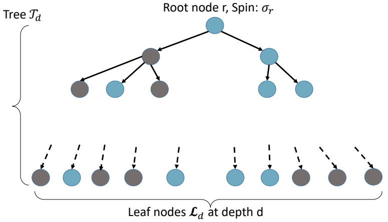
The tree reconstruction problem



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: $\mathbb{P}(\sigma_{\mathcal{L}_d} = s_{\mathcal{L}_d} | \mathcal{T}, \sigma_{V_{d-1}}) = \prod_{i \in \mathcal{L}_d} P_{\sigma_{p(i)} s_i}$ where P: stochastic matrix, assumed irreducible, with invariant distribution ν on $[q]$

The tree reconstruction problem

Assume root spin $\sigma_r \sim \nu$. Then $\mathbb{P}(\sigma_{V_d} = s_{V_d} | \mathcal{T}) = \nu_{s_r} \prod_{(i,j)\in E_d, i=p(j)} P_{s_i s_j}$

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

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* $\lim_{d\to\infty} I(\sigma_r; X_d) = 0$, *i.e. census reconstruction fails.*

Remark

Result still true for more general branching processes. It holds for instance with $Z \equiv \alpha \in \mathbb{N}^*$.

Proof Elements

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* $\sigma_r = \tau \in [q]$, $\{\alpha^{-d/2}(X_{s,d} - \alpha^d \nu_s)\}_{s \in [q]} \underset{d \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(m, \Sigma)$, *where* m, Σ *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 $t \in \{\tau, \tau'\}$, $\mathcal{L}(X_{d+1}^{(t)} \mid X_d^{(t)}) = \otimes_{s\in[q]}\mathrm{Poi}(M_s^{(t)})$, where $M_s^{(t)} = \alpha \sum_{s' \in [q]} X_{s',d}^{(t)} P_{s's}$.

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, $\exists \alpha_d \overset{d\to\infty}{\to} 0$ such that $\forall s \in [q],\ \mathbb{P}(\epsilon_s \le \alpha_d) \overset{d\to\infty}{\to} 1$.

Lemma

Variation distance $d_{var}(\mu, \nu) := 2\sup_A |\mu(A) - \nu(A)|$ *also equals* $2 \inf_{(X,Y) \text{ coupling of } (\mu,\nu)} \mathbb{P}(X \ne Y)$.

Corollary

$$d_{var}(\otimes_{s\in[q]}\mu^{(s)}, \otimes_{s\in[q]}\nu^{(s)}) \le \sum_{s\in[q]} d_{var}(\mu^{(s)}, \nu^{(s)}).$$

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

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

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

Thus: $\lim_{d\to\infty} \mathbb{E}(A_{s,>}) = 0$.

$$A_{s,\leq} \leq \sum_{k:|k-M_s|\leq\omega_d\sqrt{M_s}} e^{-M_s}\frac{M_s^k}{k!}\left|e^{-\epsilon_s\sqrt{M_s}}(1+\frac{\epsilon_s}{\sqrt{M_s}})^k - e^{\epsilon_s\sqrt{M_s}}(1-\frac{\epsilon_s}{\sqrt{M_s}})^k\right|$$

On the event $\{\epsilon_s \leq \alpha_d\}$, for $k : |k-M_s| \leq \omega_d\sqrt{M_s}$, one has:

$$\begin{aligned} e^{\pm\epsilon_s\sqrt{M_s}}(1\mp\tfrac{\epsilon_s}{\sqrt{M_s}})^k &= e^{\pm\epsilon_s\sqrt{M_s}+k(\mp\epsilon_s/\sqrt{M_s}+O(\epsilon_s^2/M_s))} = e^{O(\epsilon_s\omega_d)} \\ &= 1+O(\sqrt{\alpha_d}). \end{aligned}$$

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

By Jensen's inequality
$d_{var}(X_{d+1}^{(\tau)}, X_{d+1}^{(\tau')}) \leq \mathbb{E}[d_{var}(X_{d+1}^{(\tau)}, X_{d+1}^{(\tau')} \mid X_d^{(\tau)}, X_d^{(\tau')})]$

Thus $d_{var}(X_{d+1}^{(\tau)}, X_{d+1}^{(\tau')}) \leq \sum_{s\in[q]}\mathbb{E}(A_{s,>}+A_{s,\leq}) \overset{d\to\infty}{\to} 0$.

