Individual Infection Risk Prediction: A Graph Learning Perspective

Galen Harrison[1], Jiangzhuo Chen[2], Henning Mortveit[2], Stefan Hoops[2], Przemyslaw Porebski[2], Dawen Xie[2], Mandy Wilson[2], Parantapa Bhattacharya[2]
Anil Vullikanti[1,2], Li Xiong[3]

[1] Department of Computer Science, University of Virginia
[2] Biocomplexity Institute & Initiative, University of Virginia
[3] Department of Computer Science, Emory University

Network Systems Science & Advanced Computing

Problem: Predicting risk of infection for an individual person

Given observations of infected individuals and a known contact graph, can we predict how likely it is person *i* will become infected within the next week?

Formal Statement: Given graph *G(V,E)*, of *n* individuals, each with time-varying state $s_i(t)$ *{S, I, R}.* Denote the state over time of all the nodes up to time T by ***X****(T),* denote the state of all nodes at time *T+1* by *S(T+1).* Our goal is to predict, for each individual *i*, for *t* in *[T+1, T+D]*, *Pr*[$s_i(t)$ *= I,* | ***X****(T), G*]

This problem (surprisingly) difficult

Note: this setting extends the task of contact tracing from single links to multiple links.

Results from Rosenkrantz et al. (PNAS, 2022) for known disease process suggests that for D > 1, precise prediction is #P-Hard, and approximation is similarly difficult.

Can we find techniques that work well in practice? Certain graph processes can be learned (Narasimhan et al., NeurIPS ’15), and there is some work suggesting that predictions of node state is feasible (Qiu et al. KDD ’18)

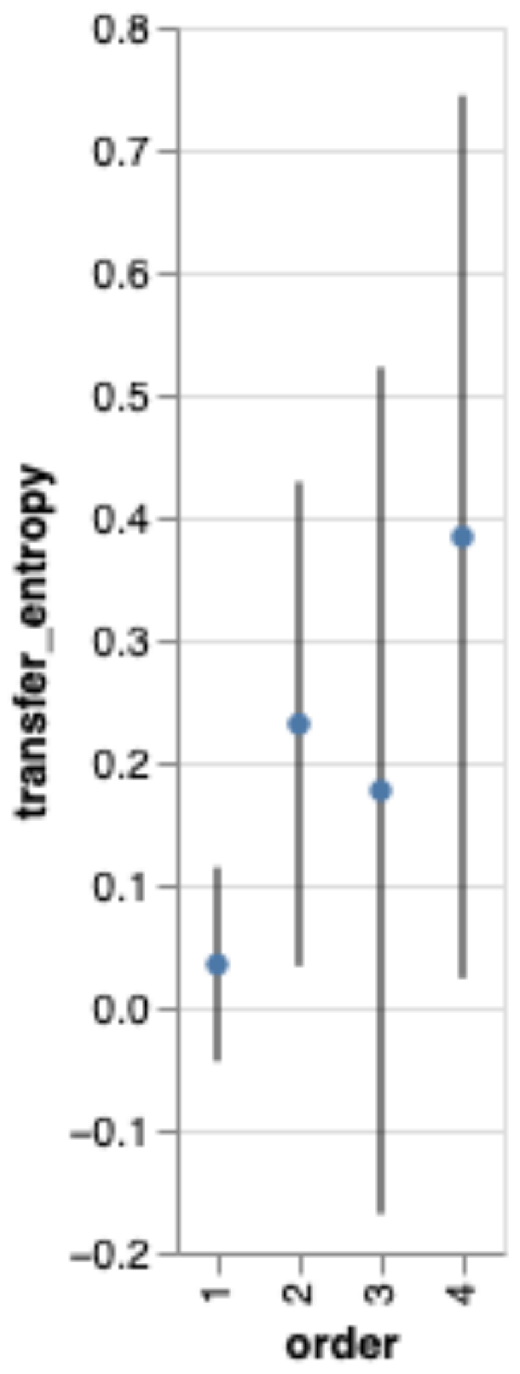
Global infections != local state

Can we infer anything about the risk of a single person or a single group of people from aggregate statistics?

