Finding Patient Zero: Learning Contagion Source with Graph Neural Networks

Sumanth Varambally[5], Chintan Shah[1], Nima Dehmany[2], Nicola Perra[3], Matteo Chinazzi[1], Albert-László Barabasi[1,4], Alessandro Vespignani[1], Rose Yu[5]

[1] Northeastern University, Boston, MA [2] Northwestern University, Evanston, IL [3] Greenwich University, London [4] Harvard University, Boston, MA [5] University of California, San Diego

svarambally@ucsd.edu, shah.ch@northeastern.edu, nimadt@bu.edu, n.perra@qmul.ac.uk, m.chinazzi@northeastern.edu, a.barabasi@northeastern.edu, a.vespignani@northeastern.edu, roseyu@ucsd.edu

Problem

- We are given an interaction graph, whose nodes indicate people, and an edge between two nodes indicates an interaction.
- We are also given a snapshot, indicating the status of each node as susceptible, infected or recovered
- Our goal is to **determine the source of the contagion** (also known as patient zero).

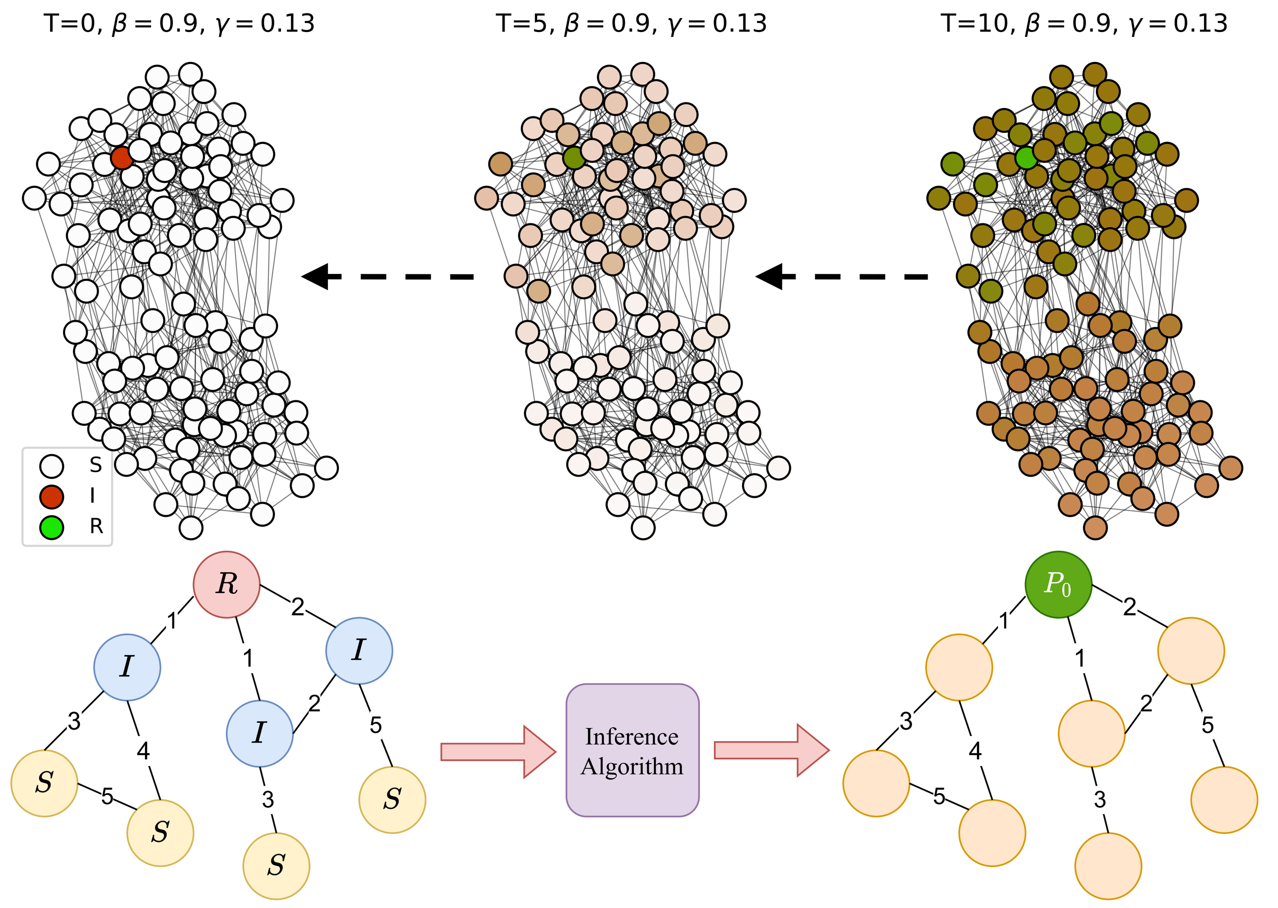


Figure: Graphical representation of our problem

Our Results

We examine GNNs (Graph Neural Networks) as an efficient and flexible inference algorithm for this problem. In particular, we:

- Detail how with residual connections, the standard Graph Convolutional Network (GCN) model can be better adapted to the problem;
- Compare against other state-of-the-art methods on different choices of graph topology for the interaction graph;
