The graph alignment problem: a local point of view

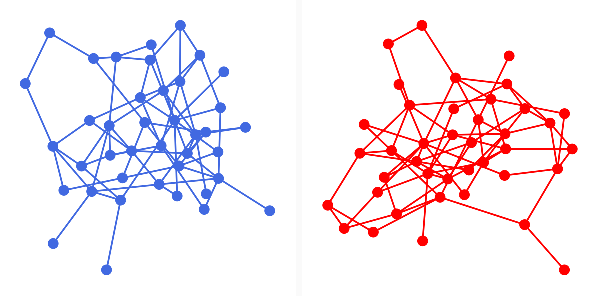
Luca Ganassali
INRIA, DI/ENS, PSL Research University
joint work with Laurent Massoulié and Marc Lelarge

Colloque Jeunes Probabilistes et Statisticiens - Saint Pierre d'Oléron - Oct. 2021

The graph alignment problem

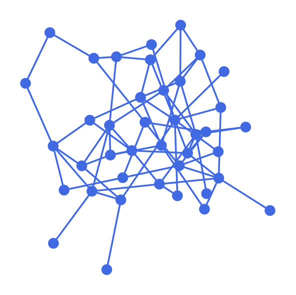
The graph alignment problem

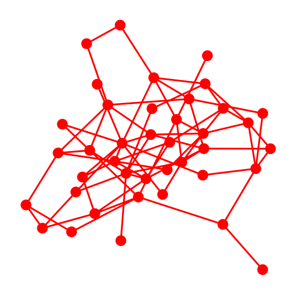
Question : What is the 'best way' to match the nodes of two graphs $\mathcal{G}, \mathcal{H}$?



The graph alignment problem

Question : What is the 'best way' to match the nodes of two graphs $\mathcal{G}, \mathcal{H}$?



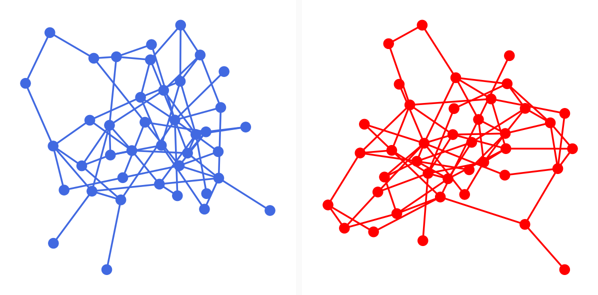


Mathematically :
if $|V(\mathcal{G})| = |V(\mathcal{H})|$, find a bijection $f : V(\mathcal{G}) \to V(\mathcal{H})$ minimizing :

$$\sum_{i,j \in V(\mathcal{G})} \left(\mathbf{1}_{(i,j) \in E(\mathcal{G})} - \mathbf{1}_{(f(i),f(j)) \in E(\mathcal{H})} \right)^2 =: \sharp \textit{edge disagreements}$$

The graph alignment problem

Question : What is the 'best way' to match the nodes of two graphs $\mathcal{G}, \mathcal{H}$?



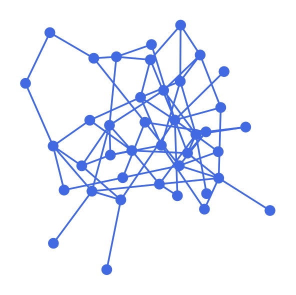
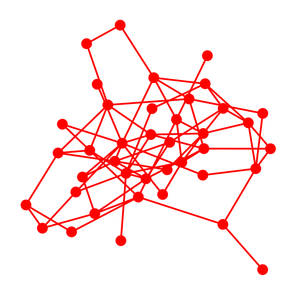
Mathematically :

Equivalently, if $V(\mathcal{G}) = V(\mathcal{H}) = [n]$, solve

$$\arg\max_{\Pi \in \mathcal{S}_n} \mathrm{Tr}\left(A_{\mathcal{G}} \Pi A_{\mathcal{H}} \Pi^T\right)$$

The graph alignment problem

Question : What is the 'best way' to match the nodes of two graphs $\mathcal{G}, \mathcal{H}$?

Mathematically :
Equivalently, if $V(\mathcal{G}) = V(\mathcal{H}) = [n]$, solve

$$\underbrace{\arg\max_{\Pi \in \mathcal{S}_n} \mathrm{Tr}\left(A_{\mathcal{G}} \Pi A_{\mathcal{H}} \Pi^T\right)}$$

NP-hard...