Transfer entropy provides a measure of how much you can predict one time series from another.

$$H(X \to Y) = H(Y_t|Y_{t-1:t-L}) - H(Y_t|Y_{t-1:t-L}, X_{t-1:t-L})$$

Total infections over time does not appear to reduce uncertainty about total infections within a subgraph



Using synthetic data and disease model, compute TE from total infections, to total infections within neighborhood of randomly chosen individual *i*

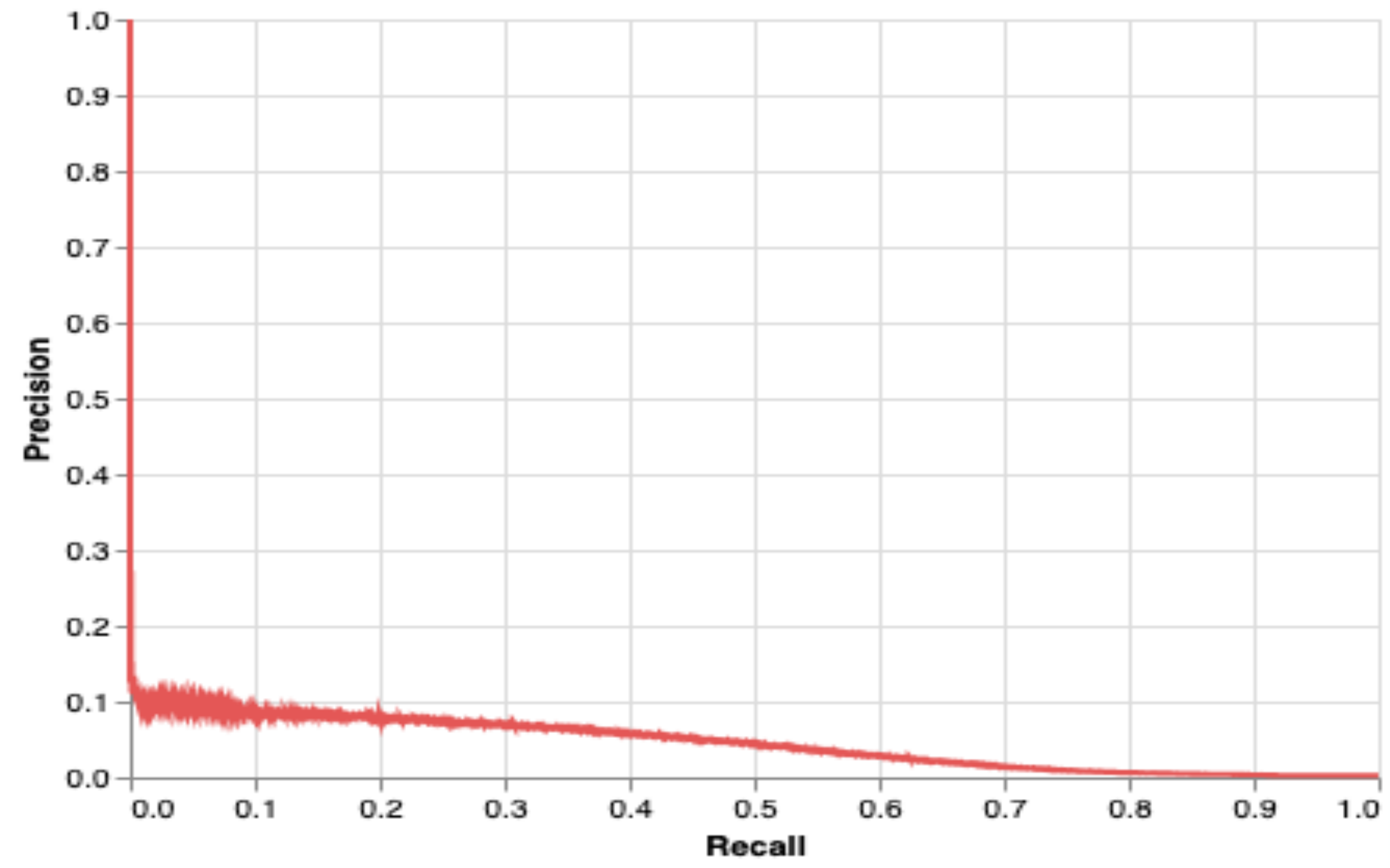
Low TE as neighborhoods get smaller, low amount overall suggest that spread is not “uniform”

