Graph matching: from fundamental limits to algorithms

Marc Lelarge (Inria and ENS Paris)

Alignment of graphs consists in finding a mapping between the nodes of two graphs which preserves most of the edges. For two correlated Erdős–Rényi, we present information-theoretic results in the sparse regime. We then propose an algorithm based on local comparisons and prove theoretical guarantees. Finally, we show the empirical success of learning algorithms.

Graph matching: from fundamental limits to algorithms

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

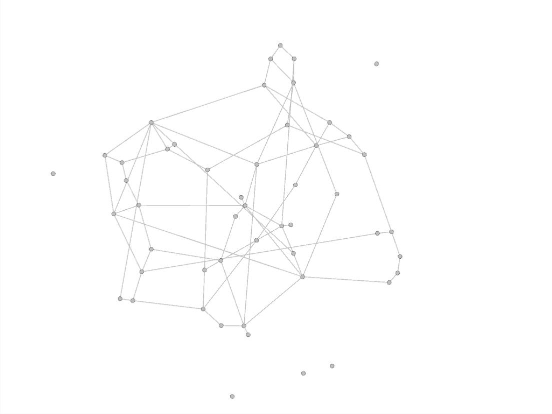
statistical learning on LARge scale GRaphs - March 2023

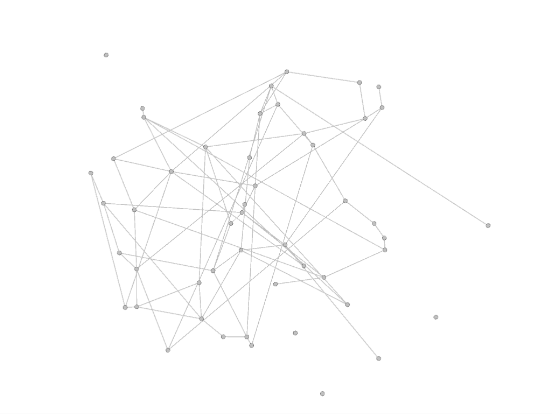
Alignment of Graphs

Problem : alignment of graphs

From graph 1 (on the left), put indices on its vertices, perturb the graph by adding and removing a few edges and remove indices to obtain graph 2 (on the right).

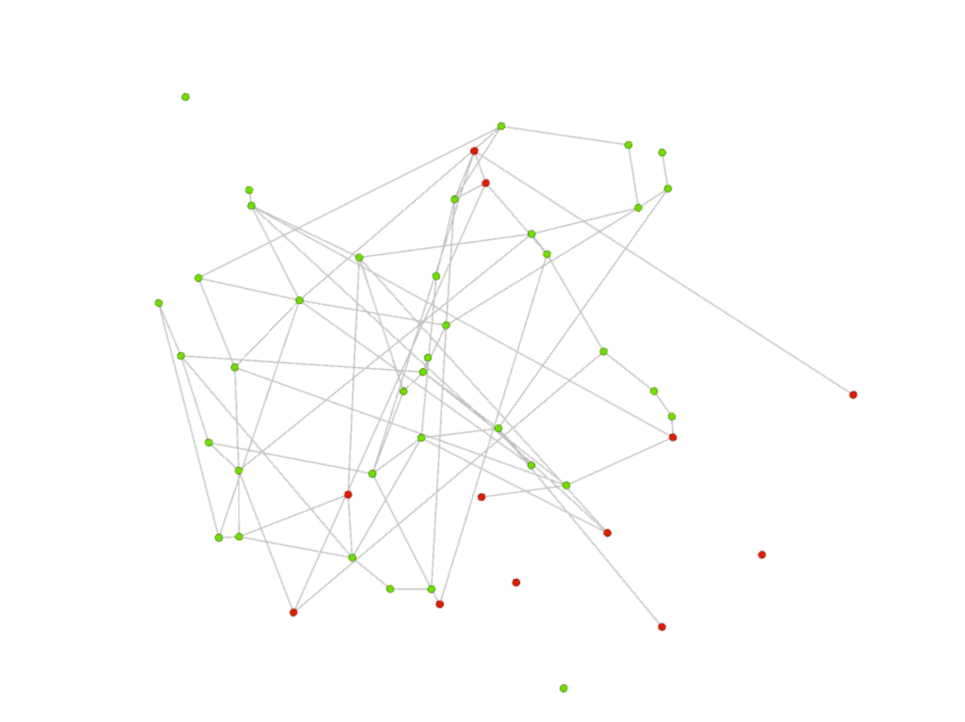
Task : recover the indices on vertices of graph 2.





