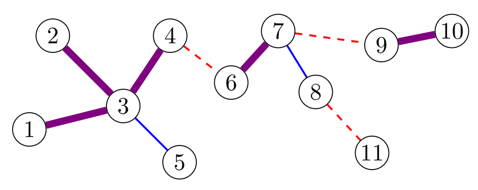
Correlation detection in trees for planted graph alignment

L. Ganassali, M. Lelarge, L. Massoulié
INRIA, DI/ENS, PSL Research University, Paris, France.

Planted Erdős-Rényi graph alignment

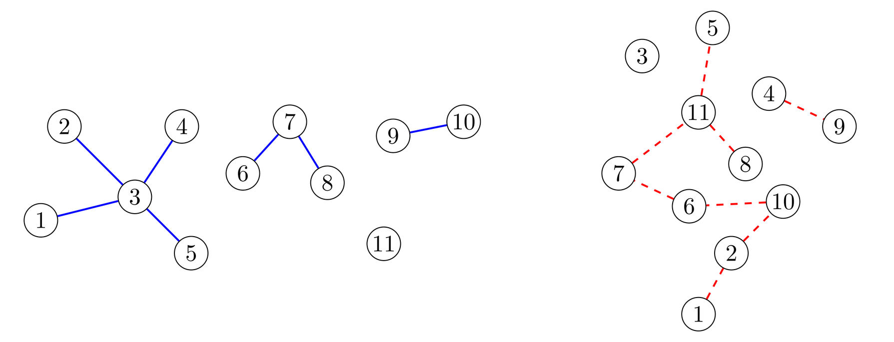
(1) Draw $\mathcal{G}, \mathcal{G}'$ two graphs with same node set $[n]$ s.t. for all $\{i,j\} \in \binom{[n]}{2}$,

$$\left(\mathbf{1}_{i \underset{\mathcal{G}}{\sim} j}, \mathbf{1}_{i \underset{\mathcal{G}'}{\sim} j}\right) = \begin{cases} (1,1) & \text{w.p. } qs \\ (1,0),(0,1) & \text{w.p. } q(1-s) \\ (0,0) & \text{w.p. } 1-q(2-s). \end{cases}$$



Sparse setting: $q = \boldsymbol{\lambda}/n$, correlation parameter $s \in (0,1)$.

(2) Relabel the vertices of $\mathcal{G}'$ with a uniform planted permutation π^*: $\mathcal{H} := \mathcal{G}' \circ \pi^*$.



Goal: Upon observing $\mathcal{G}$ and $\mathcal{H}$, find an estimator $\hat{\pi}$ that recovers π^* w.h.p., that is such that $\mathrm{ov}(\hat{\pi}, \pi^*) := \sum_{1 \le i \le n} \mathbf{1}_{\hat{\pi}(i) = \pi^*(i)}$ is $\ge \alpha n$ for some $\alpha > 0$.

Local point of view: detecting correlation in trees

For $i \in V(\mathcal{G}), u \in V(\mathcal{H})$, look at the neighborhoods $\mathcal{N}_i$ and $\mathcal{N}_u$ at depth d:

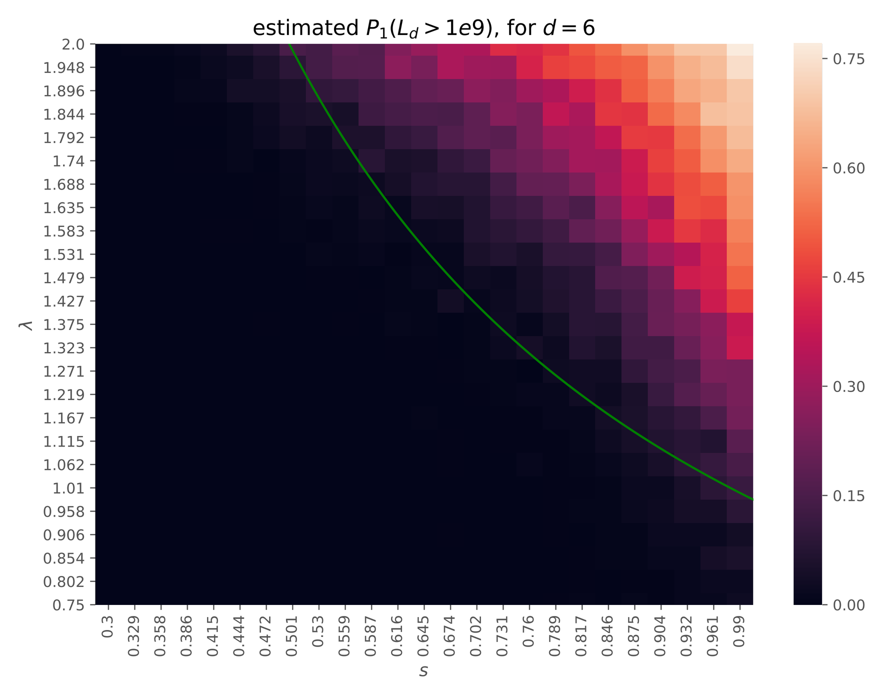
- if $u = \pi^*(i)$, $(\mathcal{N}_i, \mathcal{N}_u) \simeq$ GW trees of offspring $\mathrm{Poi}(\lambda)$, with intersection of offspring $\mathrm{Poi}(\lambda s)$ (model $\mathbb{P}_{1,d}$);
- if $u \neq \pi^*(i)$, $(\mathcal{N}_i, \mathcal{N}_u) \simeq$ independent GW trees of offspring $\mathrm{Poi}(\lambda)$ (model $\mathbb{P}_{0,d}$).

Hypothesis testing: Can we test $\mathbb{P}_{1,d}$ versus $\mathbb{P}_{0,d}$? For two trees of depth d, the likelihood ratio $L_d(t,t') := \frac{\mathbb{P}_{1,d}(t,t')}{\mathbb{P}_{0,d}(t,t')}$ verifies

$$L_d(t,t') = \sum_{k=0}^{c \wedge c'} \psi(k,c,c') \sum_{\substack{\sigma \in \mathcal{S}(k,c) \\ \sigma' \in \mathcal{S}(k,c')}} \prod_{i=1}^{k} L_{d-1}(t_{\sigma(i)}, t'_{\sigma'(i)}),$$

where $\psi(k,c,c') = e^{\lambda s} \times \frac{s^k \bar{s}^{c+c'-2k}}{\lambda^k k!}$, and $\mathcal{S}(k,\ell)$ denotes the set of injective mappings from $[k]$ to $[\ell]$.

One-sided tests: tests $\mathcal{T}_d : \mathcal{X}_d \times \mathcal{X}_d \to \{0,1\}$ such that $\mathbb{P}_{0,d}(\mathcal{T}_d = 0) = 1 - o(1)$ and $\liminf_d \mathbb{P}_{1,d}(\mathcal{T}_d = 1) > 0$ (i.e. vanishing type I error and non vanishing power).



Estimated $\mathbb{P}_{1,d}(L_d > 10^9)$ for $d = 6$ (green curve: $\lambda s = 1$).

General result for hypothesis testing in trees

Theorem 1 (Correlation detection in trees). *Let* $\mathrm{KL}_d := \mathrm{KL}(\mathbb{P}_{1,d} \| \mathbb{P}_{0,d}) = \mathbb{E}_{1,d}[\log(\mathbf{L}_d)]$ *. Then the following propositions are equivalent:*

(i) There exists a one-sided test for deciding $\mathbb{P}_{0,d}$ versus $\mathbb{P}_{1,d}$,

(ii) $\lim_{d \to \infty} \mathrm{KL}_d = +\infty$ and $\lambda s > 1$,

(iii) There exists $(a_d)_d$ such that $a_d \to \infty$, $\mathbb{P}_{0,d}(\mathbf{L}_d > a_d) \to 0$ and $\liminf_d \mathbb{P}_{1,d}(\mathbf{L}_d > a_d) > 0$.

