Link Prediction in Networks with Core-Fringe Data

Austin R. Benson and Jon Kleinberg

arb@cs.cornell.edu, kleinber@cs.cornell.edu

Cornell University

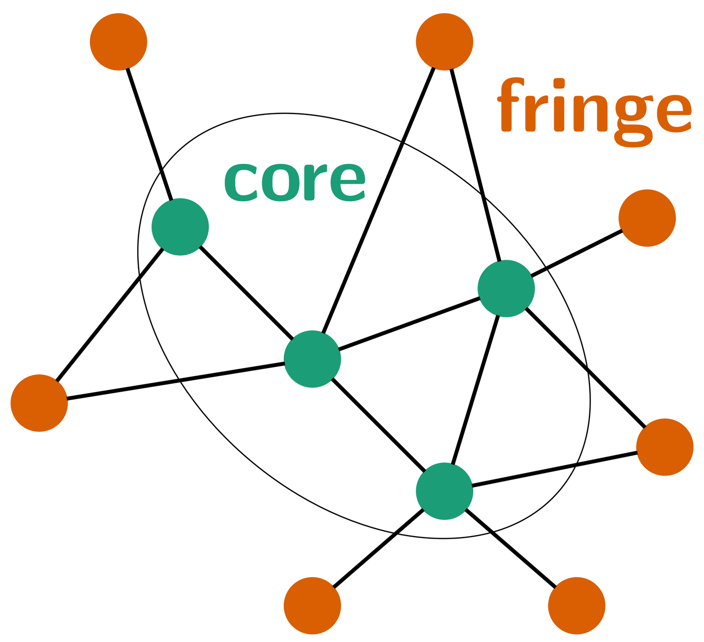
Code & data ⟶ https://github.com/arbenson/cflp

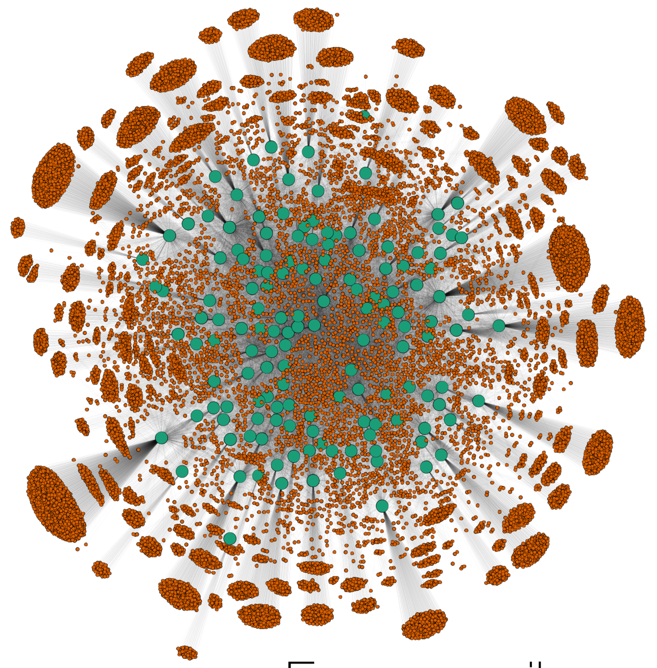
Partially measured graphs & core-fringe data.

We often measure graph data by recording interactions involving a *core* set of nodes:

- Email of company employees
- Phone calls of all customers of a service provider

We end up with a dataset that includes the core along with a potentially much larger set of *fringe* nodes.