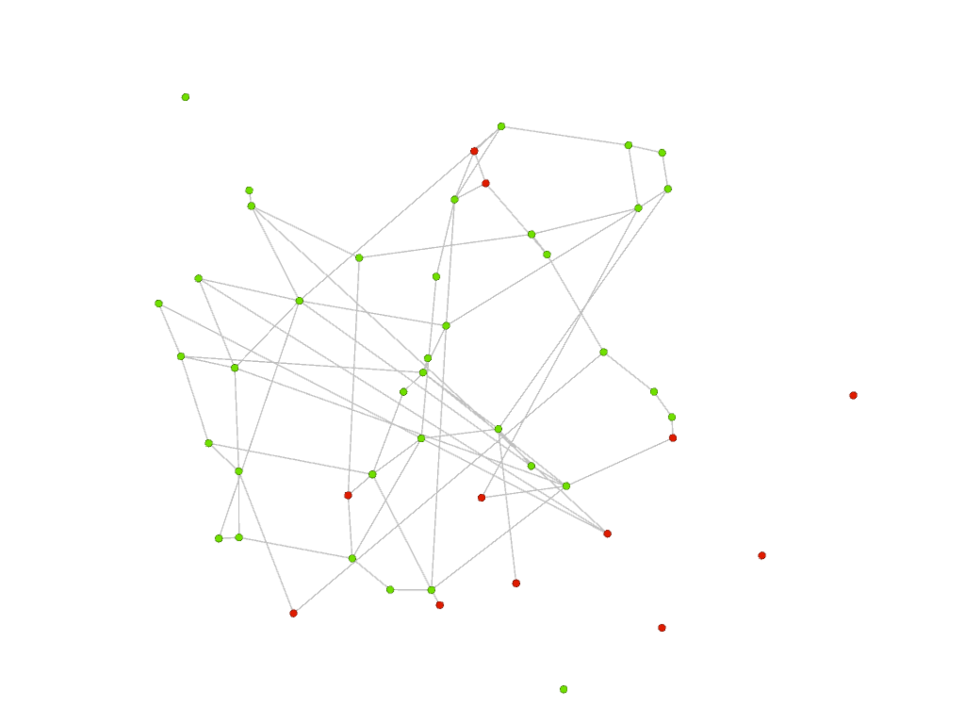
Result with FGNN

Green vertices are good predictions. Red vertices are errors (graph 2).



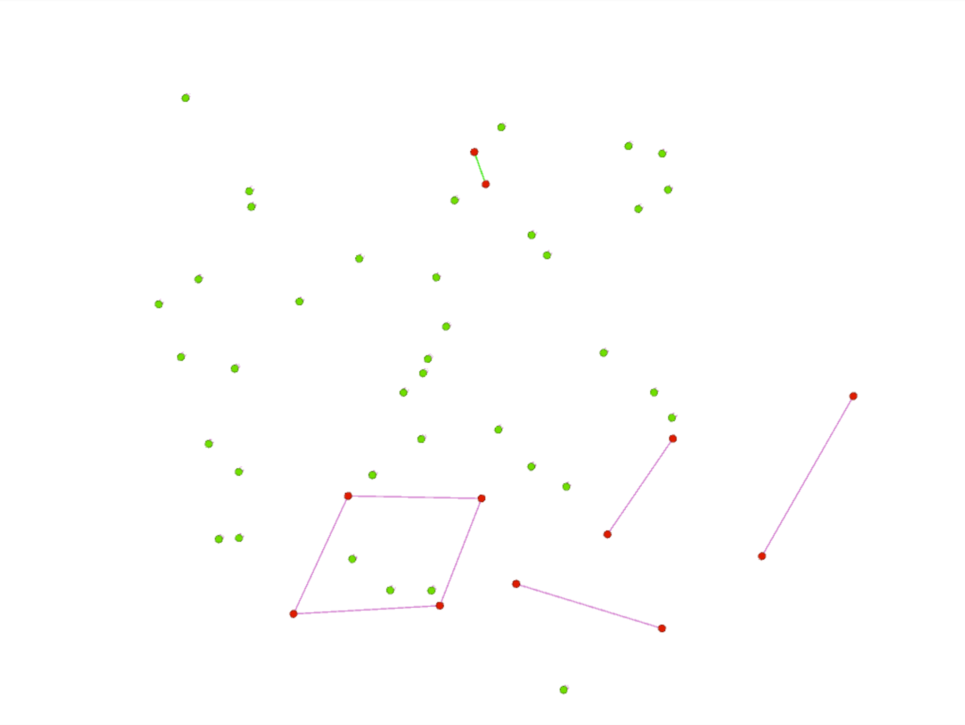
Result with FGNN

Green vertices are good predictions. Red vertices are errors (graph 1).



