The tree reconstruction problem

Tree $\mathcal{T}$, root $r$. $\mathcal{L}_d$: nodes in generation $d$ (at distance $d$ from $r$).
Tree of nodes of generations $0, \ldots, d$: $\mathcal{T}_d = (V_d, E_d)$.

$\sigma_i \in [q]$: “trait” of individual $i$. $p(i)$: parent of $i$.

Probabilistic transmission: $P(\sigma_{\mathcal{L}_d} = s_{\mathcal{L}_d} | \mathcal{T}, \sigma_{\mathcal{V}_{d-1}}) = \prod_{i \in \mathcal{L}_d} P_{\sigma_{p(i)}s_i}$ where $P$: stochastic matrix, assumed irreducible, with invariant distribution $\nu$ on $[q]$.
The tree reconstruction problem

Assume root spin $\sigma_r \sim \nu$. Then $\mathbb{P}(\sigma_V = s_V | \mathcal{T}) = \nu_s \prod_{(i,j) \in E, i = p(j)} P_{s_i s_j}$

→ A tree Markov field.

Special case: $P_{\tau\tau} = p$, $P_{\tau s} = \frac{1-p}{q-1}$, $s \neq \tau$: symmetric Potts model ($q = 2$: Ising model).

Let $\mathcal{F}_d = \sigma(\mathcal{T}_d, \sigma_V_d)$, $\mathcal{G}_d = \sigma(\mathcal{T}_d, \sigma_L_d)$, $\hat{\nu}_{s,d} = \mathbb{P}(\sigma_r = s | \mathcal{G}_d)$, $s \in [q]$.

**Definition**

tree reconstruction is feasible if and only if $\lim_{d \to \infty} I(\sigma_r; \mathcal{G}_d) > 0$. 
Census reconstructibility and Kesten-Stigum threshold

Define generation \( d \)'s census: \( X_d = \{X_{s,d}\}_{s \in [q]} \) where \( X_{s,d} := \sum_{i \in \mathcal{L}_d} \mathbb{I}_{\sigma_i = s} \).

**Definition**

Census reconstructibility holds if \( \lim_{d \to \infty} I(\sigma_r; X_d) > 0 \).

Assume \( \mathcal{T} \): Galton-Watson, with r.v. \( Z \): number of children verifying \( \mathbb{E} Z = \alpha > 1 \) and \( \mathbb{E} Z^2 < \infty \).

For transition matrix \( P_{s \tau} := \mathbb{P}(\sigma_i = \tau | \sigma_{p(i)} = s) \), let \( \lambda_2(P) \): eigenvalue of \( P \) with second largest modulus (\( \lambda_1(P) = 1 \)).

**Theorem**

If \( \alpha |\lambda_2(P)|^2 > 1 \), census reconstructibility holds.
Census reconstructibility and Kesten-Stigum threshold

Theorem

Reciprocally, for $Z \sim Poi(\alpha)$ with $\alpha > 1$ such that $\alpha|\lambda_2(P)|^2 < 1$, then
\[
\lim_{d \to \infty} I(\sigma_r; X_d) = 0, \text{ i.e. census reconstruction fails.}
\]

Remark

Result still true for more general branching processes. It holds for instance with $Z \equiv \alpha \in \mathbb{N}^*$. 
Theorem (Kesten-Stigum, “Additional limit theorems for indecomposable multidimensional G-W processes”, 1966)

Below threshold, i.e. when $\alpha|\lambda_2|^2 < 1$, conditional on $\sigma_r = \tau \in [q]$, 
\[ \left\{ \alpha^{-d/2}(X_{s,d} - \alpha^d \nu_s) \right\}_{s \in [q]} \xrightarrow{d \to \infty} \mathcal{N}(m, \Sigma), \]
where $m, \Sigma$ do not depend on $\tau \in [q]$.