Enron emails

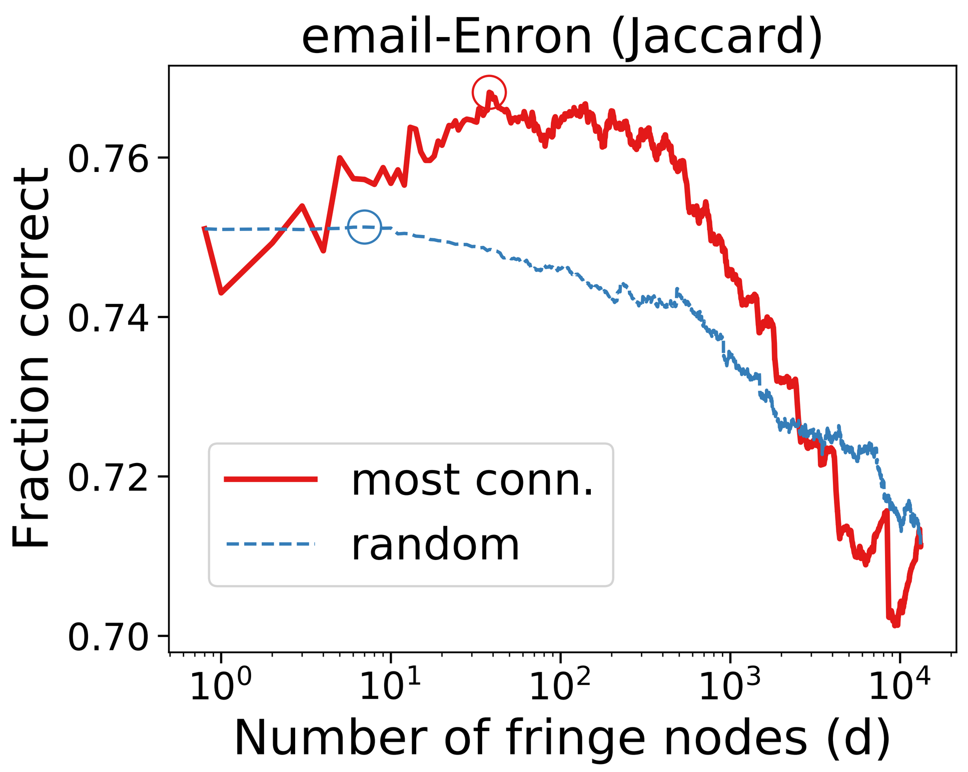
Does knowing the fringe help with link prediction in the core?

1. Fix a link prediction algorithm

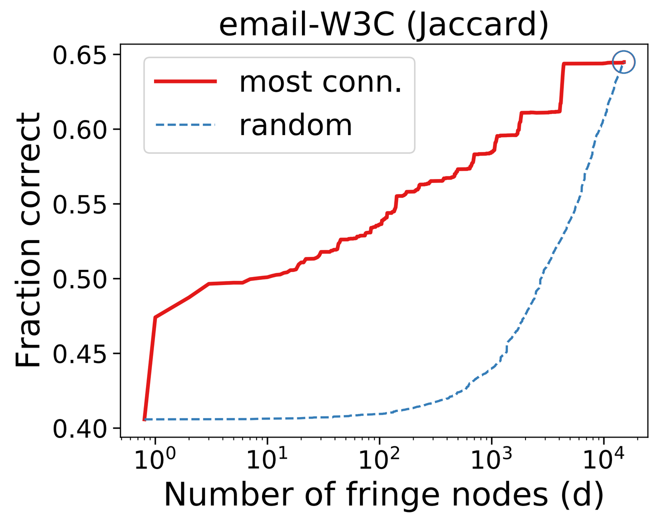
$$\text{CommonNeighbors}(u, v) = |N(u) \cap N(v)|$$

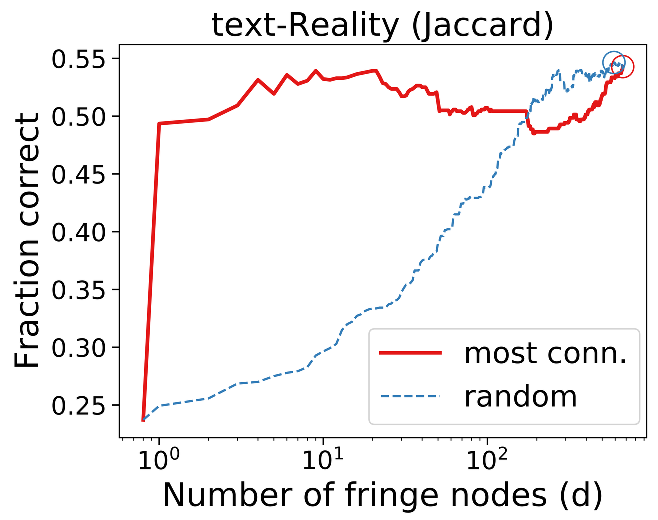
$$\text{Jaccard}(u, v) = |N(u) \cap N(v)| / |N(u) \cup N(v)|$$