Theorem then follows from

Lemma

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

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

$$
\begin{aligned}
I(\sigma_r; X_d) &= \sum_{s\in[q],x\in\mathbb{N}^q} \nu_s \mathbb{P}(X_d = x|\sigma_r = s) \ln\left(\frac{\mathbb{P}(X_d=x|\sigma_r=s)}{\mathbb{P}(X_d=x)}\right) \\
&= \sum_{x\in\mathbb{N}^q} \mathbb{P}(X_d = x) \sum_{s\in[q]} \nu_s f_s(x) \ln(f_s(x)) \\
&\leq \sum_{x\in\mathbb{N}^q} \mathbb{P}(X_d = x) \sum_{s\in[q]} \nu_s f_s(x)[f_s(x) - 1] \\
&= \sum_{\tau\in[q]} \nu_\tau \sum_{x\in\mathbb{N}^q} \mathbb{P}(X_d = x) \sum_{s\in[q]} \nu_s f_s(x)[f_s(x) - f_\tau(x)] \\
&\leq \sum_{s,\tau\in[q]} \nu_\tau \sum_{x\in\mathbb{N}^q} \mathbb{P}(X_d = x)|f_s(x) - f_\tau(x)| \\
&= \sum_{s,\tau\in[q]} \nu_\tau d_{var}(\mathbb{P}(X_d \in \cdot|\sigma_r = s), \mathbb{P}(X_d \in \cdot|\sigma_r = \tau))
\end{aligned}
$$

Tree reconstruction threshold for symmetric case with $q = 2$

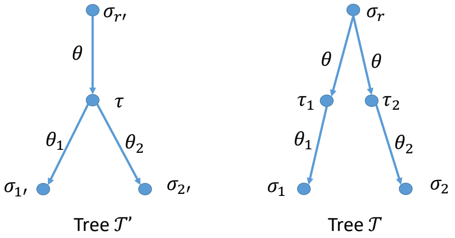
For $q = 2$, take $\sigma_i = \pm$. Symmetry: $P_{++} = P_{--} = 1 - \epsilon$, $P_{-+} = P_{+-} = \epsilon$.
Notation: let $\theta = \lambda_2(P) = 1 - 2\epsilon$, so that $\mathbb{E}(\sigma_i | \sigma_{p(i)}) = \theta \sigma_{p(i)}$, $\mathbb{E}(\sigma_i \sigma_{p(i)}) = \theta$.

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

For symmetric $q = 2$ propagation on deterministic tree $\mathcal{T}$ such that $\limsup \frac{1}{d} \ln(|\mathcal{L}_d|) \leq \ln(\alpha)$, tree reconstruction fails when $\alpha(\lambda_2)^2 < 1$.

Corollary

For symmetric $q = 2$ propagation on Galton-Watson tree $\mathcal{T}$, 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 $\mathcal{T}$, $\mathcal{T}'$ above, where node variables are binary spins, each uniformly distributed with values ± 1, edge weights $\in [0,1]$ represent transmission probability, e.g. $\mathbb{E}(\sigma_r \tau_1) = \theta$.
Then there exists a probability transition matrix $M^0 : \{-1,1\}^2 \to \{-1,1\}^2$ such that

$$\mathbb{P}(\sigma_{r'} = s_r, \sigma_{1'} = s_1, \sigma_{2'} = s_2) = \sum_{u_1,u_2=\pm} \mathbb{P}(\sigma_r = s_r, \sigma_1 = u_1, \sigma_2 = u_2) \times \cdots$$
$$\cdots M^0_{(u_1,u_2),(s_1,s_2)}$$

