THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

Consistency of common neighbors for link prediction

Purnamrita Sarkar*, Deepayan Chakrabarti* and Peter J. Bickel

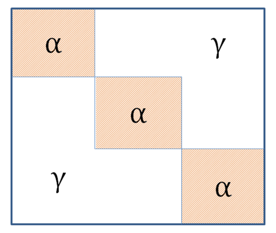
U T Austin*, U C Berkeley

email: purna.sarkar@austin.utexas.edu, deepay@utexas.edu,bickel@stat.berkeley.edu

Questions:

1. (Link prediction/recommendation). Given node i, identify at least a constant number of nodes from i's cluster.
   - Recommending a few friends on Facebook
   - Recommending next few movies to watch on Netflix
   - In these applications, its not necessary to find all nodes in C_i
2. (Local Clustering). For i, identify all nodes in i's cluster.
   - Clearly harder than the first problem.

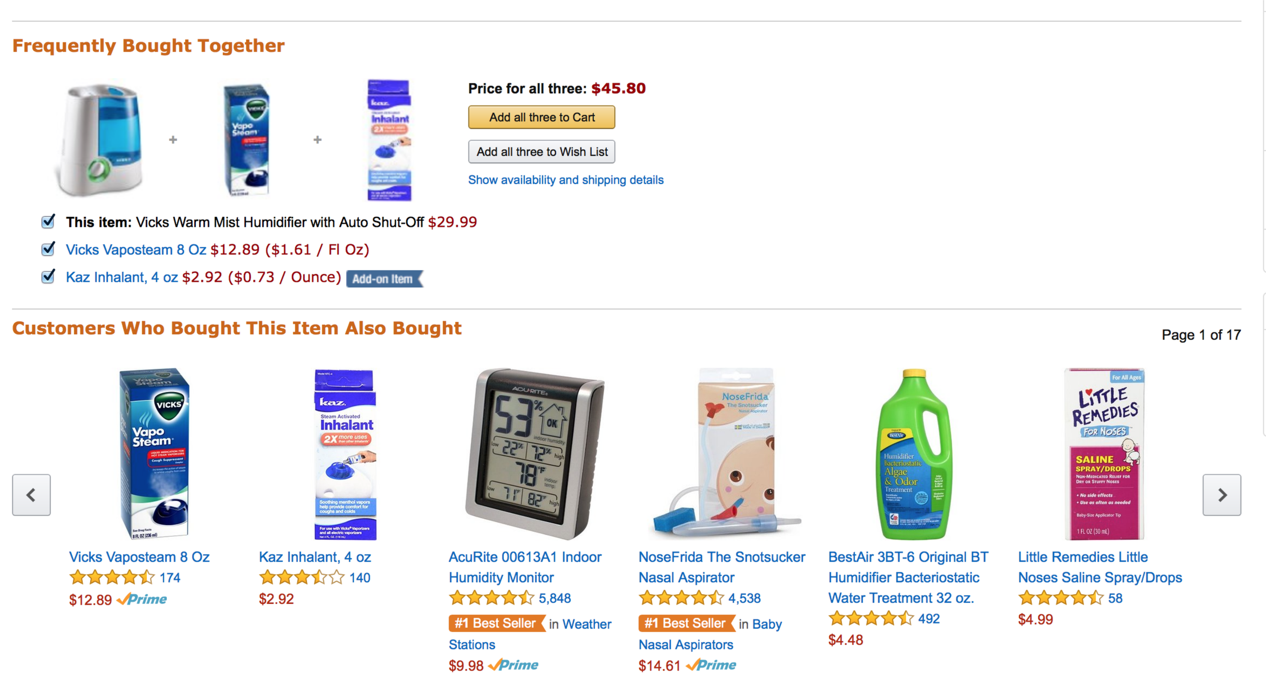
Setup of our stochastic blockmodel



- Fixed number of equal-sized blocks k.
- Assortative clusters $\alpha > \gamma$.
- $\alpha, \gamma = \Theta(\rho)$ where $\rho \to 0$. ρ controls sparsity.