Corollary (Kesten-Stigum, Coupling)

For all $d \in \mathbb{N}, \tau, \tau' \in [q]$ there exists coupling of census vectors $X_d^{(\tau)}, X_d^{(\tau')}$ corresponding to $\sigma_r = \tau, \tau'$ respectively such that 
\[ \forall \epsilon > 0, \lim_{d \to \infty} \mathbb{P}\left( \left\| X_d^{(\tau)} - X_d^{(\tau')} \right\| \geq \epsilon \alpha^{d/2} \right) = 0. \]
For \( t \in \{\tau, \tau'\} \), \( \mathcal{L}(X_{d+1}^{(t)} \mid X_d^{(t)}) = \otimes_{s \in [q]} \text{Poi}(M_s^{(t)}) \), where

\[
M_s^{(t)} = \alpha \sum_{s' \in [q]} X_{s',d} \, P_{s's}.
\]

Let \( M_s = \frac{1}{2}(M_s^{(\tau)} + M_s^{(\tau')}) \) and \( \epsilon_s = \frac{1}{2}|M_s^{(\tau)} - M_s^{(\tau')}| M_s^{-1/2} \).

By [Kesten-Stigum, Coupling] Corollary, \( \exists \alpha_d \xrightarrow{d \rightarrow \infty} 0 \) such that \( \forall s \in [q], \mathbb{P}(\epsilon_s \leq \alpha_d) \xrightarrow{d \rightarrow \infty} 1 \).

**Lemma**

*Variation distance*

\[
d_{\text{var}}(\mu, \nu) := 2 \sup A |\mu(A) - \nu(A)| \quad \text{also equals} \quad 2 \inf_{(X,Y) \text{ coupling of } (\mu, \nu)} \mathbb{P}(X \neq Y).
\]

**Corollary**

\[
d_{\text{var}}(\otimes_{s \in [q]} \mu^{(s)}, \otimes_{s \in [q]} \nu^{(s)}) \leq \sum_{s \in [q]} d_{\text{var}}(\mu^{(s)}, \nu^{(s)}).
\]
Hence $d_{\text{var}}(X_{d+1}^{(\tau)}, X_{d+1}^{(\tau')}, X_{d}^{(\tau)}, X_{d}^{(\tau')}) \leq \cdots \\
\cdots \sum_{s \in [q]} \sum_{k \geq 0} |\text{Poi}_{M_s^{(\tau)}}(k) - \text{Poi}_{M_s^{(\tau')}}(k)| =: \sum_{s \in [q]} A_s$

Split sums $A_s$ according to whether $|M_s - k| \leq \omega_d \sqrt{M_s}$ or not, where $\omega_d = \frac{1}{\sqrt{\alpha_d}}$, i.e. $A_s = A_{s, \leq} + A_{s, >}$.

Write

$$A_{s, >} \leq \mathbb{P}(|\text{Poi}_{M_s^{(\tau)}}(k) - M_s| \geq \omega_d \sqrt{M_s}) + \mathbb{P}(|\text{Poi}_{M_s^{(\tau')}}(k) - M_s| \geq \omega_d \sqrt{M_s})$$

Note that $M_s^{(\tau)} = M_s \pm \epsilon_s \sqrt{M_s}$ so that on event $\{\epsilon_s \leq \alpha_d\}$,

w.h.p. $|\text{Poi}_{M_s^{(\tau)}}(k) - M_s| < \omega_d \sqrt{M_s}$.

Thus: $\lim_{d \to \infty} E(A_{s, >}) = 0$. 
\[ A_{s, \leq} \leq \sum_{k: |k - M_s| \leq \omega_d \sqrt{M_s}} e^{-M_s \frac{M_s^k}{k!}} \left| e^{-\varepsilon_s \sqrt{M_s}} \left(1 + \frac{\varepsilon_s}{\sqrt{M_s}}\right)^k - e^{\varepsilon_s \sqrt{M_s}} \left(1 - \frac{\varepsilon_s}{\sqrt{M_s}}\right)^k \right| \]

On the event \( \{\varepsilon_s \leq \alpha_d\} \), for \( k: |k - M_s| \leq \omega_d \sqrt{M_s} \), one has:

\[ e^{\pm \varepsilon_s \sqrt{M_s}} \left(1 \mp \frac{\varepsilon_s}{\sqrt{M_s}}\right)^k = e^{\pm \varepsilon_s \sqrt{M_s} + k(\mp \varepsilon_s / \sqrt{M_s} + O(\varepsilon_s^2 / M_s))} = e^{O(\varepsilon_s \omega_d)} \]

\[ = 1 + O(\sqrt{\alpha_d}). \]

Thus \( A_{s, \leq} \leq |1 + O(\sqrt{\alpha_d}) - 1 - O(\sqrt{\alpha_d})| = O(\sqrt{\alpha_d}). \)