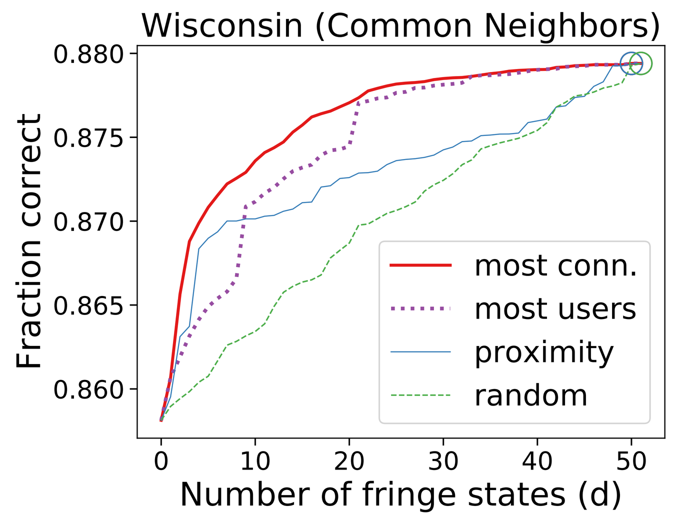
2. Include fringe nodes and connections in some order, *possibly changing the algorithm predictions*.
3. Measure link prediction accuracy as a function of the number of fringe nodes included.



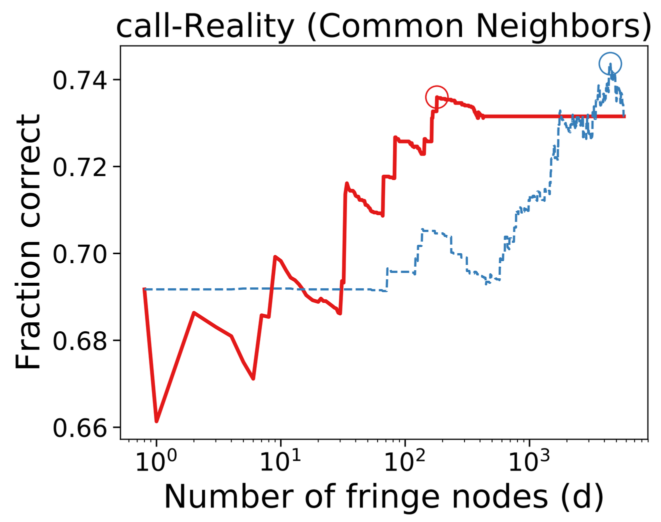
Sometimes, we want all of the information from the fringe.

