Inference in Sparse Graphs with Pairwise Measurements and Side Information

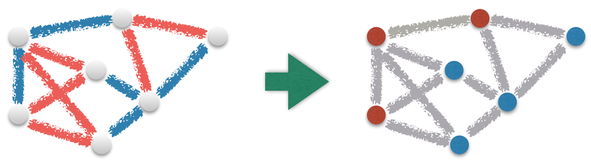
Dylan Foster, Daniel Reichman, and Karthik Sridharan

djfoster@cs.cornell.edu, daniel.reichman@gmail.com, sridharan@cs.cornell.edu

Motivation

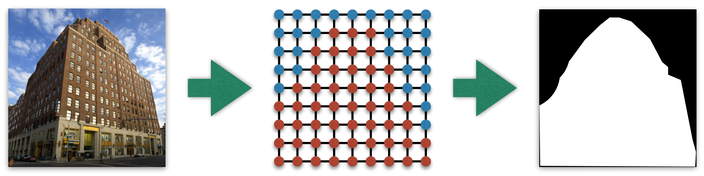
Basic problem: Recover latent node variables using noisy measurements on edges of a graph $G = (V, E)$.

- Community detection



Censored block model: [Abbe et al.`14] [Saade et al. `15]

- Inference for structured prediction (e.g. image segmentation)



[Chen et al. `15] [Globerson,Yildirim, Roughgarden,Sontag`15]

- Alignment/registration/synchronization, correlation clustering, genome assembly, ... many more!

[Joachims/ Hopcroft`05]

Huge body of work on solving / approximating MAP, MLE, etc., but how to establish tight bounds on statistical performance?

Model

Introduced in [Globerson-Roughgarden-Sontag-Yildirim'15].

- Fixed graph $G = (V, E)$, ($|V| = n, |E| = m$).
- Ground truth labels $Y \in \{\pm 1\}^V$.
- Observe *noisy edge labels* $X \in \{\pm 1\}^E$:

$$X_{uv} = \begin{cases} Y_u Y_v, & \text{with prob. } (1-p) \\ -Y_u Y_v, & \text{with prob. } p \end{cases}$$

- Observe *noisy vertex labels* $Z \in \{\pm 1\}^V$:

$$Z_u = \begin{cases} Y_u, & \text{with prob. } (1-q) \\ -Y_u, & \text{with prob. } q \end{cases}$$

Goal: Obtain small *Hamming error*:

$$\mathcal{E}(\widehat{Y}) \triangleq \sum_{v \in V} \mathbb{1}\left\{\widehat{Y}_v(X, Z) \neq Y_v\right\}.$$

aka partial recovery.

Goal: **Error** $= O(h(p)n)$, with $h(p) \to 0$ as $p \to 0$.

Contributions

Our question: How do recovery prospects change with addition of side information?

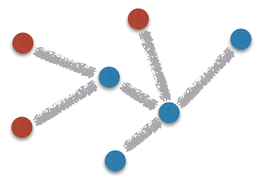
Contributions

- Characterize optimal recovery rates for trees.
- Lift result to general graphs via tree decomposition.
- Non-trivial recovery rates for all connected graphs, including sparse graphs where recovery without side information is impossible.
- All rates finite sample and high probability.
- All achieved efficiently.

Result: Trees

Theorem: Optimal recovery for trees

When G is a tree:

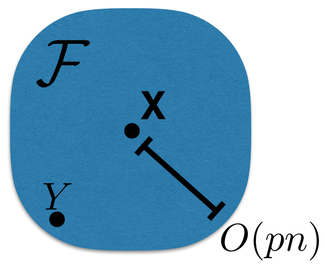


- Efficient algorithm $\widehat{Y}$ with Hamming error **Error** $\leq \widetilde{O}(pn)$ w.h.p.
- Lower bound of $\Omega(pn)$.
- $\to$ **Error** $= \widetilde{O}(pn)$ for all connected G!

Proof sketch:

How to take advantage of Chernoff?

$$\sum_{uv \in E} \mathbb{1}\{X_{uv} \neq Y_u Y_v\} \leq 2pn + O(1) \text{ w.h.p.}$$



Define hypothesis class:

$$\mathcal{F}(X) \triangleq \left\{ Y' \in \{\pm 1\}^V \mid \sum_{uv \in E} \mathbb{1}\{X_{uv} \neq Y'_u Y'_v\} \leq 2pn + O(1) \right\}.$$

Then we have $Y \in \mathcal{F}(X)$ w.h.p., and $|\mathcal{F}(X)| \approx O\left(\left(\frac{1}{p}\right)^{pn}\right)$

Statistical learning reduction:
Take $\widehat{Y}$ to be the empirical risk minimizer:

$$\widehat{Y} = \arg\min_{Y' \in \mathcal{F}(X)} \sum_{v \in V} \mathbb{1}\{Y'_v \neq Z_v\}.$$

Rate for ERM:

$$\sum_{v \in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\} \leq \tilde{O}(\log|\mathcal{F}(X)|/\epsilon^2) \text{ w.h.p. over } Z.$$

Result: General Graphs

For graph $G = (V, E)$, let $\mathcal{W} = \{W_1, \ldots, W_N\}$ be a collection of subsets of V and let $T = (\mathcal{W}, F)$ and be a tree graph over $\mathcal{W}$.

T is a tree decomposition if

1. $\bigcup_{W \in \mathcal{W}} W = V$.
2. For each $uv \in E$, some $W \in \mathcal{W}$ has $u, v \in W$.
3. For $W_1, W_2, W_3 \in \mathcal{W}$, if W_2 on path from W_1 to W_3, need $W_1 \cap W_3 \subseteq W_2$.

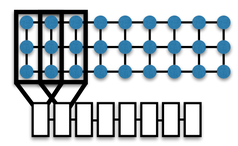
Main theorem: Recovery from tree decomposition
Suppose we have

- $G' = (V, E')$, $E' \subseteq E$.
- Tree decomp. $T = (\mathcal{W}, F)$ for G' w/ constant width, overlap.
- $\mathrm{mincut}(G'(W)) \gtrsim \Delta$ for each $W \in \mathcal{W}$.

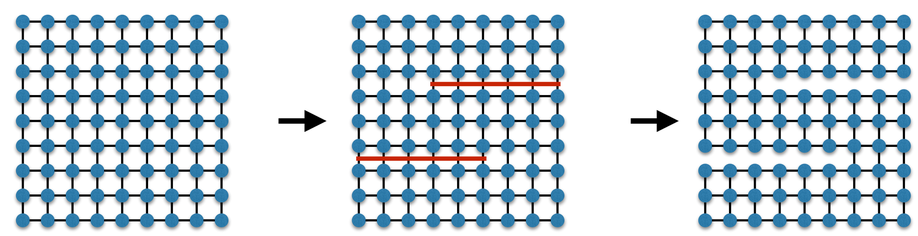
Then there is efficient $\widehat{Y}$ s.t.

$$\textbf{Error} \leq \widetilde{O}(p^{\lceil \Delta/2 \rceil} n).$$

- For grids of size $c \times n$, optimal **Error** $= \Theta(p^2 n)$.



- $\sqrt{n} \times \sqrt{n}$ grid: We recover $O(p^2 n)$, but like so:



Further examples: Hypergrids, lattices, Newman-Watts — see paper for more.