By Jensen’s inequality

\[ d_{\text{var}}(X^{(\tau)}_{d+1}, X^{(\tau')}_{d+1}) \leq \mathbb{E}[d_{\text{var}}(X^{(\tau)}_{d+1}, X^{(\tau')}_{d+1} | X^{(\tau)}_d, X^{(\tau')}_d)] \]

Thus \( d_{\text{var}}(X^{(\tau)}_{d+1}, X^{(\tau')}_{d+1}) \leq \sum_{s \in [q]} \mathbb{E}(A_{s, >} + A_{s, \leq}) \xrightarrow{d \to \infty} 0. \)
Theorem then follows from

**Lemma**

*Mutual information* $I(\sigma_r; X_d)$ *is upper-bounded by* $q \times \sup_{s, \tau \in[q]} d_{var}(P(X_d \in \cdot | \sigma_r = s), P(X_d \in \cdot | \sigma_r = \tau))$.

Lemma’s proof: define $f_s(x) = P(X_d = x | \sigma_r = s)/P(X_d = x)$, $x \in \mathbb{N}^q$. It verifies: $\sum_{\tau \in [q]} \nu_\tau f_\tau(x) \equiv 1$.

Write:

$$I(\sigma_r; X_d) = \sum_{s \in[q], x \in \mathbb{N}^q} \nu_s P(X_d = x | \sigma_r = s) \ln \left( \frac{P(X_d = x | \sigma_r = s)}{P(X_d = x)} \right)$$

$$= \sum_{x \in \mathbb{N}^q} P(X_d = x) \sum_{s \in[q]} \nu_s f_s(x) \ln(f_s(x))$$

$$\leq \sum_{x \in \mathbb{N}^q} P(X_d = x) \sum_{s \in[q]} \nu_s f_s(x) [f_s(x) - 1]$$

$$= \sum_{\tau \in [q]} \nu_\tau \sum_{x \in \mathbb{N}^q} P(X_d = x) \sum_{s \in[q]} \nu_s f_s(x) [f_s(x) - f_\tau(x)]$$

$$\leq \sum_{s, \tau \in[q]} \nu_\tau \sum_{x \in \mathbb{N}^q} P(X_d = x) |f_s(x) - f_\tau(x)|$$

$$= \sum_{s, \tau \in[q]} \nu_\tau d_{var}(P(X_d \in \cdot | \sigma_r = s), P(X_d \in \cdot | \sigma_r = \tau))$$
Tree reconstruction threshold for symmetric case with $q = 2$

For $q = 2$, take $\sigma_i = \pm$. Symmetry: $P_{++} = P_{--} = 1 - \epsilon$, $P_{-+} = P_{+-} = \epsilon$.

Notation: let $\theta = \lambda_2 (P) = 1 - 2\epsilon$, so that $\mathbb{E}(\sigma_i | \sigma_p(i)) = \theta \sigma_p(i)$, $\mathbb{E}(\sigma_i \sigma_p(i)) = \theta$.

Theorem (Evans et al., Broadcasting on trees and the Ising model, 2000)

For symmetric $q = 2$ propagation on deterministic tree $T$ such that

$$\limsup \frac{1}{d} \ln(|L_d|) \leq \ln(\alpha),$$

tree reconstruction fails when $\alpha (\lambda_2)^2 < 1$.

Corollary

For symmetric $q = 2$ propagation on Galton-Watson tree $T$,

Kesten-Stigum threshold provides necessary and sufficient condition for tree reconstruction (ignoring equality case $\alpha (\lambda_2)^2 = 1$).
Lemma (Evans et al.’00)

Consider trees $T$, $T'$ above, where node variables are binary spins, each uniformly distributed with values $\pm 1$, edge weights $\in [0, 1]$ represent transmission probability, e.g. $E(\sigma_r \tau_1) = \theta$.