Baseline models

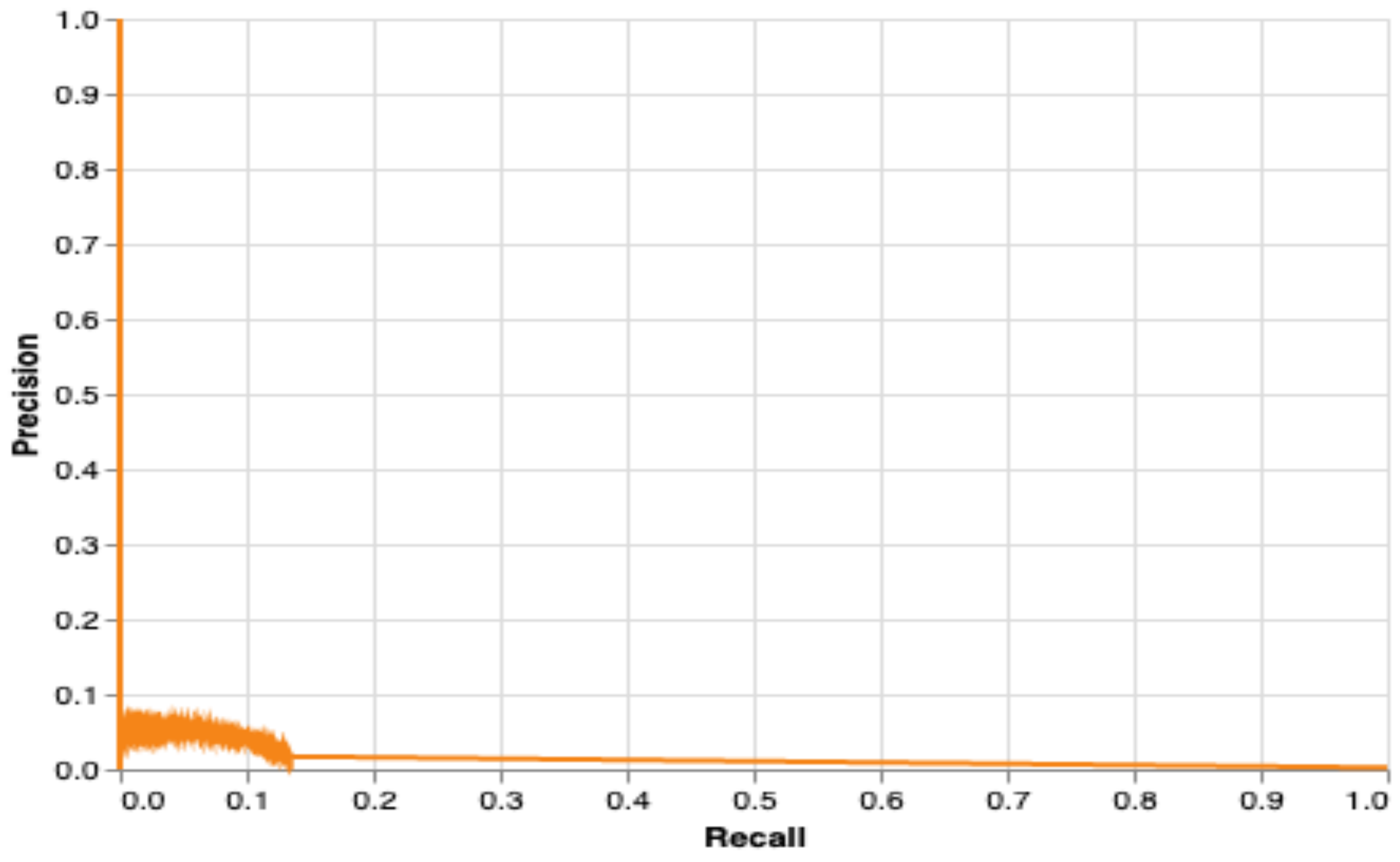
How well do standard models from literature work?

Logistic Regression: LR that takes as input the total number of infected neighbors of *v* at time *t*, the weighted duration of the paths of infected nodes to *v*, labels are whether *v* is infected at any time between *t+1* and *t+D*

DeepInf: A Graph Neural Network using an attention mechanism. Features are embeddings of subgraphs of area around *v*, sampled using random walks, states of each node in subgraph, adjacency matrix of subgraph, labels are whehter *v* is infected at any time between *t+1* and *t+D*



Precision/Recall for *LR*



Precision/Recall for *GNN*

Baseline methods perform above null, however performance is not exactly *good*

Also suffer from scalability issues

University *of* Virginia
BIOCOMPLEXITY INSTITUTE