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

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

$$\begin{aligned}
\mathbb{P}(\sigma'_r = s, (X', Y') = (x', y')) &= \sum_{s'_1, s'_2=\pm} \mathbb{P}(U = s'_1 x', V = s'_2 y') \times \cdots \\
&\cdots \mathbb{P}(\sigma_r = s, \sigma'_1 = s'_1, \sigma'_2 = s'_2) \\
&= \sum_{s_1,s_2,s'_1,s'_2} \mathbb{P}(U = s'_1 x', V = s'_2 y') \times \cdots \\
&\cdots M^0_{(s_1,s_2),(s'_1,s'_2)} \mathbb{P}(\sigma_r = s, \sigma_1 = s_1, \sigma_2 = s_2)
\end{aligned}$$

Lemma (sub-additivity of mutual information)

Assume that $Y_1, \ldots, Y_m$ are independent conditionally on X. Then $I(X; Y_1^m) \leq \sum_{i=1}^m I(X; Y_i)$.

Proof: By conditional independence,
$I(X; Y_1^m) = H(Y_1^m) - \sum_{i=1}^m H(Y_i|X)$.
By sub-additivity of entropy (which follows from non-negativity of entropy and of mutual information), $H(Y_1^m) \leq \sum_{i=1}^m H(Y_i)$, hence the result.

Corollary

For symmetric binary tree transmission, with arbitrary transmission parameters $\theta_{(p(i),i)} \in [-1, 1]$ for all edges $(p(i), i)$, $I(\sigma_r; \sigma_{\mathcal{L}_d}) \leq \sum_{j \in \mathcal{L}_d} I(\sigma_r; \sigma_j)$.

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.

End of proof

For each $i \in \mathcal{L}_d$, channel between σ_r and σ_i: binary symmetric channel, with $\mathbb{E}(\sigma_r \sigma_i) = \theta^d = \lambda_2^d$.

Equivalently, $\mathbb{P}(\sigma_i = \sigma_r) = \frac{1+\lambda_2^d}{2}$. Thus

$$\begin{aligned} I(\sigma_r;\sigma_i) &= \sum_{s,t=\pm} \frac{1}{2}\frac{1+st\lambda_2^d}{2}\ln(1+st\lambda_2^d) \\ &\leq \sum_{s,t=\pm} \frac{1}{2}\frac{1+st\lambda_2^d}{2} st\lambda_2^d \\ &= \lambda_2^{2d}. \end{aligned}$$

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 $\alpha(\lambda_2)^2 < 1$,

$I(\sigma_r;\sigma_{\mathcal{L}_d}) \leq e^{d[\ln(\alpha\lambda_2^2)+o(1)]} \overset{d\to\infty}{\to} 0.$

Tree reconstruction threshold, general case

$\hat{\nu}_{s,d} = \mathbb{P}(\sigma_r = s | \mathcal{G}_d)$ determines $I(\sigma_r; \mathcal{G}_d)$.

Notations: For $i \in V_d$, $\mathcal{L}_{i,d}$: vertices in $\mathcal{L}_d$ that admit i as ancestor.
$\mathcal{G}_{i,d} = \sigma(\mathcal{T}_d, \sigma_{\mathcal{L}_{i,d}})$, $\nu_s^{i,d} = \mathbb{P}(\sigma_i = s | \mathcal{G}_{i,d})$.
For node i, $C_i = \{j : p(j) = i\}$ children of j.

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

$\rightarrow$ BP Equations admit $\{\nu_s\}$ as trivial fixed point.