Then there exists a probability transition matrix $M^0 : \{-1, 1\}^2 \to \{-1, 1\}^2$ such that

$$\mathbb{P}(\sigma_r' = s_r, \sigma_1' = s_1, \sigma_2' = s_2) = \sum_{u_1, u_2 = \pm} \mathbb{P}(\sigma_r = s_r, \sigma_1 = u_1, \sigma_2 = u_2) \times \cdots \times M^0_{(u_1, u_2), (s_1, s_2)}$$
Lemma (channel between trees)

For two random vectors $U \in \{\pm 1\}^a$, $V \in \{\pm 1\}^b$, mutually independent and independent of the spins of the two trees on previous Figure, let $X = \sigma_1 U$, $Y = \sigma_2 V$, $X' = \sigma_1' U$, $Y' = \sigma_2' V$. Then there is a probability transition matrix $M$ on $\{\pm 1\}^{a+b}$ such that

$$P(\sigma'_r = s, (X', Y') = (x', y')) = \sum_{x,y} P(\sigma_r = s, (X, Y) = (x, y)) M_{(x,y);(x',y')}$$

Proof: for vectors $(x, y) \in (\pm)^{a+b}$, define

$$M_{(x,y),(x',y')} = \sum_{t_1, t_2, s_1, s_2 = \pm} \mathbb{I}_{x' = t_1 s_1 x, y' = t_2 s_2 y} M_0^{(t_1, t_2), (s_1, s_2)}$$

Verify that $M$ satisfies condition by writing

$$P(\sigma'_r = s, (X', Y') = (x', y')) = \sum_{s_1', s_2' = \pm} P(U = s_1' x', V = s_2' y') \times \cdots$$

$$\cdots P(\sigma_r = s, \sigma'_1 = s_1', \sigma'_2 = s_2')$$

$$= \sum_{s_1, s_2, s_1', s_2'} P(U = s_1' x', V = s_2' y') \times \cdots$$

$$\cdots M_0^{(s_1, s_2), (s_1', s_2')} P(\sigma_r = s, \sigma_1 = s_1, \sigma_2 = s_2)$$
Lemma (sub-additivity of mutual information)

Assume that $Y_1, \ldots, Y_m$ are independent conditionally on $X$. Then

$$I(X; Y^m_1) \leq \sum_{i=1}^m I(X; Y_i).$$

Proof: By conditional independence,

$$I(X; Y^m_1) = H(Y^m_1) - \sum_{i=1}^m H(Y_i|X).$$

By sub-additivity of entropy (which follows from non-negativity of entropy and of mutual information), $H(Y^m_1) \leq \sum_{i=1}^m H(Y_i)$, hence the result.

Corollary

For symmetric binary tree transmission, with arbitrary transmission parameters $\theta_{(p(i),i)} \in [-1,1]$ for all edges $(p(i), i)$,

$$I(\sigma_r; \sigma_{L_d}) \leq \sum_{j \in L_d} I(\sigma_r; \sigma_j).$$

Proof: by induction on number of edges in tree. If root degree $> 1$, use [sub-additivity] lemma. If root degree $= 1$, and degree of root’s child equals 1, concatenate top-two edges. If root degree $= 1$, and degree of root’s child $> 1$, use: i) “channel-between-trees” lemma, ii) Data Processing Inequality, then iii) sub-additivity lemma.
For each $i \in \mathcal{L}_d$, channel between $\sigma_r$ and $\sigma_i$: binary symmetric channel, with $\mathbb{E}(\sigma_r \sigma_i) = \theta^d = \lambda_2^d$.

Equivalently, $P(\sigma_i = \sigma_r) = \frac{1+\lambda_2^d}{2}$. Thus
\[
I(\sigma_r; \sigma_i) = \sum_{s,t=\pm} \frac{1+st\lambda_2^d}{2} \ln(1 + st\lambda_2^d)
\leq \sum_{s,t=\pm} \frac{1+st\lambda_2^d}{2} \cdot st\lambda_2^d
= \lambda_2^{2d}.
\]

By previous lemma, $I(\sigma_r; \sigma_{\mathcal{L}_d}) \leq |\mathcal{L}_d| \lambda_2^{2d}$.

Under hypotheses $|\mathcal{L}_d| \leq e^{d[\ln(\alpha) + o(1)]}$ and $\alpha(\lambda_2)^2 < 1$,
\[
I(\sigma_r; \sigma_{\mathcal{L}_d}) \leq e^{d[\ln(\alpha\lambda_2^2) + o(1)]} \xrightarrow{d \to \infty} 0.
\]
Tree reconstruction threshold, general case

\[ \hat{\nu}_{s,d} = \mathbb{P}(\sigma_r = s | G_d) \] determines \( I(\sigma_r; G_d) \).

**Notations:** For \( i \in V_d, L_{i,d} \): vertices in \( L_d \) that admit \( i \) as ancestor.
\[ G_{i,d} = \sigma(T_d, \sigma_{L_{i,d}}), \nu_{s,i,d} = \mathbb{P}(\sigma_i = s | G_{i,d}). \]
For node \( i \), \( C_i = \{ j : p(j) = i \} \) children of \( j \).

