Statistical and Computational limits for sparse graph alignment

Luca Ganassali

Junior Conference on Random networks and interacting particle systems

Joint work with Laurent Massoulié and Marc Lelarge

INRIA, DI/ENS, PSL Research University, Paris, France

Graph alignment

Question: Given two graphs $G = (V, E)$ and $G' = (V', E')$ with $|V| = |V'|$, *what is the best way to match nodes of* G *with nodes of* G'*?*

Question: Given two graphs $G = (V, E)$ and $G' = (V', E')$ with $|V| = |V'|$, *what is the best way to match nodes of G with nodes of G'?*

Minimizing disagreements: Find a bijection $f : V \to V'$ that minimizes

$$\sum_{(i,j)\in V^2} \left(\mathbf{1}_{(i,j)\in E} - \mathbf{1}_{(f(i),f(j))\in E'}\right)^2 ,$$

or, equivalently solve

$$\max_{\Pi} \mathrm{Tr}\left(G\Pi G'\Pi^\top\right) ,$$

where Π runs over all permutation matrices.

Graph alignment

Question: Given two graphs $G = (V, E)$ and $G' = (V', E')$ with $|V| = |V'|$, *what is the best way to match nodes of G with nodes of G'?*

Minimizing disagreements: Find a bijection $f : V \to V'$ that minimizes

$$\sum_{(i,j)\in V^2} \left(\mathbf{1}_{(i,j)\in E} - \mathbf{1}_{(f(i),f(j))\in E'}\right)^2 ,$$

or, equivalently solve

$$\max_{\Pi} \operatorname{Tr}\left(G\Pi G'\Pi^\top\right) ,$$

where Π runs over all permutation matrices. ⟵ *NP-hard in the worst case*

Planted graph alignment

Planted setting: we observe some random graphs $\mathcal{G}, \mathcal{H}$ (*n* nodes) correlated with a *planted* alignment π^* (hidden). We want to 'recover' π^*, with high probability when $n \to \infty$.

Planted graph alignment

Planted setting: we observe some random graphs $\mathcal{G}, \mathcal{H}$ (n nodes) correlated with a *planted* alignment π^* (hidden). We want to 'recover' π^*, with high probability when $n \to \infty$.

Another measure of performance: for any $[n]$-valued estimator $\hat{\pi}(\mathcal{G}, \mathcal{H})$, define its *overlap* with the planted permutation π^*

$$\mathrm{ov}(\hat{\pi}, \pi^*) := \sum_{i=1}^{n} \mathbf{1}_{\hat{\pi}(i)=\pi^*(i)}.$$

Planted graph alignment

Planted setting: we observe some random graphs $\mathcal{G}, \mathcal{H}$ (n nodes) correlated with a *planted* alignment π^* (hidden). We want to 'recover' π^*, with high probability when $n \to \infty$.

Another measure of performance: for any $[n]$-valued estimator $\hat{\pi}(\mathcal{G}, \mathcal{H})$, define its *overlap* with the planted permutation π^*

$$\mathrm{ov}(\hat{\pi}, \pi^*) := \sum_{i=1}^{n} \mathbf{1}_{\hat{\pi}(i)=\pi^*(i)}.$$

Definitions We say that $\hat{\pi}$ achieves:

- *Exact recovery* if
$$\mathbb{P}(\hat{\pi} = \pi^*) \to 1.$$
- *Partial recovery* if
$$\mathbb{P}(\mathrm{ov}(\hat{\pi}, \pi^*) > \alpha n) \to 1.$$

Planted graph alignment

Planted setting: we observe some random graphs $\mathcal{G}, \mathcal{H}$ (n nodes) correlated with a *planted* alignment π^* (hidden). We want to 'recover' π^*, with high probability when $n \to \infty$.