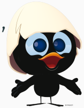
Planted graph alignment

Planted graph alignment

Idea : Study the problem in the mean-case setting (on *random instances*), *planting* a solution π^* in the model, and try to recover it *w.h.p.*

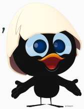
Planted graph alignment

Idea : Study the problem in the mean-case setting (on *random instances*), *planting* a solution π^* in the model, and try to recover it *w.h.p.*



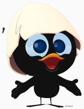
Planted graph alignment

Idea : Study the problem in the mean-case setting (on *random instances*), *planting* a solution π^* in the model, and try to recover it *w.h.p.*



Correlated Erdős-Rényi model

Planted graph alignment



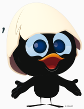
Idea : Study the problem in the mean-case setting (on *random instances*), *planting* a solution π^* in the model, and try to recover it *w.h.p.*

Correlated Erdős-Rényi model

1. Two graphs $\mathcal{G}$ (blue) and $\mathcal{G}'$ (red) with same node set $[n]$, with edges sampled independently as follows :

- with probability $\lambda s/n$ to get two-colored edges ;
- with probability $\lambda(1-s)/n$ to get a blue monochromatic (resp. red monochromatic) edge ;
- with remaining probability to get a non-edge,

Planted graph alignment



Idea : Study the problem in the mean-case setting (on *random instances*), *planting* a solution π^* in the model, and try to recover it *w.h.p.*

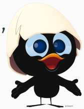
Correlated Erdős-Rényi model

1. Two graphs $\mathcal{G}$ (blue) and $\mathcal{G}'$ (red) with same node set $[n]$, with edges sampled independently as follows :

- with probability $\lambda s/n$ to get two-colored edges ;
- with probability $\lambda(1-s)/n$ to get a blue monochromatic (resp. red monochromatic) edge ;
- with remaining probability to get a non-edge,

2. Relabel the vertices of the red graph $\mathcal{G}'$ with an uniform independent permutation $\pi^* \in \mathcal{S}_n$. We observe $\mathcal{G}$ and $\mathcal{H} := \mathcal{G}' \circ \pi^*$.

Planted graph alignment



Idea : Study the problem in the mean-case setting (on *random instances*), *planting* a solution π^* in the model, and try to recover it *w.h.p.*

Correlated Erdős-Rényi model

1. Two graphs $\mathcal{G}$ (blue) and $\mathcal{G}'$ (red) with same node set $[n]$, with edges sampled independently as follows :

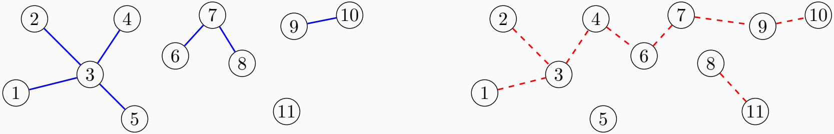
- with probability $\lambda s/n$ to get two-colored edges ;
- with probability $\lambda(1-s)/n$ to get a blue monochromatic (resp. red monochromatic) edge ;
- with remaining probability to get a non-edge,

2. Relabel the vertices of the red graph $\mathcal{G}'$ with an uniform independent permutation $\pi^* \in \mathcal{S}_n$. We observe $\mathcal{G}$ and $\mathcal{H} := \mathcal{G}' \circ \pi^*$.

(The marginals $\mathcal{G}, \mathcal{H}$ are Erdős-Rényi random graphs with average degree λ).

Planted graph alignment : Correlated Erdős-Rényi model

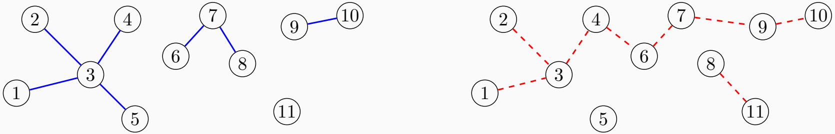
1.



$\mathcal{G}, \mathcal{G}'$

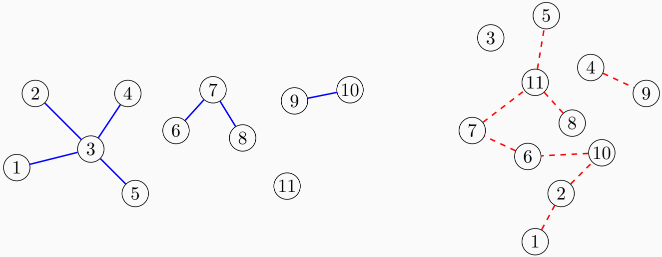
Planted graph alignment : Correlated Erdős-Rényi model

1.