**Belief Propagation:**
Initialize for \( i \in L_d \) by \( \nu_{s,i,d} = \mathbb{I}_{s=\sigma_i} \);
Propagate towards \( r \), for \( i \in V_{d-1} \) by Equation
\[ \nu_{s,i,d} = \frac{1}{Z_{i,d}} \nu_s \prod_{j \in C(i)} \sum_{s_j \in [q]} \frac{\nu_{s_j}^{j,d}}{\nu_{s_j}} P_{ss_j}. \]

→ BP Equations admit \( \{ \nu_s \} \) as trivial fixed point.
Belief Propagation as an analysis tool

Let \( p_k := \mathbb{P}(Z = k) \) (e.g. \( e^{-\alpha_k^\alpha k} / k! \) for \( \text{Poi}_\alpha \) offspring)

\( M([q]) \): probability distributions on \([q]\)

\[ F_k : M([q])^k \rightarrow M([q]) \]

\((\eta_1, \ldots, \eta_k) \rightarrow \left\{ \frac{1}{Z_k(\eta_1^k)} \nu_s \prod_{j=1}^{k} \sum_{s_j \in [q]} \frac{\eta_j(s_j)}{\nu_{s_j}} P_{s_{s_j}} \right\}_{s \in [q]} \]

Let \( Q_{\tau,d} \): law on \( M([q]) \) of \( \{ \mathbb{P}(\sigma_r = s | \mathcal{G}_d) \}_{s \in [q]} \) conditionally on \( \sigma_r = \tau \).

Density Evolution Equation (conditional version): for \( \phi : M([q]) \rightarrow \mathbb{R} \),

\[
\int_{M([q])} \phi(\eta) Q_{\tau,d+1}(d\eta) = \sum_{k \geq 0} p_k \int_{M([q])^k} \phi(F_k(\eta_1, \ldots, \eta_k)) \cdots \cdots \prod_{\ell=1}^{k} \sum_{s_\ell \in [q]} P_{\tau s_\ell} Q_{s_\ell,d}(d\eta_\ell) 
\]
Let \( \hat{Q}_d \): unconditional law on \( M([q]) \) of \( \{ \mathbb{P}(\sigma_r = s | G_d) \}_{s \in [q]} \).

**Density Evolution Equation** (unconditional version): for \( \phi : M([q]) \to \mathbb{R} \),

\[
\int_{M([q])} \phi(\eta) \hat{Q}_{d+1}(d\eta) = \sum_{\tau \in [q]} \nu_{\tau} \sum_{k \geq 0} p_k \int_{M([q])^k} \phi(F_k(\eta_1, \ldots, \eta_k)) \cdots \cdot \prod_{\ell=1}^k \sum_{s_{\ell} \in [q]} P_{\tau s_{\ell}} \frac{\eta_{\ell}(s_{\ell})}{\nu_{s_{\ell}}} \hat{Q}_d(d\eta_{\ell})
\]

\( \rightarrow \) Formally, \( \hat{Q}_{d+1} = \Psi(\hat{Q}_d) \).

**Trivial fixed point for \( \Psi \):** Dirac mass \( \delta_{\{\nu_s\}_{s \in [q]}} \).

**Theorem (see lecture notes)**

*Tree reconstruction problem is feasible if and only if \( \Psi \) admits at least two fixed points (i.e., admits a non-trivial fixed point).*

Proof by Mézard-Montanari’06 for case \( \nu_s \equiv \frac{1}{q} \)
Remark

For \( b \)-ary trees, \( q \geq 4 \), and symmetric Potts model, reconstruction is feasible strictly below Kesten-Stigum threshold, i.e. for parameters such that \( b \times (\lambda_2)^2 < 1 \).

Hence census reconstructibility does not in general coincide with reconstructibility.