Speed vs. Accuracy

1. **Spectral clustering** yields *strongly consistent* results if:
   - $\frac{\alpha-\gamma}{\sqrt{\alpha}} > C\sqrt{\log n/n}$
   - Average degree grows faster than poly-logarithmic powers of n.
   - Relatively slow for very large graphs.
2. A popular alternative is counting **common neighbors**.

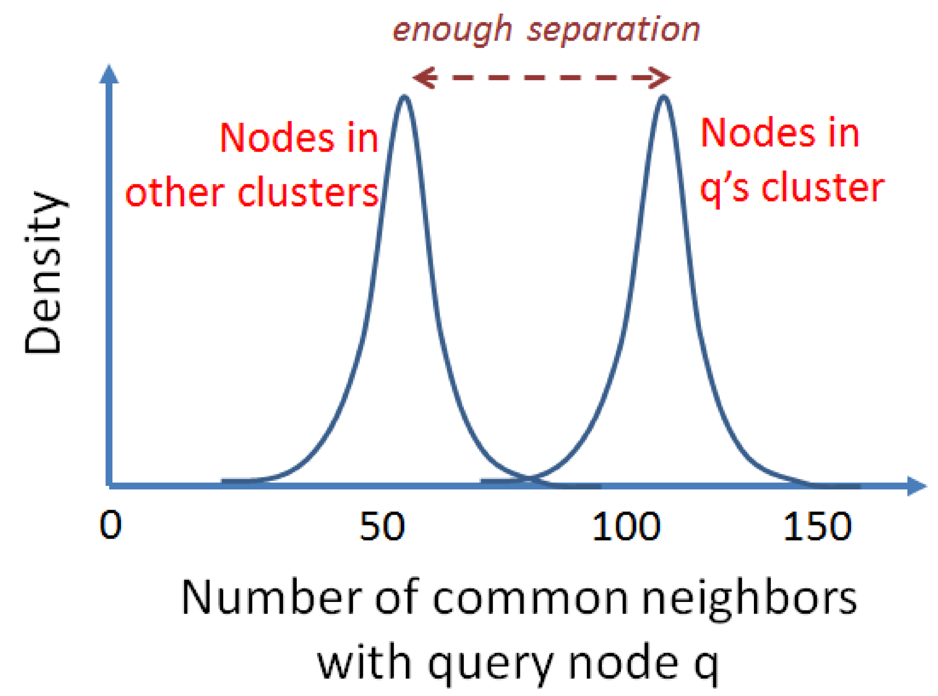


3. **Counting common neighbors** is fast:
   - Only requires database join operations.
   - Works pretty well empirically—here we investigate this formally.

Theory: basic setup

1. Sanity Check:
   - Fix the query node q. Let X_i denote the number of common neighbors between q and any other node i.
   - $E[X_i|C_i = C_q] - E[X_i|C_i \neq C_q] = n\pi(\alpha-\gamma)^2 > 0$.
2. Difficulties:
   - X_i and X_j are dependent quantities – use a conditioning argument.
   - X_i only concentrates when average degree grows faster than $\sqrt{n\log n}$. This is a fairly dense regime. We show that a further preprocessing (*cleaning*) step can recover the entire cluster w.h.p.