- Demonstrate the robustness of GNNs to incomplete node and edge information;
- Examine the topological entropy as a proxy measure for predicting inference performance.

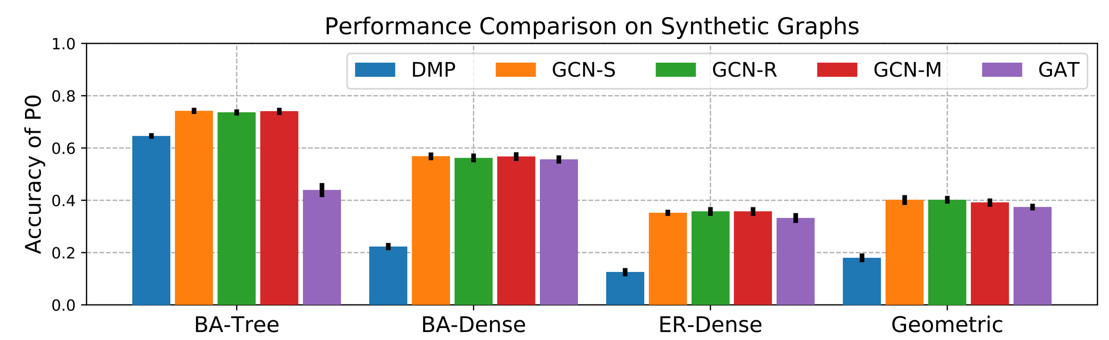
Other applications

- Rumor Source Detection on Social Networks
- Bitcoin Deanonymization
- Viruses spreading over cyber-physical networks

Related Work

- Baseline methods we compared against include Dynamic Message Passing (DMP) and Belief Propagation (`sib`).
- These methods model the likelihood of each node being the origin of the outbreak using the epidemic parameters. However, they do not scale well to large and dense graphs.
- GNNs, on the other hand, do not require knowledge of the epidemic parameters. They are also fast and efficient, and scale well with the number of nodes.
- We note here that both the message passing and belief propagation algorithms can leverage knowledge about the time of each interaction, while the GNNs do not.

Results - Accuracy on different synthetic graph topologies



Inference times

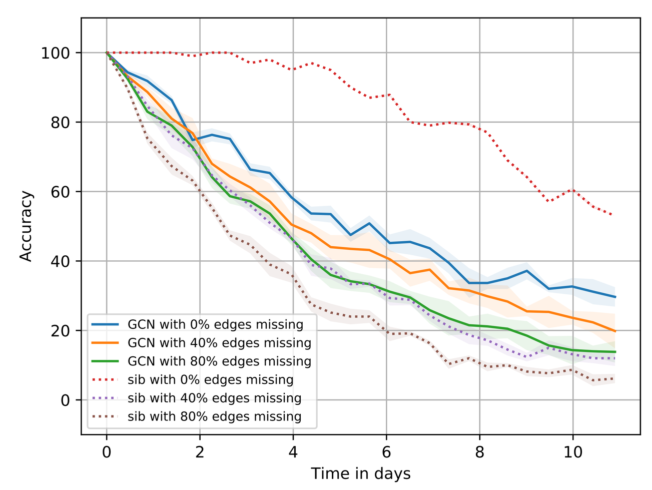
Dataset	DMP	GCN	GAT
BA-Tree	14.40 hr	3.89s	**3.18s**
BA-Dense	77.04 hr	**4.91s**	8.19s
ER-Dense	71.77 hr	**4.93s**	9.66s
Geometric	70.35 hr	**5.34s**	10.87s

The GNN-based methods perform better than the classical message-passing algorithm DMP in all settings while being several orders of magnitude faster.