Remark

Density Evolution Equation an important tool in:
- Statistical Physics for several other problems (underlies so-called cavity method);
- Theory of Error Correcting Codes.
Community Detection for Sparse Stochastic Block Models

**Sparse SBM** $G(n, P, \alpha)$:
Let $P$: stochastic matrix on $[q]$, assumed irreducible and reversible for stationary measure $\nu$, i.e. $\nu_s P_{st} = \nu_t P_{ts}$.
Model: $n$ vertices, spins $\sigma_i$: i.i.d., $\sim \nu$.

$$
\mathbb{P}((i,j) \in E \mid \sigma[n]) = \frac{R_{\sigma_i \sigma_j}}{n} = \alpha \frac{P_{\sigma_i \sigma_j}}{\nu_{\sigma_j}} \frac{1}{n}
$$

where $R_{st} := \alpha \frac{P_{st}}{\nu_t}$ symmetric, by reversibility.

Average degrees:

$$
\mathbb{E}[\sum_{j \in [n]} \mathbb{1}(i,j) \in E \mid \sigma[n]] = \sum_{s \in [q]} \frac{R_{\sigma_i s}}{n} \sum_{j \neq i \in [n]} \mathbb{1}_{\sigma_j = s}
\approx \sum_{s \in [q]} \alpha \frac{P_{\sigma_i s}}{n \nu_s} \nu_s n
\approx \alpha,
$$
same irrespective of spin $\sigma_i$.

**Mean progeny matrix**: $M_{st} = \text{average number of spin } t\text{-neighbors of spin } s\text{-node}$. Then $M_{st} \approx \alpha P_{st}$. 
Definition

For estimates $\hat{\sigma}_i$ of spins $\sigma_i$ from observation of graph $G$, overlap:

$$\text{overlap}(\hat{\sigma}) = \max_{\pi \in S_q} \frac{1}{n} \sum_{i \in [n]} I_{\pi(\sigma_i) = \hat{\sigma}_i} - \sup_{s \in [q]} \nu_s.$$ 

Definition

Partial reconstruction is feasible (respectively, polynomial-time feasible) if

$$\exists \{\hat{\sigma}_i\} = f(G) \text{ (respectively, } = f(G) \text{ for polynomial-time computable function } f) \text{ such that for some } \epsilon > 0, \mathbb{P}(\text{overlap}(\hat{\sigma}) \geq \epsilon) \xrightarrow{n \to \infty} 1.$$ 

Remark

Zero overlap can always be achieved by $\hat{\sigma}_i \equiv 1$. In case $\nu \sim \mathcal{U}([q])$, zero overlap also achieved by taking $\hat{\sigma}_i$: i.i.d. uniform on $[q]$, independent of $G$. 
Definition

Weak partial reconstruction feasible if \( \exists \{ \hat{\sigma}_i \} = f(G) \) such that with high probability, \( \lim \inf \sum_{s,t \in [q]} p_n(s,t) \ln \left( \frac{p_n(s,t)}{\nu_s q_n(t)} \right) \geq \epsilon > 0 \), where

\[
p_n(s,t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}_{\sigma_i = s, \hat{\sigma}_i = t}, \quad q_n(t) = \sum_{s \in [q]} p_n(s,t).
\]

Remark

When \( \nu = \mathcal{U}([q]) \), weak partial reconstructibility is equivalent to partial reconstructibility ([Bordenave-Lelarge-Massoulié’18]).
Links between tree and community reconstruction

Let $B_G(i, d)$ denote the set of nodes in $G$ at graph distance at most $d$ from $i$. By abuse of notation, also denote by $B_G(i, d)$ the sub-graph of $G$ induced by $B_G(i, d)$.

Lemma (Local structure of $G(n, P, \alpha)$)

For $G \sim G(n, P, \alpha)$, $d \leq c \ln(n)$, where $c > 0$ is fixed sufficiently small, then for randomly chosen vertex $i \in [n]$,

$$d_{\text{var}} \left( \{B_G(i, d), \sigma_{B_G(i,d)}\}, \{\mathcal{T}_d, \sigma_{\mathcal{V}_d}\} \right) \xrightarrow{n \to \infty} 0,$$

where $\mathcal{T}_d = (V_d, E_d)$: Galton-Watson branching tree with offspring $\text{Poi}_\alpha$, and spin propagation mechanism driven by $P$.