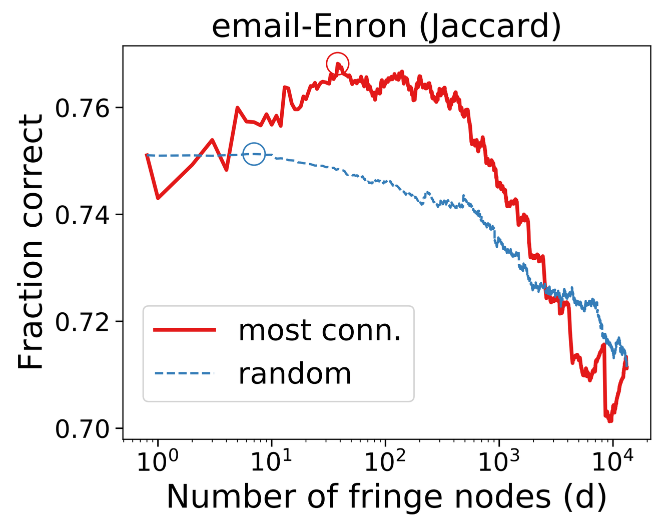


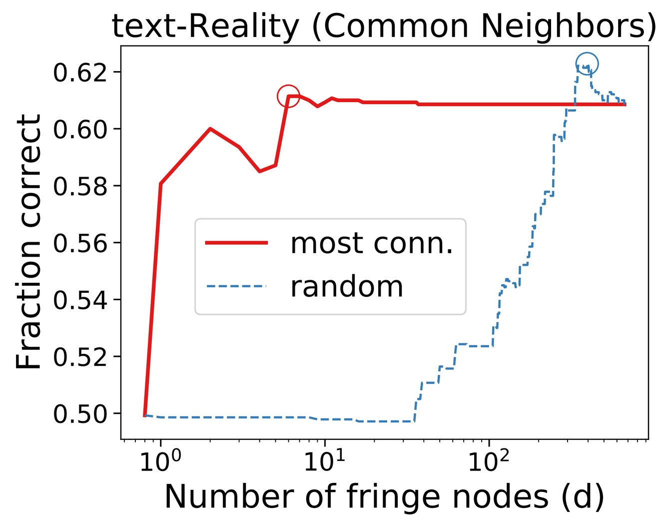




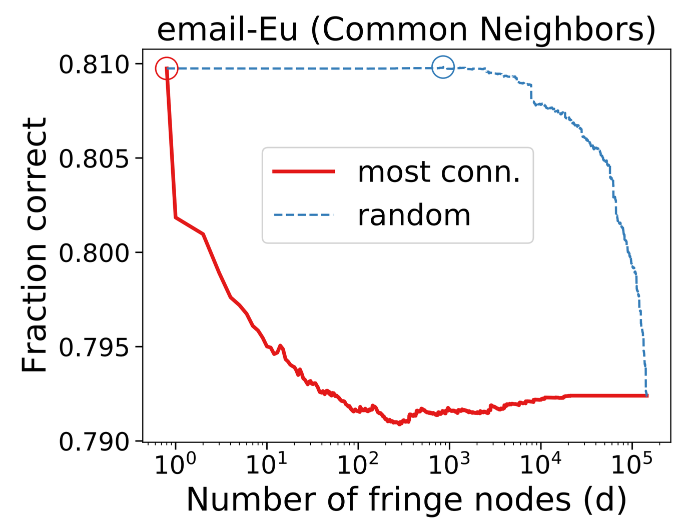
Sometimes, an intermediate amount of fringe gives the best performance.