$\mathcal{G}, \mathcal{G}'$

2.



$\mathcal{G}, \mathcal{H}$, with $\pi^* = (6)(1\ 5\ 3\ 11\ 9\ 2\ 8\ 4\ 7\ 10)$

Planted graph alignment : Correlated Erdős-Rényi model

Goal : estimate the latent vertex correspondence π^*, in a *sparse regime* where the *average degree* λ and the *correlation* $s \in [0, 1]$ are constant (not scaling with n).

Planted graph alignment : Correlated Erdős-Rényi model

Goal : estimate the latent vertex correspondence π^*, in a *sparse regime* where the *average degree* λ and the *correlation* $s \in [0,1]$ are constant (not scaling with n).

Remarks :

- The MAP estimator of π^* is $\arg\max_\Pi \langle A_\mathcal{G}, \Pi A_\mathcal{H} \Pi^T \rangle$...

Planted graph alignment : Correlated Erdős-Rényi model

<u>Goal</u> : estimate the latent vertex correspondence π^*, in a *sparse regime* where the *average degree* λ and the *correlation* $s \in [0, 1]$ are constant (not scaling with n).

<u>Remarks</u> :

- The MAP estimator of π^* is $\arg\max_\Pi \langle A_\mathcal{G}, \Pi A_\mathcal{H} \Pi^T \rangle$...
- We can only hope for *partial recovery*

Planted graph alignment : Correlated Erdős-Rényi model

Goal : estimate the latent vertex correspondence π^*, in a *sparse regime* where the *average degree* λ and the *correlation* $s \in [0, 1]$ are constant (not scaling with n).

Remarks :

- The MAP estimator of π^* is $\arg\max_\Pi \langle A_\mathcal{G}, \Pi A_\mathcal{H} \Pi^T \rangle$...
- We can only hope for *partial recovery* (isolated nodes)...

Measure of performance

For any subset $\mathcal{C} \subset [n]$, the performance of any one-to-one estimator $\hat{\pi} : \mathcal{C} \to [n]$

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

Note that the estimator $\hat{\pi}$ only consists in a partial matching. The *error fraction* of $\hat{\pi}$ with the unknown permutation π^* is defined as

$$\mathrm{err}(\pi^*, \hat{\pi}) := \frac{1}{n} \sum_{i \in \mathcal{C}} \mathbf{1}_{\hat{\pi}(i) \neq \pi^*(i)} = \frac{|\mathcal{C}|}{n} - \mathrm{ov}(\pi^*, \hat{\pi}).$$

Measure of performance

For any subset $\mathcal{C} \subset [n]$, the performance of any one-to-one estimator $\hat{\pi} : \mathcal{C} \to [n]$

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

Note that the estimator $\hat{\pi}$ only consists in a partial matching. The *error fraction* of $\hat{\pi}$ with the unknown permutation π^* is defined as

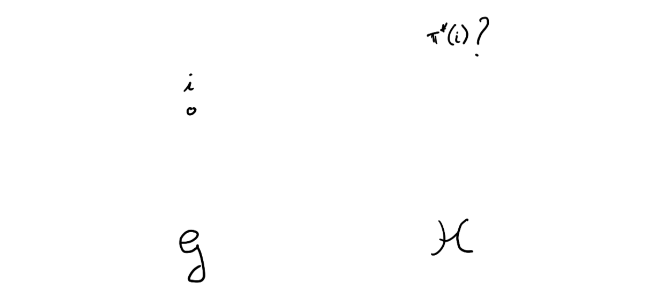
$$\mathrm{err}(\pi^*, \hat{\pi}) := \frac{1}{n} \sum_{i \in \mathcal{C}} \mathbf{1}_{\hat{\pi}(i) \neq \pi^*(i)} = \frac{|\mathcal{C}|}{n} - \mathrm{ov}(\pi^*, \hat{\pi}).$$

A sequence of injective estimators $\{\hat{\pi}_n\}_n$ is said to achieve

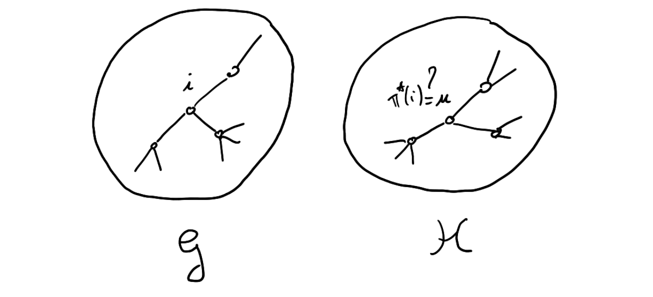
- *Partial recovery* if there exists some $\alpha > 0$ such that $\mathbb{P}(\mathrm{ov}(\pi^*, \hat{\sigma}) > \alpha) \underset{n \to \infty}{\longrightarrow} 1$,
- *One-sided partial recovery* if it achieves partial recovery and $\mathbb{P}(\mathrm{err}(\pi^*, \hat{\sigma}) = o(1)) \underset{n \to \infty}{\longrightarrow} 1$.