Belief Propagation as an analysis tool

Let $p_k := \mathbb{P}(Z = k)$ (e.g. $e^{-\alpha}\alpha^k/k!$ for Poi_α offspring)
$M([q])$: probability distributions on $[q]$

$F_k : M([q])^k \to M([q])$
$(\eta_1, \dots, \eta_k) \to \left\{ \frac{1}{Z_k(\eta_1^k)} \nu_s \prod_{j=1}^k \sum_{s_j \in [q]} \frac{\eta_j(s_j)}{\nu_{s_j}} P_{ss_j} \right\}_{s \in [q]}$

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 $\phi : M([q]) \to \mathbb{R}$,

$$\int_{M([q])} \phi(\eta) Q_{\tau,d+1}(d\eta) = \sum_{k \geq 0} p_k \int_{M([q])^k} \phi(F_k(\eta_1, \dots, \eta_k)) \cdots$$
$$\cdots \prod_{\ell=1}^{k} \sum_{s_\ell \in [q]} P_{\tau s_\ell} Q_{s_\ell, d}(d\eta_\ell)$$

Let $\hat{Q}_d$: unconditional law on $M([q])$ of $\{\mathbb{P}(\sigma_r = s|\mathcal{G}_d)\}_{s\in[q]}$.

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

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

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

Trivial fixed point for Ψ: Dirac mass $\delta_{\{\nu_s\}_{s\in[q]}}$.

Theorem (see lecture notes)

Tree reconstruction problem is feasible if and only if Ψ 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 $\mathcal{G}(n, P, \alpha)$:
Let P: stochastic matrix on $[q]$, assumed irreducible and reversible for stationary measure ν, i.e. $\nu_s P_{st} = \nu_t P_{ts}$.
Model: n vertices, spins σ_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:

$$\begin{aligned} \mathbb{E}[\textstyle\sum_{j \in [n]} \mathbb{I}_{(i,j) \in E} | \sigma_{[n]}] &= \textstyle\sum_{s \in [q]} \frac{R_{\sigma_i s}}{n} \sum_{j \neq i \in [n]} \mathbb{I}_{\sigma_j = s} \\ &\approx \textstyle\sum_{s \in [q]} \alpha \frac{P_{\sigma_j s}}{n \nu_s} \nu_s n \\ &\approx \alpha, \end{aligned}$$

same irrespective of spin σ_i.

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

Definition

For estimates $\hat{\sigma}_i$ of spins σ_i from observation of graph G, **overlap**:
$\text{overlap}(\hat{\sigma}) = \max_{\pi \in \mathcal{S}_q} \frac{1}{n} \sum_{i \in [n]} \mathbb{I}_{\pi(\sigma_i) = \hat{\sigma}_i} - \sup_{s \in [q]} \nu_s$.

Definition

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

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, $\liminf \sum_{s,t \in [q]} p_n(s,t) \ln \left(\frac{p_n(s,t)}{\nu_s q_n(t)} \right) \geq \epsilon > 0$, where $p_n(s,t) = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}_{\sigma_i = s, \hat{\sigma}_i = t}$, $q_n(t) = \sum_{s \in [q]} p_n(s,t)$.

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 $\mathcal{B}_G(i,d)$ denote the set of nodes in G at graph distance at most d from i. By abuse of notation, also denote by $\mathcal{B}_G(i,d)$ the sub-graph of G induced by $\mathcal{B}_G(i,d)$.

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

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

$d_{var}\left(\{\mathcal{B}_G(i,d), \sigma_{\mathcal{B}_G(i,d)}\}, \{\mathcal{T}_d, \sigma_{V_d}\}\right) \overset{n\to\infty}{\to} 0\,,$

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

Proof: coupling construction, using total variation bounds
$d_{var}(\mathrm{Poi}_\lambda, \mathrm{Bin}(n, \lambda/n)) \le 2\lambda/n$, $d_{var}(\mathrm{Poi}_\lambda, \mathrm{Poi}_\mu) \le 2|\lambda - \mu|$.