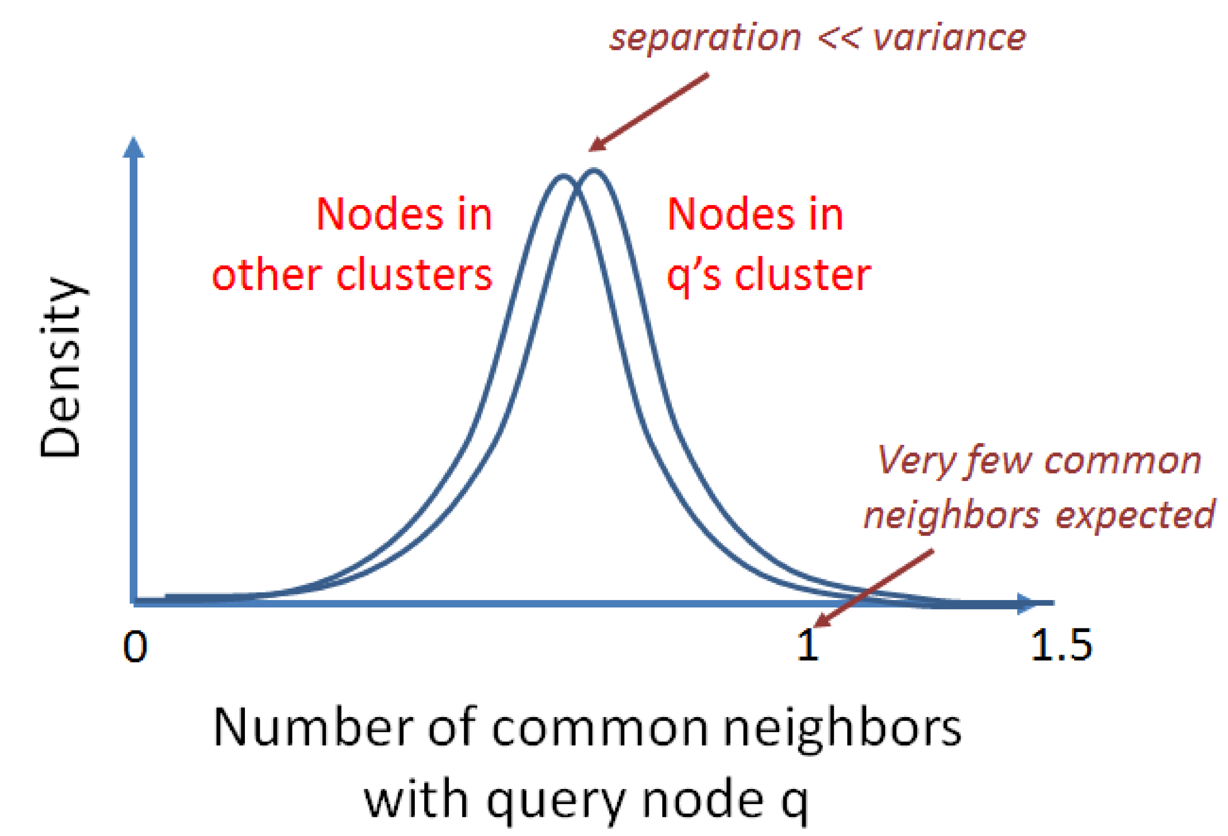
Semi-dense case



- Degree grows faster than $\sqrt{n\log n}$.
- The separation is of a larger order than the standard deviation.
- There exists a threshold t_n, s.t. $S := \{i : X_i > t_n\} = C_q$ w.h.p.
- Practical implication: clustering the X_i's works.

Theorem 1. *When average degree is growing faster than* $\sqrt{n\log n}$, *if* $\frac{\alpha-\gamma}{\alpha} > \frac{2}{\sqrt{\pi}}\left(\frac{\log n}{n\alpha^2}\right)^{1/4}$ *then* $\exists t_n$ $P(|S\cap C_q| = n\pi) \to 1$ *and* $P(|S \setminus C_q| = 0) \to 1$. *Here* $\pi = 1/k$.

Semi-sparse case



- Degree grows faster than $(n\log n)^{1/3}$.
- The separation is of a **smaller** order than the standard deviation.
- Even one common neighbor is rare.
- Let $S = \{i : X_i \geq 1\}$, $n_w := |S \cap C_q|$ (good) and $n_o := |S \setminus C_q|$ (bad).
- We can show that n_w and n_o concentrate and $E[n_w] > E[n_o]$.
- So S has more "good" nodes than "bad" nodes.

Semi-sparse: stronger results

- Use S as a filter.
- For node i, count the number of edges to S (Y_i).
- Y_i concentrates around their expectations.
- These expectations are well separated: $\exists s_n$ such that

$$E[Y_i] > s_n > E[Y_j] \quad \forall i \in C_q,\ \ j \notin C_q.$$

- Practical implication: clustering the Y_i's works.

Theorem 2. *Let* $S_1 = \{i : Y_i > s_n\}$. *When average degree is growing slower than* $\sqrt{n\log n}$ *but faster than* $(n\log n)^{1/3}$, *if* $(\pi\alpha - (1-\pi)\gamma)/(1-\pi)\gamma > 2$, *then for* $t_n = 1$, $P(|S_1 \cap C_q| = n\pi) \to 1$ *and* $P(|S_1 \setminus C_q| = 0) \to 1$.