Proof: coupling construction, using total variation bounds

$$d_{\text{var}}(\text{Poi}_\lambda, \text{Bin}(n, \lambda/n)) \leq 2\lambda/n, \quad d_{\text{var}}(\text{Poi}_\lambda, \text{Poi}_\mu) \leq 2|\lambda - \mu|.$$
Lemma (Mossel-Neeman-Sly’15)

For $i$ chosen uniformly at random in $[n]$, $d \leq c \ln(n)$, $U = B_G(i, d)$, $V = \{j \in [n] : d_G(i, j) = d + 1\}$, $W = [n] \setminus (U \cup V)$, then for all $s \in [q]$, $\epsilon > 0$,
\[
\lim_{n \to \infty} P\left( |P(\sigma_i = s | \sigma_{V \cup W}, G) - P(\sigma_i = s | \sigma_V, G|_{U \cup V})| \geq \epsilon \right) = 0.
\]

Together with local structure Lemma, implies

Corollary

If Tree reconstruction problem is not feasible, then weak partial community reconstruction is not feasible.

Proof: Tree reconstruction infeasible

$\Rightarrow P(\sigma_i = s | \sigma_{V \cup W}, G) \approx P(\sigma_r = s | T_d, \sigma_{L_d}) \approx \nu_s$.

Thus for uniform independent selection of $I, J \in [n]$,

$P(\sigma_I = s | G, \sigma_J = t) \rightarrow \nu_s$. Then for $\phi_i(G) = \mathbb{I}_{\hat{\sigma}_i = t}$,

$E[(p_n(s, t) - \nu_s q_n(t))^2] = E[(\frac{1}{n} \sum_{i=1}^{n} (\mathbb{I}_{\sigma_i = s} - \nu_s) \phi_i(G))^2]$

$= E[(\mathbb{I}_{\sigma_I = s} - \nu_s) \phi_I(G) (\mathbb{I}_{\sigma_J = s} - \nu_s) \phi_J(G)]$

$\rightarrow 0$ as $n \rightarrow \infty$. 
Then w.h.p., $p_n(s, t) = \nu_s q_n(t) + o(1)$

**Remark**

*One does not expect this sufficient condition for impossibility of weak community reconstruction to be sharp. Distinct threshold for impossibility of weak community reconstruction conjectured by statistical physicists, see notes.*
Failure of classical spectral methods for community reconstruction in sparse SBM

For Erdős-Rényi graph $G(n, \alpha/n)$, $D = c \frac{\ln(n)}{\ln(\ln(n))}$, let $Z_i = \mathbf{1}_i$: center of isolated star with $D$ branches.

Then $E(Z_i) = \left(\frac{n-1}{D}\right) \left(\frac{\alpha}{n}\right)^D (1 - \frac{\alpha}{n})^{(D+1)(n-1-D)+\binom{D}{2}} = e^{-c \ln(n)(1+o(1))}$, and $E(Z_i Z_j) = [E(Z_i)]^2 (1 + o(1))$, so that for $c < 1$, w.h.p. (by second moment method), there are isolated stars with $D$ branches in $G(n, \alpha/n)$.

$\Rightarrow$ Sparse E-R graphs have adjacency matrix with eigenvalues of order $\sqrt{D} \gg 1$.

Corresponding eigenvectors supported by $D + 1$ vertices of corresponding star, hence localized, and not reflecting global structure of graph.

Same holds for sparse SBM $G(n, P, \alpha)$. 
Spectral Redemption

BP equations for estimating node spins in SBM:
\[ \psi_{s \rightarrow j} \propto \nu_s \prod_{k \sim i, k \neq j} \sum_{s_k \in [q]} \psi_{s_k \rightarrow i} R_{ss_k}. \]

Conjecture (Decelle et al.'11): if \( \lambda_2(P)^2 \alpha > 1 \), i.e. above Kesten-Stigum threshold, BP initialized with random weights converges to limits \( \psi_{s \rightarrow j} \) such that positive overlap achieved by
\[ \hat{\sigma}_i = \arg \max \psi^i_s, \text{ where } \psi^i_s \propto \nu_s \prod_{j \sim i} \sum_{s_j \in [q]} \psi_{s_j \rightarrow i} R_{ss_j}. \]

Still open: analysis of BP on sparse graphs very challenging.

