cosasi: Graph Diffusion Source Inference in Python

Lucas H. McCabe [1,2]

1 Digital and Analytic Solutions, Logistics Management Institute **2** Department of Mathematics, The George Washington University

Summary

`cosasi` (COntagion Simulation And Source Inference) is, to the author's knowledge, the first extensible open-source framework for graph diffusion source inference that allows users to:

- **perform and evaluate** source localization using standard techniques from literature,
- **contribute** innovative algorithms to a growing core library, and
- **benchmark** new techniques against a battery of comparable schemes.

The software is currently used within the Logistics Management Institute. Additional development continues, and we welcome contribution from the broader academic and industrial communities.

Statement of Need

Because spreading phenomena - including viral epidemics, rumors, and malware - often proceed as a function of pairwise interactions, it is practical to model their propagation as diffusion processes on networks. The source inference/localization problem is that of estimating the inverse of this cascade, aimed at identifying the "patient(s) zero" from partial observations. This problem has captured the attention of epidemiologists, security researchers, social scientists, and more, dating back to Shah and Zaman's seminal work on rumor centrality (Shah & Zaman, 2011).

Since then, source inference algorithms have been developed across subject areas, with practitioners often contributing new techniques in domain-specific venues. Additionally, algorithms tend to be problem-specific, with various solutions preferable for different diffusion processes and network topologies. Finally, researchers interested in novel source localization algorithms may not have time to implement a robust battery of alternatives to compare new schemes against the state-of-the-art.

`cosasi` provides a standard framework for researchers and practitioners alike to perform graph diffusion source inference. The package implements a number of prominent techniques from literature and provides utilities for estimating the number of sources, partitioning infection subgraphs, and more. Where possible, source identification methods are extended as ranking algorithms for hypothesis comparison. `cosasi` also offers a benchmark suite, which automatically implements a battery of comparable localization methods applicable to the graph diffusion use case at hand, enabling users to easily evaluate novel techniques against appropriate baselines. Standardization is emphasized; for instance, all source inference methods return a `SourceResult` object, which provides resources for analyzing, ranking, comparing, and learning more about hypothesized sources and the techniques used.

Background

Given an undirected graph $G = (V, E)$ with vertex set V and edge set E, a diffusion process begins with a source set $S \subseteq V$ and spreads along the edges according to some (usually stochastic) propagation function. It is common for diffusion processes to invoke formalizations from epidemiology, such as the Susceptible-Infected (SI) model, which can represent information spread, or the Susceptible-Infected-Recovered (SIR) model, which can represent dynamics more evocative of viral epidemics. Even when describing metaphorical contagion, such as rumors, it is standard to refer to vertices affected by the spreading process as "infected."

The infection subgraph I_t is the subgraph of G induced by the infected vertices at time t. In the single-source SI model, I_t is guaranteed to be connected. A common setting for source localization is to infer S from some I_t. More recently, some techniques have incorporated information from a small set of observers, who record the time at which they become infected (Zhu et al., 2016).

Broadly speaking, source estimators fall into one of two categories: message-passing algorithms, such as *Short-Fat Tree* (Zhu & Ying, 2014), or spectral algorithms, such as *NETSLEUTH* (Prakash et al., 2012). An extensive overview of source localization techniques is provided by Ying and Zhu (Ying & Zhu, 2018).

Availability and Documentation

`cosasi` is available under the MIT License. The package may be cloned from the GitHub repository or via PyPI: `pip install cosasi`.

Documentation is provided via Read the Docs, including a tutorial introducing major functionality and a detailed API reference. Extensive unit testing is employed throughout the library, with ~97% code coverage.

Similar Software

To the author's knowledge, the only comparable and active source localization software is RPaSDT (Frąszczak, 2022). Here, we enumerate a handful of differences between RPaSDT and `cosasi`, which we believe make `cosasi` preferable for user accessibility, scalability, and community contribution:

- **Presentation**: RPaSDT is a GUI toolkit. `cosasi` is an importable package, with extensive documentation and unit testing.

