Sparsifying Networks While Preserving Properties

CS 328 Project Proposal

Guntas Singh Saran, 22110089
Hrriday V. Ruparel, 22110099
Yajurvedh Bodala, 19110077

March 15, 2024

Under the guidance of

Prof. Anirban Dasgupta
Computer Science and Engineering
IIT Gandhinagar

1 Abstract

In this data science project, we aim to make large graphs more manageable by sampling edges and keeping only a subset while retaining key properties. This allows us to apply community detection algorithms and analyze graph properties before and after sparsifying. Our goal is to assess performance of these sampling techniques in various datasets with community awareness.

2 Sampling Strategies

The following are some sampling strategies that we propose to employ to reduce the size of a graph while maintaining its key structural properties [1] [2]:

- **Random Sampling**: Select edges randomly from the original graph or randomly select a subset of nodes and keep all edges incident to these nodes.
- **Neighbourhood Sampling**: Start with a seed node, randomly select or bias the node choice based on metrics like degree, triangle count, etc. of one of its neighbours, and continue this process iteratively until the desired number of edges or nodes is reached.
- **Stratified Sampling**: Divide the nodes into strata based on certain node attributes (e.g., degree, community membership) and sample uniformly from each stratum.
- **Metropolis-Hastings Sampling** [3]: A Markov chain Monte Carlo method that iteratively samples edges with a probability distribution based on the current state of the chain.
- **Importance Sampling**: Bias the sampling towards edges that are likely to be important for the task at hand, such as high-degree nodes or edges connecting different communities.

3 Community Detection Strategies

These are just a few examples of community detection algorithms that we intend to check for after sparsfying the network:

- **Louvain Method** [4]: The Louvain method is a greedy algorithm for community detection in networks. It optimizes modularity, which measures the quality of the division of a network into communities.
- **Girvan-Newman Algorithm** [5]: The Girvan-Newman algorithm is a hierarchical clustering algorithm based on edge betweenness centrality which iteratively removes edges which are bridges between communities.
- **Modularity-based Clustering** [6]: Modularity-based clustering methods aim to partition a network into communities such that the modularity score is maximized, which quantifies the quality of the community structure.

- **Label Propagation Algorithm (LPA)** [7]: LPA is an algorithm for community detection based on the propagation of labels (community assignments) through the network. LPA converges when the labels stabilize, resulting in the formation of communities where nodes share the same label.
- **Infomap Algorithm** [8]: The Infomap algorithm is based on the concept of information theory and aims to find the most compact and meaningful description of network structure, resulting in a hierarchical community structure.

4 Datasets

- **Social Circles: Facebook**
- **Zachary's Karate Club Network**
- **Citeseer**
- **Amazon Product Co-Purchasing Network**
- **Twitter Follower Network**

References

[1] Mihail N. Kolountzakis et al. "Efficient Triangle Counting in Large Graphs via Degree-Based Vertex Partitioning". In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 15–24. ISBN: 9783642180095. DOI: 10.1007/978-3-642-18009-5_3. URL: http://dx.doi.org/10.1007/978-3-642-18009-5_3.

[2] Aristides Gionis et al. *Community-aware network sparsification*. 2017. arXiv: 1701.07221 [cs.SI].

[3] Christian Hübler et al. "Metropolis Algorithms for Representative Subgraph Sampling". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 283–292. DOI: 10.1109/ICDM.2008.124.

[4] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[5] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12 (June 2002), pp. 7821–7826. ISSN: 1091-6490. DOI: 10.1073/pnas.122653799. URL: http://dx.doi.org/10.1073/pnas.122653799.

[6] Ulrik Brandes et al. "On Modularity Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 20.2 (2008), pp. 172–188. DOI: 10.1109/TKDE.2007.190689.

[7] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks". In: *Physical Review E* 76.3 (Sept. 2007). ISSN: 1550-2376. DOI: 10.1103/physreve.76.036106. URL: http://dx.doi.org/10.1103/PhysRevE.76.036106.

[8] Martin Rosvall and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* 105.4 (Jan. 2008), pp. 1118–1123. ISSN: 1091-6490. DOI: 10.1073/pnas.0706851105. URL: http://dx.doi.org/10.1073/pnas.0706851105.