(iv) Denoting $\mathbb{P}_0 := \mathbb{P}_{0,\infty}$, the martingale $(\mathbf{L}_d)_d$ (w.r.t. to $\mathbb{P}_0$) is not uniformly integrable.

(v) with probability[a] $1 - p_{\mathrm{ext}}(\lambda s) > 0$, $\mathbf{L}_d$ diverges to $+\infty$ with rate $\Omega\left(\exp\left(\Omega(1) \times (\lambda s)^d\right)\right)$.

[a] probability that a Galton-Watson tree of offspring $\mathrm{Poi}(\lambda s)$ survives.

A message-passing algorithm

Message passing:

$$m^{t+1}_{i \to j, u \to v} = \sum_{k=0}^{d_i \wedge d_u - 1} k! \psi(k, d_i - 1, d_u - 1) \sum_{\substack{\{\ell_1, \ldots \ell_k\} \in \partial i \setminus j \\ \{w_1, \ldots w_k\} \in \partial u \setminus v}} \sum_{\sigma \in \mathcal{S}_k} \prod_{a=1}^{k} m^t_{\ell_a \to i, w_{\sigma(a)} \to v}. \quad (1)$$

Aggregation:

$$m^t_{i,u} = \sum_{k=0}^{d_i \wedge d_u} k! \psi(k, d_i, d_u) \sum_{\substack{\{\ell_1, \ldots \ell_k\} \in \partial i \\ \{w_1, \ldots w_k\} \in \partial u}} \sum_{\sigma \in \mathcal{S}_k} \prod_{a=1}^{k} m^t_{\ell_a \to i, w_{\sigma(a)} \to v}. \quad (2)$$

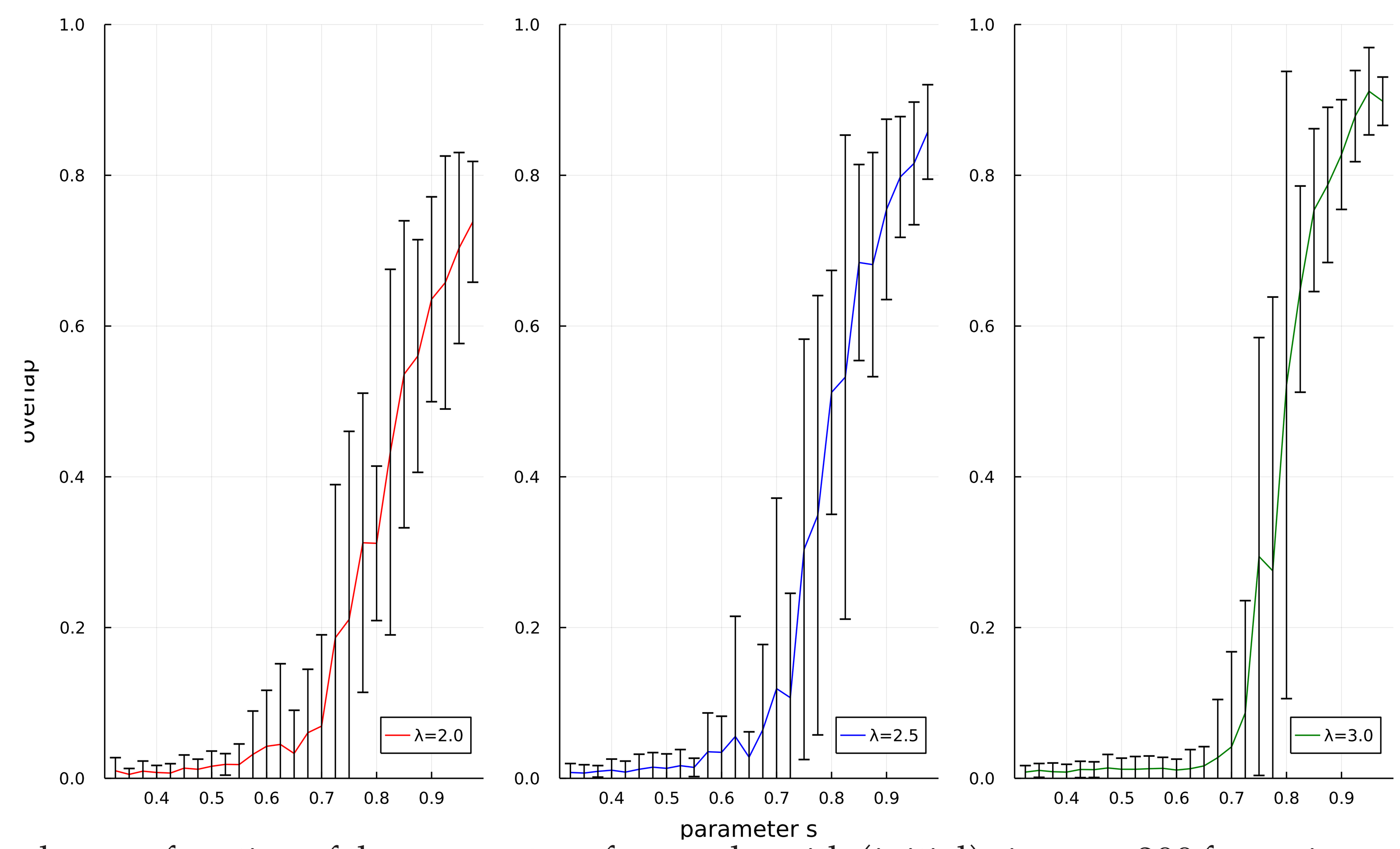
Edge score:

$$\begin{aligned} e(t) &:= \text{match-edges}(G, G', \pi^t, \sigma^t) \\ &:= \frac{1}{|E|} \sum_{(i,j) \in E} \mathbf{1}_{(\pi^t(i), \pi^t(j)) \in E'} + \frac{1}{|E'|} \sum_{(u,v) \in E'} \mathbf{1}_{(\sigma^t(u), \sigma^t(v)) \in E}. \end{aligned} \quad (3)$$

Algorithm 1: BPAlign

Input: Two connected graphs $G = (V,E)$ and $G' = (V',E')$, parameter d and parameters of the correlated Erdős-Rényi model λ (average degree) and s

for $t \in \{1, \ldots, d\}$ **do**
  compute $m^t_{i \to j, u \to v}$ thanks to (1)
  compute $m^t_{i,u}$ thanks to (2)
  compute $\pi^t : V \to V'$ as $\pi^t(i) = \arg\max(m^t_{i,\cdot})$
  compute $\sigma^t : V' \to V$ as $\sigma^t(u) = \arg\max(m^t_{\cdot,u})$
  compute $e(t) = \text{match-edges}(G, G', \pi^t, \sigma^t)$ thanks to (3)
end
$t^* = \arg\max(e(t))$
Return $\pi^{t^*}, \sigma^{t^*}, m^{t^*}$



Overlap as a function of the parameter s for graphs with (initial) size $n = 200$ for various values of λ (parameter $d = 15$). Each point is the average of 10 simulations.