- **Benchmarking**: RPaSDT does not provide automatic benchmarking, whereas this is a core feature of `cosasi`.

- **Multi-Source Capabilities**: Multi-source inference in RPaSDT is generally performed by partitioning the infection subgraph and applying single-source algorithms to each partition. `cosasi` implements this strategy, as well, but also supports "natural" multi-source inference that does not require repurposing single-source techniques.

- **Estimator Utilities**: When extending single-source algorithms to the multi-source regime (as described above), it is generally necessary to specify the number of clusters into which we partition the infection subgraph - that is, the hypothesized number of infection sources. `cosasi` provides a handful of relevant techniques for estimating this quantity, including the *Eigengap* heuristic (Von Luxburg, 2007) and *Minimum Description Length* (Prakash et al., 2012).

- **Multiple Information Types**: Some source inference algorithms require information other than an infection subgraph. For instance, *Earliest Infection First* relies on a collection

of observers, who report the time at which they become infected (Zhu et al., 2016). cosasi provides multiple methods for providing state information to the source inference modules, enabling a wider array of potential localization algorithms.

Whisper was an earlier, thematically similar web application. The project has been inactive since 2016, the web interface is no longer online, and the underlying library is less feature-rich than cosasi or RPaSDT.

A recent graph autoencoder-based approach by Ling and colleagues performs maximum a posteriori source estimation using a generative prior over diffusion sources (Ling et al., 2022). The corresponding GitHub repository implements their SL-VAE method, but is not a general-purpose diffusion source localization framework.

A handful of libraries exist to simulate diffusion processes on complex networks; OONIS (Karczmarczyk et al., 2021), EoN (Miller & Ting, 2020), contagion (McCabe, 2021), EpiModel (Jenness et al., 2018), and NDlib (Rossetti et al., 2018) are examples with tens of thousands of downloads among them. These, however, only address the *forward* problem (contagion propagation), whereas cosasi is focused on the *inverse* problem (source inference).

Acknowledgements

References

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

Jenness, S. M., Goodreau, S. M., & Morris, M. (2018). EpiModel: Epimodel: An r package for mathematical modeling of infectious disease over networks. *Journal of Statistical Software*, *84*.

Karczmarczyk, A., Jankowski, J., & Wątróbski, J. (2021). OONIS — onis—object-oriented network infection simulator. *SoftwareX*, *14*, 100675. https://doi.org/10.1016/j.softx.2021.100675

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

McCabe, L. (2021). *Lucasmccabe/contagion: v1.3.3* (Version v1.3.3) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4456181

Miller, J. C., & Ting, T. (2020). Eon (epidemics on networks): A fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks. *arXiv Preprint arXiv:2001.02436*. https://doi.org/10.21105/joss.01731

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

Rossetti, G., Milli, L., Rinzivillo, S., Sîrbu, A., Pedreschi, D., & Giannotti, F. (2018). NDlib: A python library to model and analyze diffusion processes over complex networks. *International Journal of Data Science and Analytics*, *5*(1), 61–79. https://doi.org/10.18637/jss.v084.i08

Shah, D., & Zaman, T. (2011). Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, *57*(8), 5163–5181. https://doi.org/10.1109/tit.2011.2158885

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Ying, L., & Zhu, K. (2018). Diffusion source localization in large networks. *Synthesis Lectures on Communication Networks*, *11*(1), 1–95. https://doi.org/10.1007/978-3-031-79285-4

Zhu, K., Chen, Z., & Ying, L. (2016). Locating the contagion source in networks with partial timestamps. *Data Mining and Knowledge Discovery*, *30*(5), 1217–1248. https://doi.org/10.1007/s10618-015-0435-9

Zhu, K., & Ying, L. (2014). Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Transactions on Networking*, *24*(1), 408–421. https://doi.org/10.1109/ita.2013.6502991