Another measure of performance: for any $[n]$-valued estimator $\hat{\pi}(\mathcal{G}, \mathcal{H})$, define its *overlap* with the planted permutation π^*

$$\mathrm{ov}(\hat{\pi}, \pi^*) := \sum_{i=1}^{n} \mathbf{1}_{\hat{\pi}(i)=\pi^*(i)}.$$

Definitions We say that $\hat{\pi}$ achieves:

- *Exact recovery* if
$$\mathbb{P}(\hat{\pi} = \pi^*) \to 1.$$
- *Partial recovery* if
$$\mathbb{P}(\mathrm{ov}(\hat{\pi}, \pi^*) > \alpha n) \to 1.$$

Remark: $\arg\max_{\Pi} \mathrm{Tr}\left(G\Pi G'\Pi^\top\right)$ does not coincide with π^* in general.

Planted graph alignment

Planted setting: we observe some random graphs $\mathcal{G}, \mathcal{H}$ (n nodes) correlated with a *planted* alignment π^* (hidden). We want to 'recover' π^*, with high probability when $n \to \infty$.

Another measure of performance: for any $[n]$-valued estimator $\hat{\pi}(\mathcal{G}, \mathcal{H})$, define its *overlap* with the planted permutation π^*

$$\mathrm{ov}(\hat{\pi}, \pi^*) := \sum_{i=1}^{n} \mathbf{1}_{\hat{\pi}(i)=\pi^*(i)}.$$

Definitions We say that $\hat{\pi}$ achieves:

- *Exact recovery* if
$$\mathbb{P}(\hat{\pi} = \pi^*) \to 1.$$
- *Partial recovery* if
$$\mathbb{P}(\mathrm{ov}(\hat{\pi}, \pi^*) > \alpha n) \to 1.$$

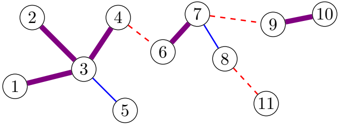
Remark: $\arg\max_{\Pi} \mathrm{Tr}\left(G \Pi G' \Pi^\top\right)$ does not coincide with π^* in general.

Some applications: de-anonymization of networks, protein classification in biology, image processing...

Correlated Erdős-Rényi model $\mathcal{G}(n, q, s)$:

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

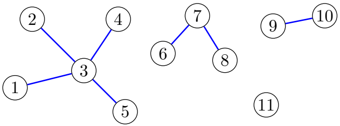
$$\left(\mathbf{1}_{i \underset{\mathcal{G}}{\sim} j}, \mathbf{1}_{i \underset{\mathcal{G}'}{\sim} j}\right) = \begin{cases} (1,1) & \text{w.p. } qs & \textit{two}-\textit{coloured} \text{ edge} \\ (0,1),(1,0) & \text{w.p. } q(1-s) & \textit{red} \text{ or } \textit{blue} \text{ edge} \\ (0,0) & \text{w.p.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

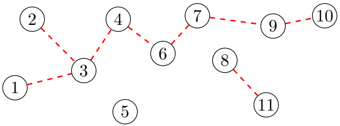


Correlated Erdős-Rényi model $\mathcal{G}(n, q, s)$:

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

$$\left(\mathbf{1}_{i \underset{\mathcal{G}}{\sim} j}, \mathbf{1}_{i \underset{\mathcal{G}'}{\sim} j}\right) = \begin{cases} (1,1) & \text{w.p. } qs & \textit{two–coloured} \text{ edge} \\ (0,1),(1,0) & \text{w.p. } q(1-s) & \textit{red} \text{ or } \textit{blue} \text{ edge} \\ (0,0) & \text{w.p. } 1-q(2-s) & \text{non-edge} \end{cases}$$



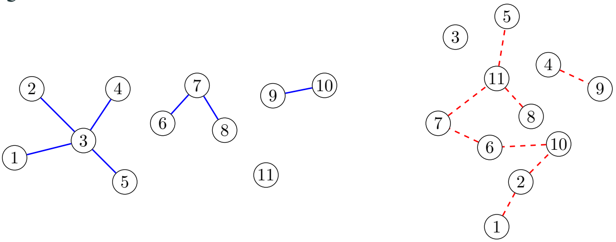


Correlated Erdős-Rényi model $\mathcal{G}(n, q, s)$:

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

$$\left(\mathbf{1}_{i \underset{\mathcal{G}}{\sim} j}, \mathbf{1}_{i \underset{\mathcal{G}'}{\sim} j}\right) = \begin{cases} (1,1) & \text{w.p. } qs & \textit{two–coloured} \text{ edge} \\ (0,1),(1,0) & \text{w.p. } q(1-s) & \textit{red} \text{ or } \textit{blue} \text{ edge} \\ (0,0) & \text{w.p. } 1-q(2-s) & \text{non-edge} \end{cases}$$

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



Planted graph alignment: general questions, state-of-the art

Goal: upon observing $\mathcal{G}$ and $\mathcal{H}$, estimate π^* with high probability.

Planted graph alignment: general questions, state-of-the art

Goal: upon observing $\mathcal{G}$ and $\mathcal{H}$, estimate π^* with high probability.

Sparse regime: $q = \lambda/n$, constant mean degree λ. Even with $s = 1$, $\Theta(n)$ isolated vertices $\longleftarrow$ *only partial alignment may be reachable* [Cullina-Kiyavash '16, '17].

Planted graph alignment: general questions, state-of-the art

Goal: upon observing $\mathcal{G}$ and $\mathcal{H}$, estimate π^* with high probability.

Sparse regime: $q = \lambda/n$, constant mean degree λ. Even with $s = 1$, $\Theta(n)$ isolated vertices $\longleftarrow$ *only partial alignment may be reachable* [Cullina-Kiyavash '16, '17].

Questions:

- Can we hope for some $\hat{\pi}$ s.t. $\mathrm{ov}(\hat{\pi}, \pi^*) > \alpha n$ w.h.p. with no computational restrictions (i.e. when is there enough signal)?
- What is the maximal fraction α?
- Can we find efficient (polynomial-time) algorithms for this task?

Planted graph alignment: general questions, state-of-the art

Goal: upon observing $\mathcal{G}$ and $\mathcal{H}$, estimate π^* with high probability.

Sparse regime: $q = \lambda/n$, constant mean degree λ. Even with $s = 1$, $\Theta(n)$ isolated vertices $\longleftarrow$ *only partial alignment may be reachable* [Cullina-Kiyavash '16, '17].

Questions:

- Can we hope for some $\hat{\pi}$ s.t. $\mathrm{ov}(\hat{\pi}, \pi^*) > \alpha n$ w.h.p. with no computational restrictions (i.e. when is there enough signal)?
- What is the maximal fraction α?
- Can we find efficient (polynomial-time) algorithms for this task?

State-of-the art: in the sparse regime where $\lambda > 0$ and $s \in [0, 1]$ are fixed constants: partial recovery is IT-feasible if $\lambda s > 4 + \varepsilon$ [Wu-Xu-Yu '21].

An impossibility result: upper bound on reachable overlap

Theorem
For $\lambda > 0$ and $s \in [0, 1]$, we have for any $\alpha > 0$, for any estimator $\hat{\pi}$:

$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^*) > (c(\lambda s) + \alpha)n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

where $c(\mu)$ is the greatest non-negative solution to the equation $e^{-\mu x} = 1 - x$.

An impossibility result: upper bound on reachable overlap

Theorem
For $\lambda > 0$ and $s \in [0, 1]$, we have for any $\alpha > 0$, for any estimator $\hat{\pi}$:

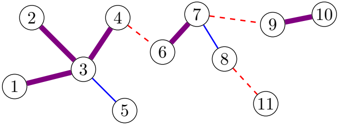
$$\mathbb{P}\left(\mathrm{ov}(\hat{\pi}, \pi^*) > (c(\lambda s) + \alpha)n\right) \underset{n \to \infty}{\longrightarrow} 0,$$

where $c(\mu)$ is the greatest non-negative solution to the equation $e^{-\mu x} = 1 - x$.

Corollary: Partial recovery is IT-infeasible if $\lambda s \leq 1$.