Phase diagram

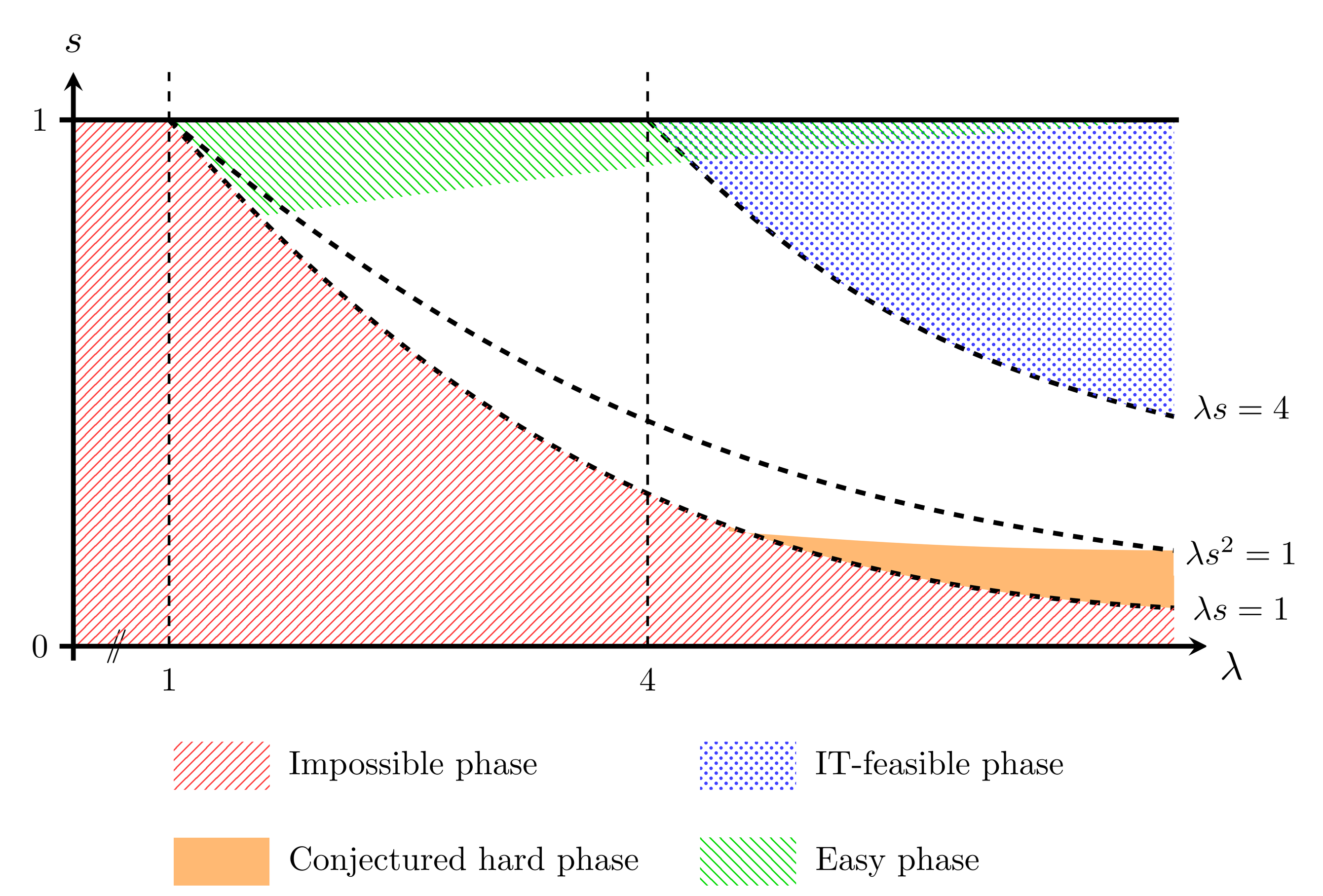


Diagram of the (λ, s) regions where partial recovery is known to be IT-impossible ([GML21b]), IT-feasible ([WXY21]), or easy ([GM20, GML21a]). In the orange region one-sided detectability is impossible in the tree correlation detection problem, and partial graph alignment is conjectured to be hard ([GML21a]).

References

[GM20] Luca Ganassali and Laurent Massoulié. From tree matching to sparse graph alignment. volume 125 of *Proceedings of Machine Learning Research*, pages 1633–1665. PMLR, 09–12 Jul 2020.

[GML21a] Luca Ganassali, Laurent Massoulié, and Marc Lelarge. Correlation detection in trees for partial graph alignment, 2021.

[GML21b] Luca Ganassali, Laurent Massoulie, and Marc Lelarge. Impossibility of partial recovery in the graph alignment problem. volume 134 of *Proceedings of Machine Learning Research*, pages 2080–2102. PMLR, 15–19 Aug 2021.

[WXY21] Yihong Wu, Jiaming Xu, and Sophie H. Yu. Settling the sharp reconstruction thresholds of random graph matching. *ArXiv*, abs/2102.00082, 2021.