SG-t-SNE-Π: Swift Neighbor Embedding of Sparse Stochastic Graphs

Nikos Pitsianis[1,2], **Dimitris Floros**[1], **Alexandros-Stavros Iliopoulos**[2], **and Xiaobai Sun**[2]

1 Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece **2** Department of Computer Science, Duke University, Durham, NC 27708, USA

DOI: 10.21105/joss.01577

Software

- Review
- Repository
- Archive

Submitted: 14 July 2019
Published: 31 July 2019

Summary

SG-t-SNE-Π is a high-performance software for swift embedding of a large, sparse, stochastic graph/network into a d-dimensional space ($d = 1, 2, 3$) on a shared-memory computer, especially on personal laptop and desktop computers. Graphs/networks are an important type of relational data, arising ubiquitously in real-world applications and various research fields. Such data include biological networks, social networks, communication networks, food webs, word co-occurrence networks, see Kovács et al. (2019) and Yang & Leskovec (2015) for more real-world networks. Graph embedding maps each vertex of the graph to a d-dimensional feature vector. Graph embedding into a d-dimensional space with $d = 1, 2, 3$ is frequently used in data-based scientific studies for visual inspection of data, interpretation of network-based analysis results, interactive inquiries and hypothesis generation.

The software SG-t-SNE-Π and its underlying algorithm are built upon precursor algorithms and software for stochastic neighbor embedding of high-dimensional data, namely the original Stochastic Neighbor Embedding (SNE) algorithm by Hinton & Roweis (2003), the algorithm for t-distributed Stochastic Neighbor Embedding (t-SNE) by van der Maaten & Hinton (2008), and their variants (Linderman, Rachh, Hoskins, Steinerberger, & Kluger, 2019; van der Maaten, 2014).[1][2] The t-SNE algorithm has successfully assisted scientific discoveries, as reported in numerous articles in Nature and Science magazines. However, previous t-SNE algorithms and software are limited in two aspects: (i) The algorithms require that the data points be in a metric space and the associated graph (internally generated) be regular with a constant degree. In many real-world networks, the vertices do not readily reside in a metric space, and their degrees vary greatly, far from constant. (ii) The software is limited in practical use either to small graphs/networks or to embedding to $d < 3$ dimensional space. We remove both limitations. SG-t-SNE-Π admits arbitrary, sparse, stochastic graphs/networks. It is demonstrated by Pitsianis, Iliopoulos, Floros, & Sun (2019) for novel, autonomous embedding of large, real-world stochastic networks. SG-t-SNE-Π also enables fast three-dimensional (3D) graph embedding, which preserves and reveals more or even critical structural information as shown by Pitsianis et al. (2019), on modern laptop and desktop computers with ease of use.

SG-t-SNE-Π is implemented in C++. It takes as input a stochastic graph and outputs d-dimensional coordinate vectors. We provide two additional interfaces. The first is to support the conventional t-SNE, with its typical interface and wrappers (van der Maaten, 2014), which converts data points in a metric space to a stochastic k-nearest neighbor graph. The second is

[1]https://github.com/lvdmaaten/bhtsne
[2]https://github.com/KlugerLab/FIt-SNE

a MATLAB interface. SG-t-SNE-Π is used to obtain all numerical experiments in the research article by Pitsianis et al. (2019) and the accompanying supplementary material.[3]

References

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems* (Vol. 15, pp. 857–864).

Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., et al. (2019). Network-based prediction of protein interactions. *Nature Communications*, *10*(1), 1240. doi:10.1038/s41467-019-09177-y

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., & Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*, *16*(3), 243–245. doi:10.1038/s41592-018-0308-4

Pitsianis, N., Iliopoulos, A.-S., Floros, D., & Sun, X. (2019). Spaceland embedding of sparse stochastic graphs. In *IEEE High Performance Extreme Computing Conference (To appear)*. Retrieved from https://arxiv.org/abs/1906.05582

van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, *15*(Oct), 3221–3245.

van der Maaten, L., & Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Yang, J., & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, *42*(1), 181–213. doi:10.1007/s10115-013-0693-z

[3] http://t-sne-pi.cs.duke.edu