Lemma (Mossel-Neeman-Sly'15)

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

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
$\Rightarrow \mathbb{P}(\sigma_i = s|\sigma_{V\cup W}, G) \approx \mathbb{P}(\sigma_r = s|\mathcal{T}_d, \sigma_{\mathcal{L}_d}) \approx \nu_s$.
Thus for uniform independent selection of $I, J \in [n]$,
$\mathbb{P}(\sigma_I = s|G, \sigma_J = t) \to \nu_s$. Then for $\phi_i(G) = \mathbb{I}_{\hat{\sigma}_i = t}$,

$$\begin{aligned}\mathbb{E}[(p_n(s,t) - \nu_s q_n(t))^2] &= \mathbb{E}[\left(\tfrac{1}{n}\sum_{i=1}^n (\mathbb{I}_{\sigma_i = s} - \nu_s)\phi_i(G)\right)^2] \\ &= \mathbb{E}[(\mathbb{I}_{\sigma_I = s} - \nu_s)\phi_I(G)(\mathbb{I}_{\sigma_J = s} - \nu_s)\phi_J(G)] \\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

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 $\mathcal{G}(n, \alpha/n)$, $D = c\frac{\ln(n)}{\ln(\ln(n))}$, let $Z_i = \mathbb{I}_{i:\text{ center of isolated star with } D \text{ branches}}$.

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

$\Rightarrow$ 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 $\mathcal{G}(n, P, \alpha)$.

Spectral Redemption

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

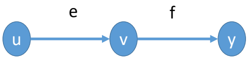
Conjecture (Decelle et al.'11): if $\lambda_2(P)^2\alpha > 1$, i.e. above Kesten-Stigum threshold, BP initialized with random weights converges to limits $\psi_s^{i\to j}$ such that positive overlap achieved by
$\hat{\sigma}_i = \arg\max \psi_s^i$, where $\psi_s^i \propto \nu_s \prod_{j\sim i} \sum_{s_j\in[q]}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

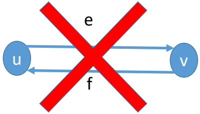
Still open: analysis of BP on sparse graphs very challenging.

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

Non-backtracking matrix

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





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

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 $e = u \to v \in \vec{E}$, 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

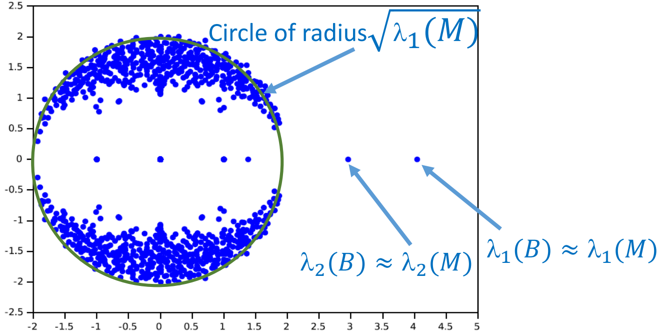
Let $r_0 = \sup\{i \in [q] : \lambda_i(M)^2 > \lambda_1(M)\}$.
(Note: $r_0 \geq 2 \Leftrightarrow \alpha\lambda_2(P)^2 > 1$*, i.e. above Kesten-Stigum threshold).*
Then $\forall i \in [r_0]$*, eigenpair* $(\lambda_i(B), \xi_i)$ *of* B *verifies:*

$\lambda_i(B) \underset{n\to\infty}{\overset{proba.}{\to}} \lambda_i(M)$.

$\exists x_i \in \mathbb{R}^q$*: 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 fo $q = 2$, above Kesten-Stigum threshold



Corollary

When above Kesten-Stigum threshold, from eigenvector ξ_2 *of* B*, compute* $\phi \in \mathbb{R}^n : \phi(u) = \sum_{v \sim u} \xi_2(v \to u)$*, 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.

References:

[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 $\mathcal{G}(n, P, \alpha)$ below KS threshold, such that non-polynomial time methods can achieve positive overlap have been identified.

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