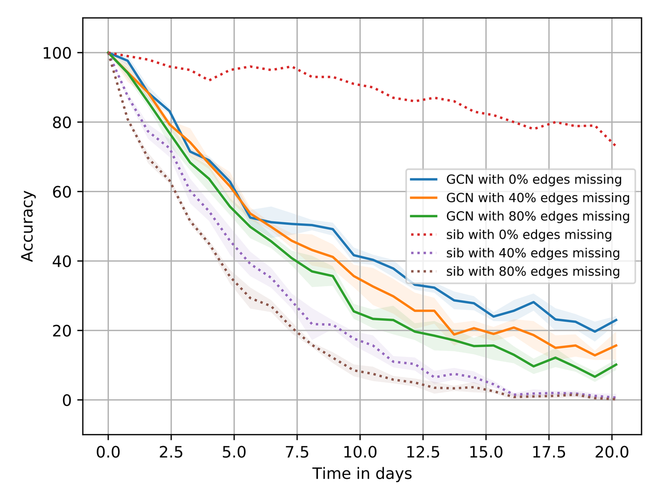
Results - Robustness

Motivation: Most real-world data is incomplete and/or inaccurately reported, e.g. missed interactions, asymptomatic cases, etc. We examine the effectiveness of the inference algorithms despite the missing information.

Missing Edge Information: Remove a proportion p of the edges, chosen randomly, from the interaction graph which is sent as input to the inference algorithm.



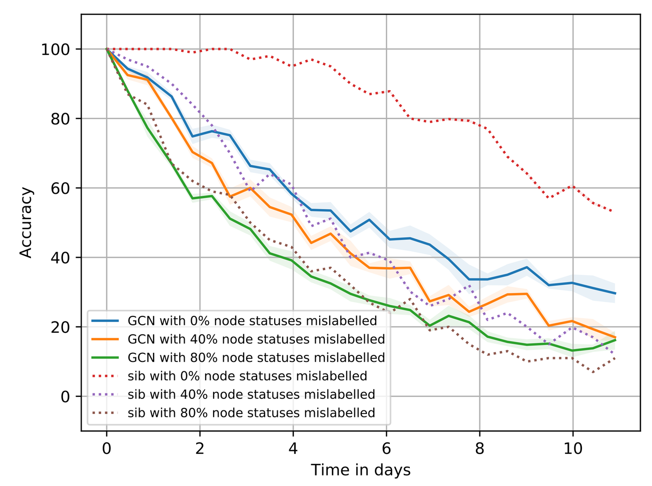
(a) Barabasi-Albert graph with $m = 7$



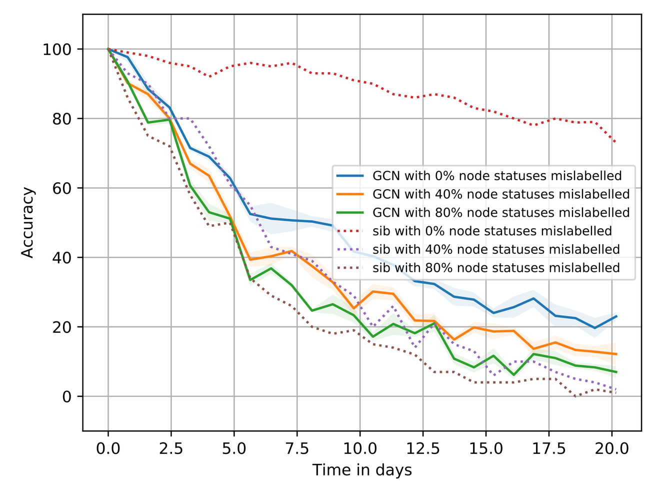
(b) Barabasi-Albert graph with $m = 3$

In figures (a)-(e), `sib` refers to the belief propagation algorithm from [1]. This method leverages time-steps from the social interaction graph during inference (which is not the case with GNNs). With 0 missing edges, `sib` outperforms the GNN methods, however, we notice that even with a small proportion of missing edges, `sib` is outperformed by the GNNs.

Missing node information: Select a proportion p of non-susceptible nodes and mark them as susceptible before the graph is sent to the inference algorithm.



