**Motivation**

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

- Community detection
- Inference for structured prediction (e.g., image segmentation)
- Alignment/registration/synchronization, correlation clustering, genome assembly, ... many more!

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

**Result: Trees**

**Theorem: Optimal recovery for trees**

When \( G \) is a tree:

- Efficient algorithm \( \hat{Y} \) with Hamming error \( \text{Error} \leq O(pm) \) w.h.p.
- Lower bound of \( \Omega(pm) \).
- \( \rightarrow \text{Error} = \Omega(pm) \) for all connected \( G \)

**Proof sketch:**

How to take advantage of Chernoff?

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

Define hypothesis class:

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

Then we have \( Y \in \mathcal{F}(X) \) w.h.p., and \( |\mathcal{F}(X)| \approx O\left(\left(\frac{1}{2}\right)^{pn}\right)\)

Statistical learning reduction:

Take \( \hat{Y} \) to be the empirical risk minimizer:

\[
\hat{Y} = \arg \min_{Y' \in \mathcal{F}(X)} \sum_{v \in V} \mathbb{I}\{Y'_v \neq Y_v\}
\]

Rate for ERM:

\[
\sum_{v \in V} \mathbb{P}(\hat{Y}_v \neq Y_v) \leq \tilde{O}(\log |\mathcal{F}(X)|/\epsilon^2) \text{ w.h.p. over } Z.
\]

**Result: General Graphs**

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

\( T \) is a tree decomposition if

1. \( \bigcup_{W \in W} W = V \).
2. For each \( uv \in E \), some \( W \in W \) has \( u, v \in W \).
3. For \( W_1, W_2, W_3 \in W \), 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

- \( G' = (V', E') \subseteq E \).
- Tree decomp. \( T = (W, F) \) for \( G' \) w/ constant width, overlap.
- mincut\((G'(W)) \geq \Delta \) for each \( W \in W \).

Then there is efficient \( \hat{Y} \) s.t.

\[
\text{Error} \leq \tilde{O}(p^{\Delta/2} n).
\]

- For grids of size \( c \times n \), optimal \( \text{Error} = \Theta(p^2 n) \).
- \( \sqrt{n} \times \sqrt{n} \) grid: We recover \( O(p^2 n) \), but like so:

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

**Contributions**

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

- 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.

**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:

\[
\mathcal{E}(\hat{Y}) \triangleq \sum_{v \in V} \mathbb{I}\{\hat{Y}_v(X, Z) \neq Y_v\}.
\]

aka partial recovery.

Goal: \( \text{Error} = O(h(p)n) \), with \( h(p) \to 0 \) as \( p \to 0 \).