A local approach

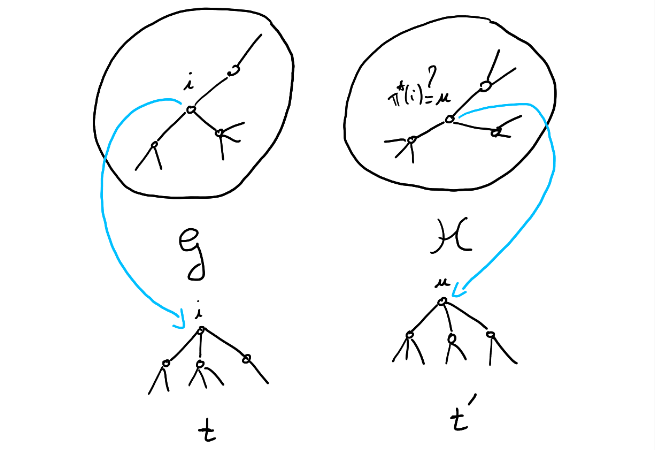
A local approach



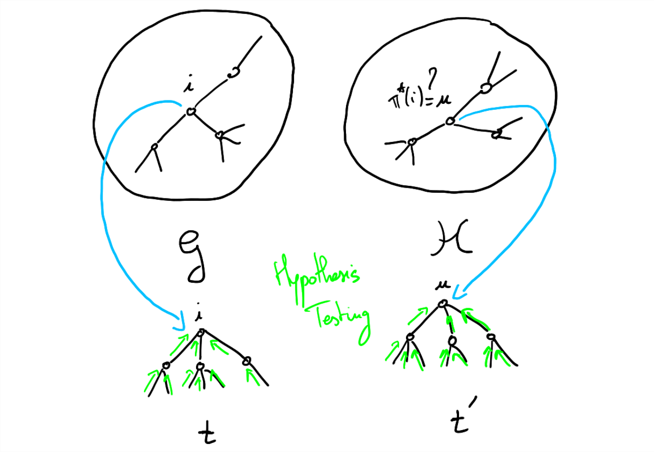
A local approach



A local approach



A local approach



From graphs to 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 $u = \pi^*(i)$, $(\mathcal{N}_i, \mathcal{N}_u) \simeq \mathrm{GW}$ 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 $\mathbb{P}_{0,d}$).

From graphs to 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 $u = \pi^*(i)$, $(\mathcal{N}_i, \mathcal{N}_u) \simeq \mathrm{GW}$ 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 $i \in V(\mathcal{G}), u \in V(\mathcal{H})$, look at the neighborhoods $\mathcal{N}_i$ and $\mathcal{N}_u$ at depth d :

- if $u = \pi^*(i)$, $(\mathcal{N}_i, \mathcal{N}_u) \simeq \mathrm{GW}$ 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 $\mathbb{P}_{0,d}$? $\to$ likelihood ratio

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

Computing the likelihood ratio

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

Results

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).

Theorem
Let

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

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*) *with probability* $1 - p_{\mathrm{ext}}(\lambda s) > 0$, L_d *diverges to* $+\infty$ *with rate* $\Omega\left(\exp\left(\Omega(1) \times (\lambda s)^d\right)\right)$.