(a) Barabasi-Albert graph with $m = 7$



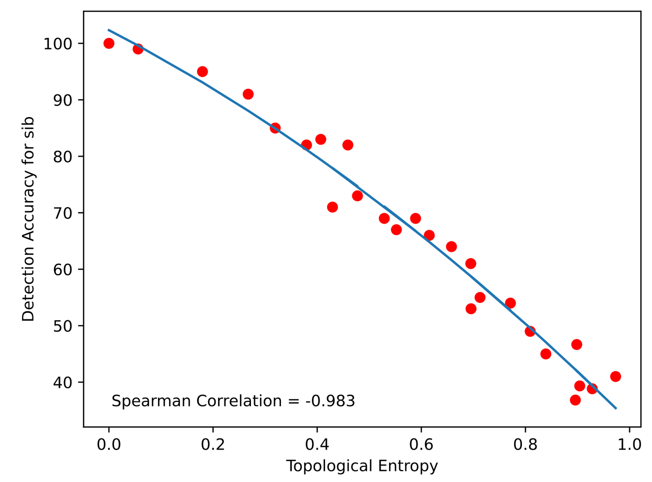
(b) Barabasi-Albert graph with $m = 3$

Again, we notice that despite superior performance when no information is missing, `sib` performs roughly as well as the GNN methods with missing node information, despite leveraging additional temporal information.

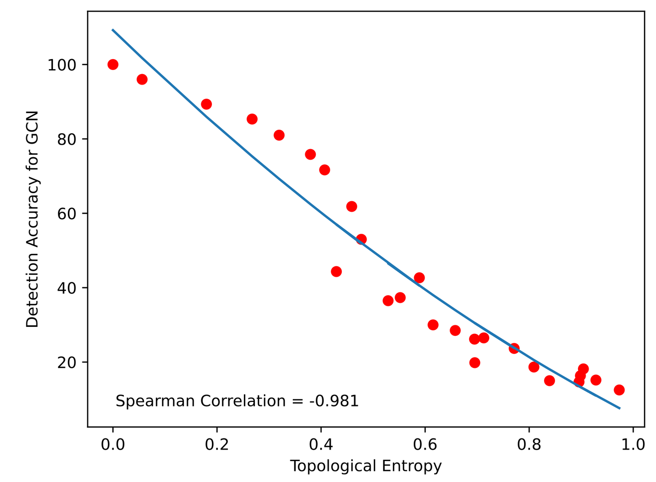
Investigating graph structure vs accuracy

Motivation:

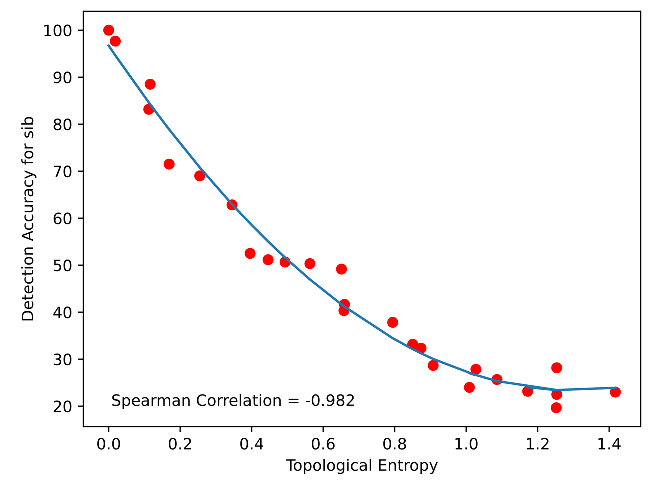
- The accuracy of the inference algorithm at different timescales depends on the topology of the underlying social interaction graph.
- We investigate what property of the graph the inference algorithm is most affected by. We examine the relationship between *topological entropy* (defined as $|\lambda|$ where λ is the largest eigenvalue of the graph adjacency matrix) [2] and patient zero detection accuracy.