Intuition for impossibility result: exchanging small tree components

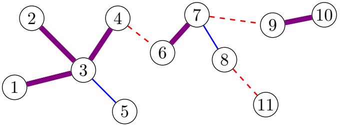
1. Information contained in the **intersection graph** $\mathcal{G} \wedge \mathcal{G}'$:



In our model $\mathcal{G} \wedge \mathcal{G}'$ is an Erdős-Rényi graph: $\mathcal{G} \wedge \mathcal{G}' \sim G(n, \lambda s/n)$.

Intuition for impossibility result: exchanging small tree components

1. Information contained in the **intersection graph** $\mathcal{G} \wedge \mathcal{G}'$:

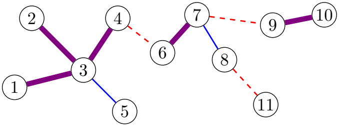


   In our model $\mathcal{G} \wedge \mathcal{G}'$ is an Erdős-Rényi graph: $\mathcal{G} \wedge \mathcal{G}' \sim G(n, \lambda s/n)$.

2. [Erdős, Rényi, Bollobás] typical fraction $c(\lambda s)$ of nodes in the giant component of $\mathcal{G} \wedge \mathcal{G}' \rightarrow$ the remaining $(1 - c(\lambda s))n$ nodes are almost all on **small tree components**.

Intuition for impossibility result: exchanging small tree components

1. Information contained in the **intersection graph** $\mathcal{G} \wedge \mathcal{G}'$:



   In our model $\mathcal{G} \wedge \mathcal{G}'$ is an Erdős-Rényi graph: $\mathcal{G} \wedge \mathcal{G}' \sim G(n, \lambda s/n)$.

2. [Erdős, Rényi, Bollobás] typical fraction $c(\lambda s)$ of nodes in the giant component of $\mathcal{G} \wedge \mathcal{G}' \rightarrow$ the remaining $(1 - c(\lambda s))n$ nodes are almost all on **small tree components**.
3. For any small tree **T**, a large number of copies of **T** will appear in $\mathcal{G} \wedge \mathcal{G}'$. **Reshuffle them at random in** $\mathcal{G} \rightarrow$ a lot of 'unnoticed' corrupted candidates for $\hat{\pi}$ that are far from π^*.

Positive result: testing tree correlation

Crucial remark: in the sparse regime, $\mathcal{G}, \mathcal{H}$ are locally **tree-like**.

Positive result: testing tree correlation

Crucial remark: in the sparse regime, $\mathcal{G}, \mathcal{H}$ are locally **tree-like**.

Recall that $(\mathcal{G}, \mathcal{H}) \sim \mathcal{G}(n, q = \lambda/n, s)$ with planted permutation π^*. Then, locally:

- if $u = \pi^*(i)$, the neighborhoods at depth d, $\mathcal{N}_\mathcal{G}(i)$ and $\mathcal{N}_\mathcal{H}(u) \simeq$ Galton-Waston trees of offspring $\mathcal{P}(\lambda)$, with intersection of offspring $\mathcal{P}(\lambda s)$.

Positive result: testing tree correlation

Crucial remark: in the sparse regime, $\mathcal{G}, \mathcal{H}$ are locally **tree-like**.

Recall that $(\mathcal{G}, \mathcal{H}) \sim \mathcal{G}(n, q = \lambda/n, s)$ with planted permutation π^*. Then, locally:

- if $u = \pi^*(i)$, the neighborhoods at depth d, $\mathcal{N}_\mathcal{G}(i)$ and $\mathcal{N}_\mathcal{H}(u) \simeq$ Galton-Waston trees of offspring $\mathcal{P}(\lambda)$, with intersection of offspring $\mathcal{P}(\lambda s)$.
- if $u \neq \pi^*(i)$, $\mathcal{N}_\mathcal{G}(i)$ and $\mathcal{N}_\mathcal{H}(u) \simeq$ independent Galton-Waston trees of offspring $\mathcal{P}(\lambda)$.

Positive result: testing tree correlation

Crucial remark: in the sparse regime, $\mathcal{G}, \mathcal{H}$ are locally **tree-like**.