Linearization of BP equations around trivial fixed point \( \psi_{s \rightarrow j} = \nu_s \):
For \( \psi_{s \rightarrow j} = \nu_s (1 + \epsilon_{s \rightarrow j}) \), gives
\[ \epsilon_{s \rightarrow j} \leftarrow \sum_{k \sim i, k \neq j} \sum_{s_k \in [q]} \epsilon_{s_k \rightarrow i} P_{ss_k}, \text{ or equivalently for } \]
\[ \epsilon = \{ \epsilon_{s \rightarrow j} \} (i \rightarrow j) \in \bar{E}, s \in [q], \bar{E} : \text{edges of } G \text{ with orientation,} \]
\[ \epsilon \leftarrow (B^\top \otimes P) \epsilon \text{ where } B: \text{non-backtracking matrix of } G \]
Non-backtracking matrix

$B$: $2m \times 2m$ matrix where $m$: number of edges of $G$, defined as
$B_{i\rightarrow j, k\rightarrow \ell} = \mathbb{I}_{j=k} \mathbb{I}_{i\neq\ell}$.

Allows counting of non-backtracking paths in $G$: $(B^t)_{i\rightarrow j, k\rightarrow \ell} = \cdots \cdot \cdot \cdot |\{\text{NB paths with } t + 1 \text{ edges, started at } i \rightarrow j, \text{ ending at } k \rightarrow \ell\}|$.

Spectrum of $B$: $\lambda_1(B) \geq |\lambda_2(B)| \geq \cdots \geq |\lambda_{2m}(B)|$. 
Spectrum of NBM $B$ for sparse SBM $G \sim \mathcal{G}(n, P, \alpha)$

Mean progeny matrix $M = \alpha P$, spectrum:
\[ \lambda_1(M) = \alpha \geq |\lambda_2(M)| = \alpha |\lambda_2(P)| \geq \cdots \geq |\lambda_q(M)| = \alpha |\lambda_q(P)|. \]

Let $x_i \in \mathbb{R}^q$: eigenvector of $M$ associated with $\lambda_i(M)$.
For $e = u \rightarrow v \in \tilde{E}$, define $y_i(e) = x_i(\sigma_u)$.
For $\ell = c \ln(n)$, $c > 0$ fixed constant, let $z_i = B^{\ell} B^{\top} \ell y_i$.

**Theorem**

Let $r_0 = \sup\{i \in [q] : \lambda_i(M)^2 > \lambda_1(M)\}$.
(Note: $r_0 \geq 2 \iff \alpha \lambda_2(P)^2 > 1$, i.e. above Kesten-Stigum threshold).

Then $\forall i \in [r_0]$, eigenpair $(\lambda_i(B), \xi_i)$ of $B$ verifies:

- $\lambda_i(B) \xrightarrow{\text{proba.}} \lambda_i(M)$.
- $\exists x_i \in \mathbb{R}^q$: eigenvector of $M \leftrightarrow \lambda_i(M)$ such that for associated $z_i \in \mathbb{R}^{2m}$,
  \[ \lim_{n \to \infty} \frac{\langle z_i, \xi_i \rangle}{\|z_i\| \|\xi_i\|} = 1. \]

For $i > r_0$, $|\lambda_i(B)| \leq \sqrt{\lambda_1(M)} + o(1)$. 
Spectrum of NBM for $q = 2$, above Kesten-Stigum threshold

Circle of radius $\sqrt{\lambda_1(M)}$

$\lambda_2(B) \approx \lambda_2(M)$  $\lambda_1(B) \approx \lambda_1(M)$
Corollary

When above Kesten-Stigum threshold, from eigenvector $\xi_2$ of $B$, compute $\phi \in \mathbb{R}^n : \phi(u) = \sum_{v \sim u} \xi_2(v \rightarrow u)$, normalized so that $\|\phi\| = \sqrt{n}$.

Then in case where $\nu_s \equiv \frac{1}{q}$, positive overlap achieved by partitioning nodes $u \in [n]$ at random into $I^+, I^-$ by setting

$$\mathbb{P}(v \in I^+ | \phi) = \frac{1}{2} + \frac{1}{2K} \phi(v) \mathbb{I}_{|\phi(v)| \leq K},$$

where $K$: a constant chosen sufficiently large.

Thus, partial reconstruction is polynomial-time feasible when above Kesten-Stigum threshold.
[Krzakala et al.’13] conjecture “spectral redemption”, i.e. possibility to achieve positive overlap based on NBM matrix above KS threshold

[Bordenave-Lelarge-M.’16,18]: proofs of NBM spectral properties. Extensions in [Stephan-M.’19].

For $q \geq 4$, instances of $G(n, P, \alpha)$ below KS threshold, such that non-polynomial time methods can achieve positive overlap have been identified.

**Common belief:** for sparse SBM $G(n, P, \alpha)$, KS threshold is the boundary for polynomial-time community reconstruction.