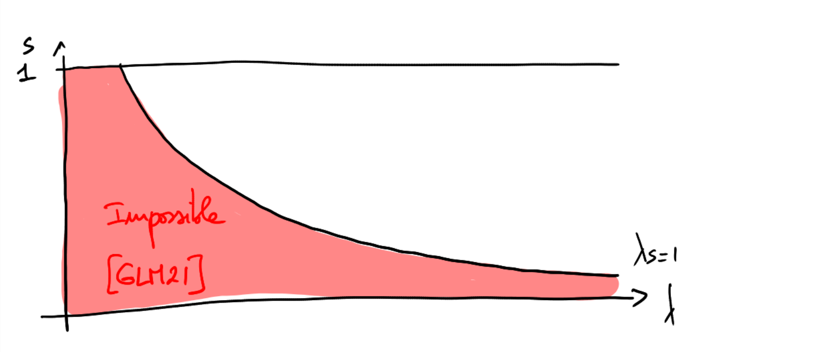
Recall : estimator $\hat{\pi} : \mathcal{C} \to [n]$ is said to achieve

- *Partial recovery* if there exists some $\epsilon > 0$ such that $\mathbb{P}(\mathrm{ov}(\pi^*, \hat{\sigma}) > \epsilon) \underset{n\to\infty}{\longrightarrow} 1$,
- *One-sided partial recovery* if it achieves partial recovery and $\mathbb{P}(\mathrm{err}(\pi^*, \hat{\sigma}) = o(1)) \underset{n\to\infty}{\longrightarrow} 1$.

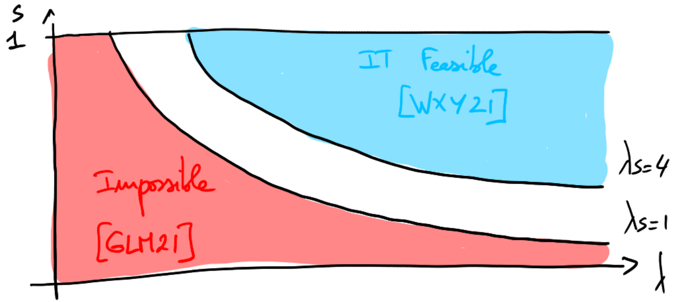
Theorem
For given (λ, s), if one-sided correlation detection is feasible, then one-sided partial alignment in the correlated Erdős-Rényi model $\mathcal{G}(n, \lambda/n, s)$ is achieved in polynomial time by our belief propagation algorithm.

Phase diagram

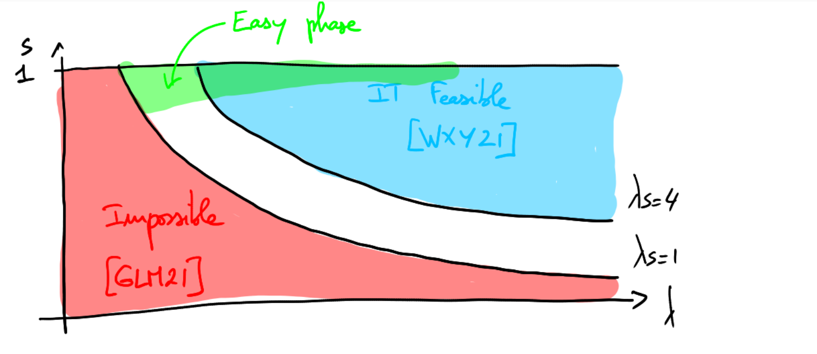
Partial alignment in the regime with constant mean degree



Partial alignment in the regime with constant mean degree

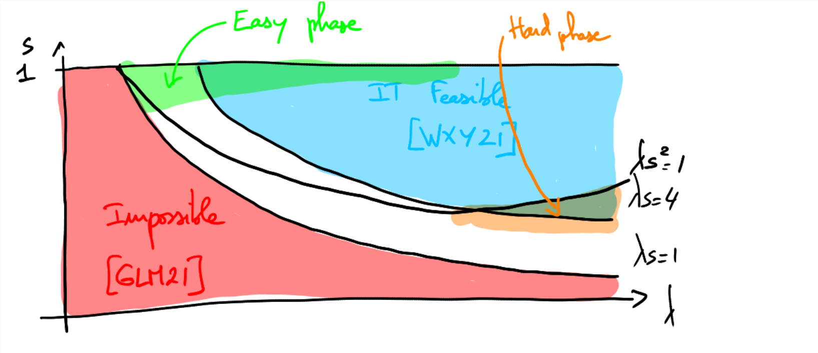


Partial alignment in the regime with constant mean degree



Sufficient conditions for the existence of one-sided test based on the Kullback-Leibler divergence or the number of automorphisms of GW trees.

Partial alignment in the regime with constant mean degree



Conjectured hard phase based on the impossibility of one-sided test because KL_d is bounded.

Conclusion

Conclusion

- Graph alignment is hard in general, we study its planted version.
- In a sparse regime, we establish a link between graph alignment and the correlation detection problem on trees.
- A belief-propagation algorithm can reach good performances, and seems to exhibit a hard phase for this problem.

Thank you!