Recall that $(\mathcal{G}, \mathcal{H}) \sim \mathcal{G}(n, q = \lambda/n, s)$ with planted permutation π^*. Then, locally:

- if $u = \pi^*(i)$, the neighborhoods at depth d, $\mathcal{N}_\mathcal{G}(i)$ and $\mathcal{N}_\mathcal{H}(u) \simeq$ Galton-Waston trees of offspring $\mathcal{P}(\lambda)$, with intersection of offspring $\mathcal{P}(\lambda s)$.
- if $u \neq \pi^*(i)$, $\mathcal{N}_\mathcal{G}(i)$ and $\mathcal{N}_\mathcal{H}(u) \simeq$ independent Galton-Waston trees of offspring $\mathcal{P}(\lambda)$.

New problem on trees: upon observing two unlabeled, rooted trees t, t' up to depth d, we want to be able to test:

$$(t, t') \sim \mathbb{P}_1 \quad \text{vs} \quad (t, t') \sim \mathbb{P}_0$$

with $\mathbb{P}_1 := s-$ correlated $GW_{\lambda,d}$ trees and $\mathbb{P}_0 := GW_{\lambda,d} \otimes GW_{\lambda,d}$.

Positive result: testing tree correlation

One sided test: test $\mathcal{T}_d : \mathcal{X}_d \times \mathcal{X}_d \to \{0, 1\}$ such that

$$\mathbb{P}_0(\mathcal{T}_d = 0) \to 1 \quad \text{and} \quad \liminf_{d \to \infty} \mathbb{P}_1(\mathcal{T}_d = 1) > 0.$$

Positive result: testing tree correlation

One sided test: test $\mathcal{T}_d : \mathcal{X}_d \times \mathcal{X}_d \to \{0, 1\}$ such that

$$\mathbb{P}_0(\mathcal{T}_d = 0) \to 1 \quad \text{and} \quad \liminf_{d\to\infty} \mathbb{P}_1(\mathcal{T}_d = 1) > 0.$$

Likelihood ratio: For $t, t' \in \mathcal{X}_d$,

$$L_d(t, t') := \frac{\mathbb{P}_{1,d}(t, t')}{\mathbb{P}_{0,d}(t, t')}.$$

Recursive computation: if c (resp. c') is the root degree in $\mathcal{T}$ (resp. $\mathcal{T}'$)

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

where $\mathcal{S}(k, \ell)$ is the set of injective mappings from $[k]$ to $[\ell]$, and

$$\psi(k, c, c') := \frac{\pi_{\lambda s}(k)\pi_{\lambda\bar{s}}(c-k)\pi_{\lambda\bar{s}}(c'-k)}{\pi_\lambda(c)\pi_\lambda(c')} \times \frac{(c-k)! \times (c-k')!}{c! \times c'!}$$

$$= e^{\lambda s} \times \frac{s^k \bar{s}^{d+d'-2k}}{\lambda^k k!}.$$

Positive result: testing tree correlation

Martingale properties: under $\mathbb{P}_0$, $(L_d)_d$ is a martingale w.r.t. to $\mathcal{F}_d := \sigma(t_{|d}, t'_{|d})$, and converges a.s. to L_∞.

Positive result: testing tree correlation

Martingale properties: under $\mathbb{P}_0$, $(L_d)_d$ is a martingale w.r.t. to $\mathcal{F}_d := \sigma(t_{|d}, t'_{|d})$, and converges a.s. to L_∞.

Sufficient condition: There exists a one sided test as soon as

$$\exists \varepsilon > 0, \ \forall a > 0, \ \liminf_{d \to \infty} \mathbb{P}_1(L_d > a) \geq \varepsilon > 0.$$

Positive result: testing tree correlation

Martingale properties: under $\mathbb{P}_0$, $(L_d)_d$ is a martingale w.r.t. to $\mathcal{F}_d := \sigma(t_{|d}, t'_{|d})$, and converges a.s. to L_∞.