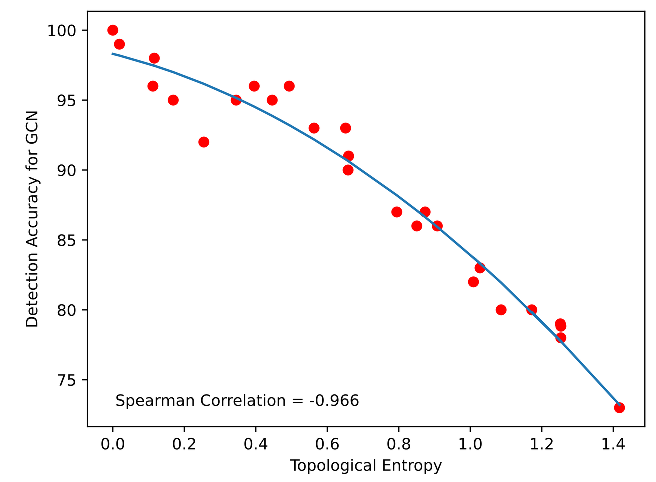
(a) Regular graph + `sib`



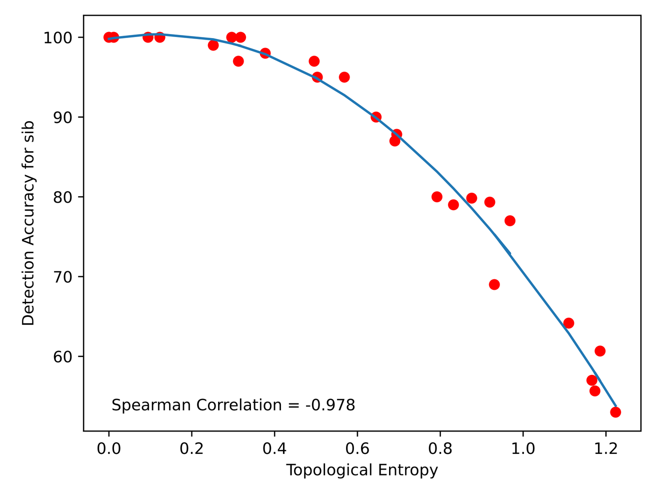
(b) Regular graph + GCN



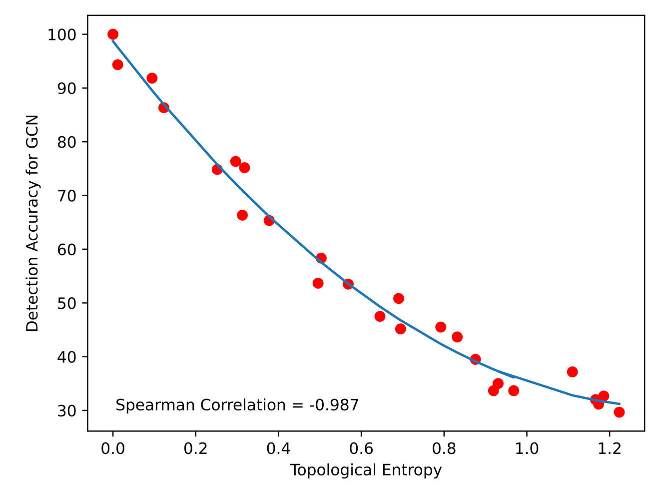
(c) Barabasi-Albert graph ($m = 3$) + `sib`



(d) Barabasi-Albert graph ($m = 3$) + GCN



(e) Barabasi-Albert graph ($m = 7$) + `sib`



(f) Barabasi-Albert graph ($m = 7$) + GCN

We notice a clear trend indicating that *increasing entropy of the social interaction graph decreases the detection accuracy* for both `sib` and GCN. Combined over all datasets, the Spearman correlation between the entropy and detection accuracy is -0.82 for `sib` and -0.86 for GCN, clearly highlighting the inverse relationship between entropy and detection accuracy.

References

[1] A. Braunstein and A. Ingrosso. Inference of causality in epidemics on temporal contact networks. *Scientific reports*, 6(1):1–10, 2016.

[2] L. Demetrius and T. Manke. Robustness and network evolution—an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3):682–696, 2005.

Acknowledgments We thank Brennan Klein, Timothy LaRock, Stefan McCabe, Leo Torres, Lisa Friedland, and Maciej Kos for the help provided in processing and preparing the mobility data. We thank Brennan Lake, Filippo Privitera, and Zachary Cohen for their continuous support. M.C. and A.V.acknowledge support from Google Cloud and Google Cloud Research Credits program to fund this project. This work was supported in part by NSF #1850394, ONR-OTA (N00014-18-9-0001), Google Faculty Research Award and Adobe Data Science Research Award.