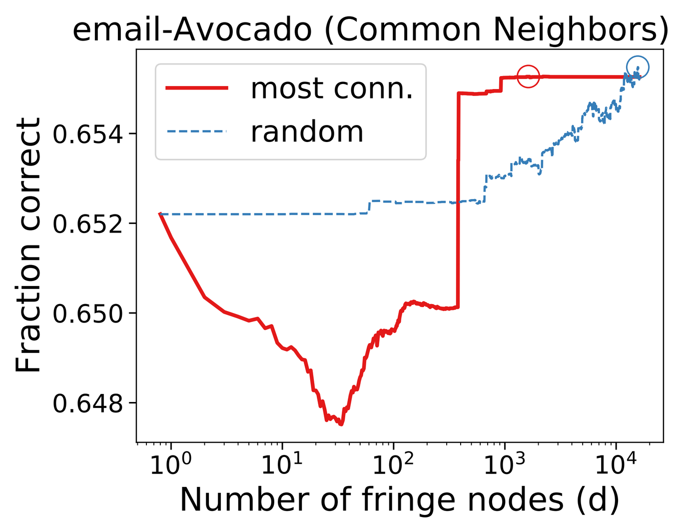


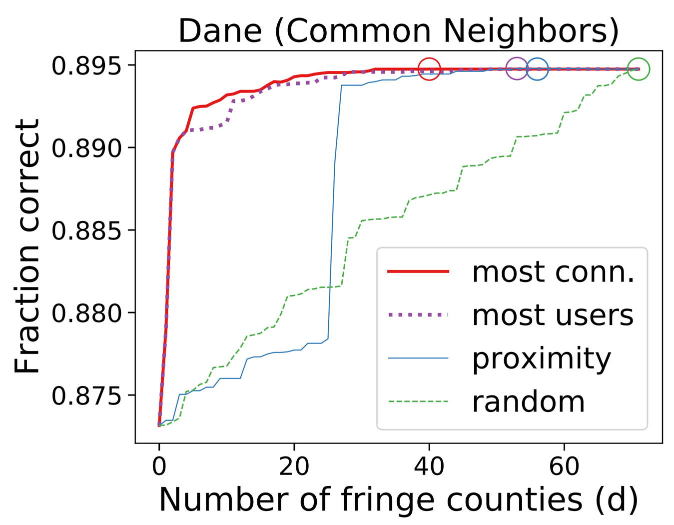


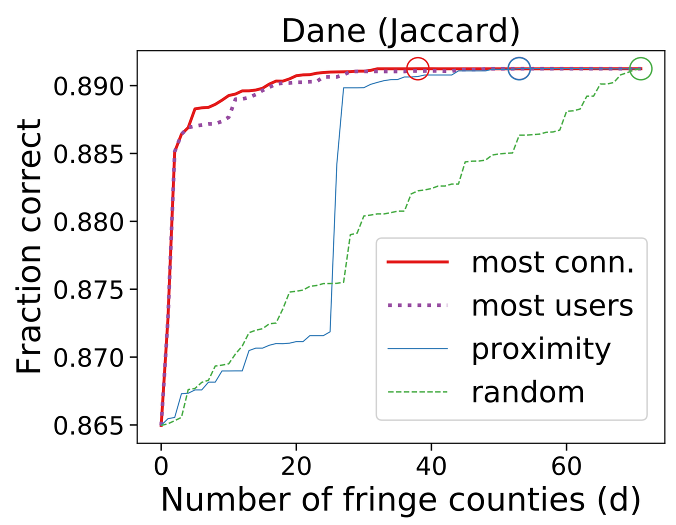
Sometimes, any fringe information hurts.



Sometimes, performance saturates with more fringe information.







Acknowledgements. This research was supported by NSF Awards DMS-1830274, CCF-1740822, and SES-1741441; a Simons Investigator Award; and ARO Award W911NF-19-1-0057.

Random graph models can explain the diversity of behaviors.

- Fix algorithm as number of common neighbors.
- Random graph model where edge {u, v} is more likely than edge {w, z} (latent random variables).
- Algorithm can use fringe info., parameterized by d.
  X_d = CommonNeighbors(u, v)
  Y_d = CommonNeighbors(w, z)
- Goal is to maximize Prob($X_d > Y_d$).
- Solution is

$$\max_d \text{SNR}(Z_d) = \frac{\mathbb{E}(Z_d)}{\sqrt{\mathbb{V}(Z_d)}}, \quad Z_d = X_d - Y_d$$

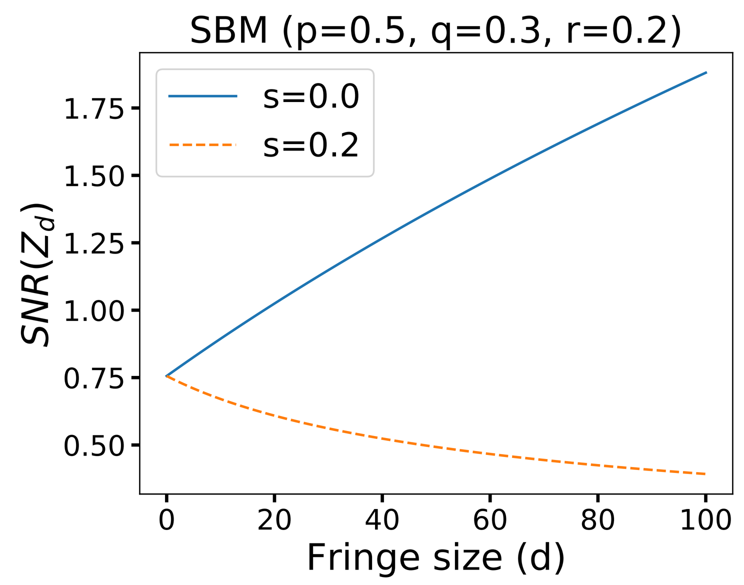
Core-fringe stochastic block model

$$\begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} p & q & r & s \\ q & p & s & r \\ r & s & 0 & 0 \\ s & r & 0 & 0 \end{bmatrix} \end{array}$$

- $p > q$; core is blocks 1 & 2
- u, v, w in block 1; z in block 2
- d is number of fringe nodes included from blocks 3 and 4.

Lemma (no-fringe optimality). If $r = s$, then SNR(Z_d) decreases monotonically in d.

Lemma (all-fringe optimality). If $r > 0$, $s = 0$, then SNR(Z_d) increases monotonically in d.



Core-fringe small-world model

- 1-D lattice of nodes, Prob(edge (i, j)) = 1 / | $i - j$ |
- Core is {-c, -c + 1, …, c – 1, c}
- d includes fringe {-(c + d), …, -(c + 1), c + 1, … c + d}

Theorem (saturation).

$$\lim_{d \to \infty} \text{SNR}(Z_d) = S^* > 0$$

Theorem (intermediate-fringe optimality).

If SNR(Z_0) < SNR(Z_1), then $d^* = \arg\max_d \text{SNR}(Z_d)$ satisfies $0 < d^* < \infty$.