Sufficient condition: There exists a one sided test as soon as

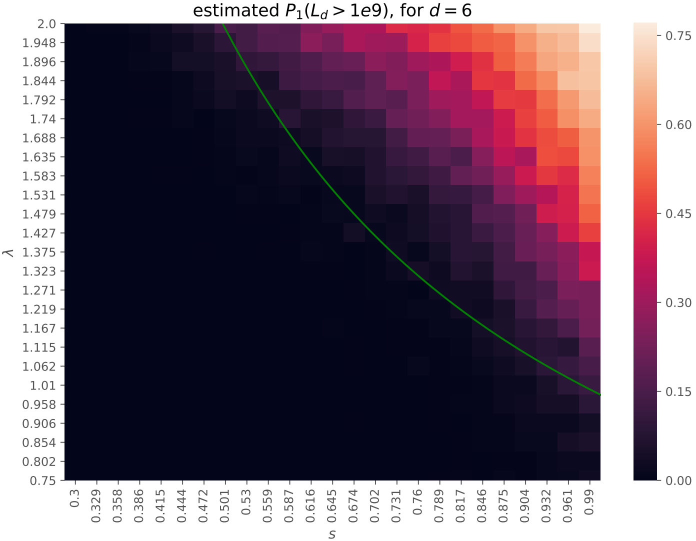
$$\exists \varepsilon > 0, \ \forall a > 0, \ \liminf_{d \to \infty} \mathbb{P}_1(L_d > a) \geq \varepsilon > 0.$$

KL - divergence:

$$KL_d := KL(\mathbb{P}_{1,d} || \mathbb{P}_{0,d}) = \mathbb{E}_1 \left[\log(L_d) \right].$$

$KL_d \to \infty$ and $\lambda s > 1 \implies$ one-sided test exists $\implies KL_d \to \infty$

Positive result: testing tree correlation



Positive result: testing tree correlation

Theorem (positive results, ongoing work)
Assume that one of the following holds:

(i) $\lambda s > 1$ *and*

$$KL_1 > \frac{1}{\lambda s - 1}\left[\lambda s(\log(\lambda/s) - 1) - 2\lambda(1-s)\log(1-s)\right]$$

(ii) $\lambda s > r_0$ *(r_0 large constant) and*

$$1 - s \leq \frac{1}{3+\eta}\sqrt{\frac{\log(\lambda s)}{\lambda^3 s}}$$

then one-sided testability holds.

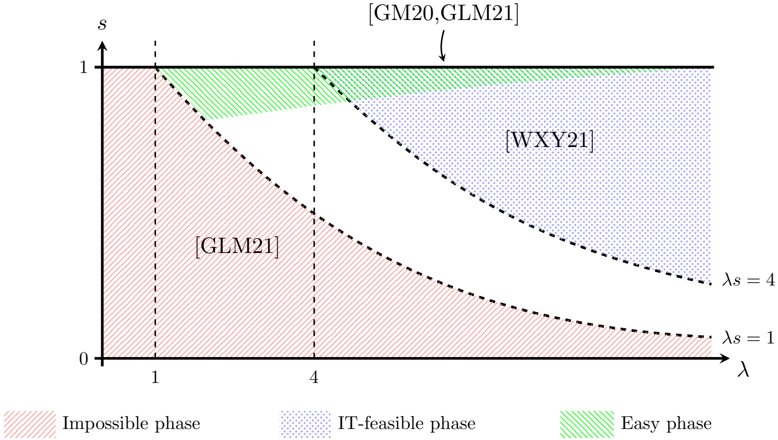
Positive result: testing tree correlation

Theorem (negative results, ongoing work)
If $\lambda s^2 < 1$, then for sufficiently large λ,

$$\limsup_d KL_d < \infty,$$

so that one-sided testability fails.

Conclusion: diagram for partial recovery



Concluding remarks

- Sparse graph alignment can be locally rephrased as an hypothesis testing problem: detecting correlation in (unlabeled, rooted) trees.
- The recursion computation of the likelihood ratio gives a natural belief-propagation method, running in polynomial-time.
- Future work:
  - $\lambda s = 1$ seems to be the sharp IT threshold.
  - Hard phase tight characterization still open.
  - Other random graph models, labeled version.

Thank you!