10.1609/aaai.v31i1.10746