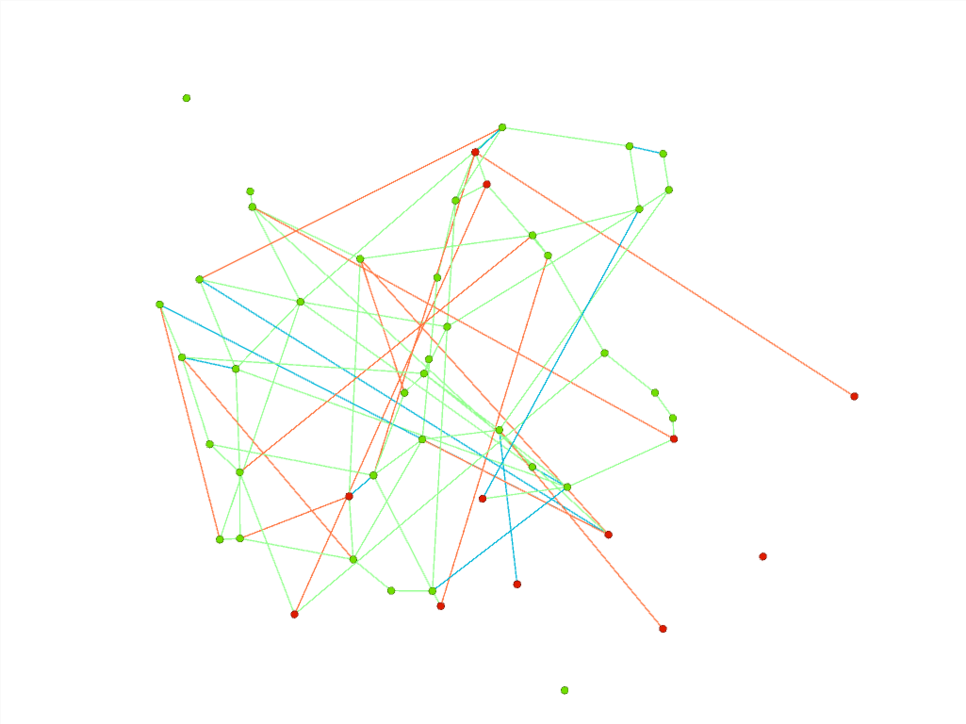
Result with FGNN

Here are the 'wrong' matchings or cycles.



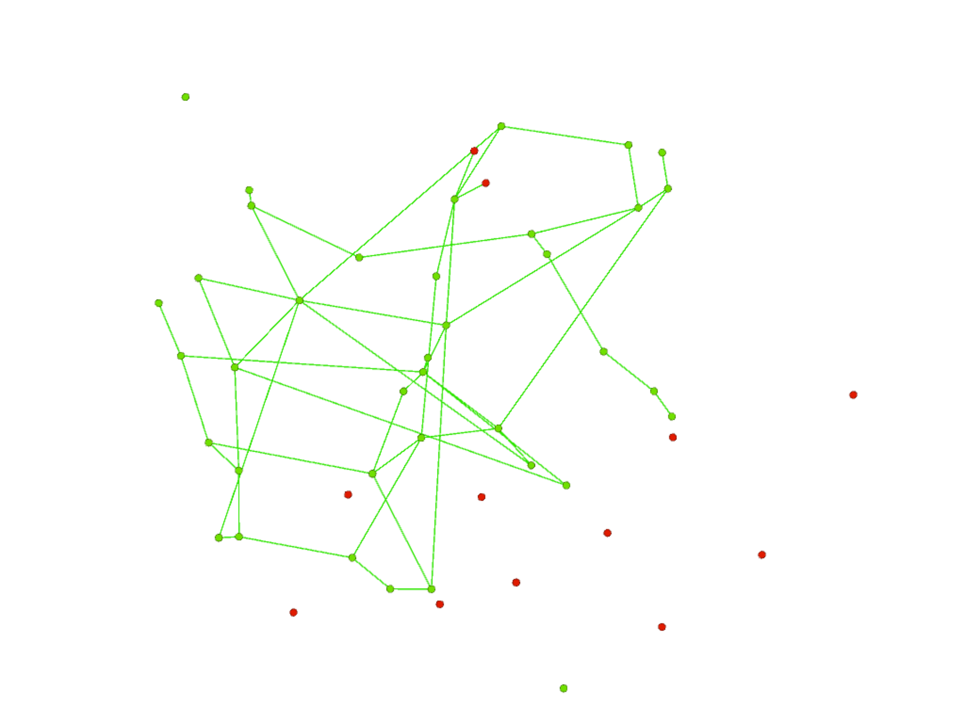
Result with FGNN

Superposing the 2 graphs : green edges in both, orange and blue edges in graph 1 and 2 resp.



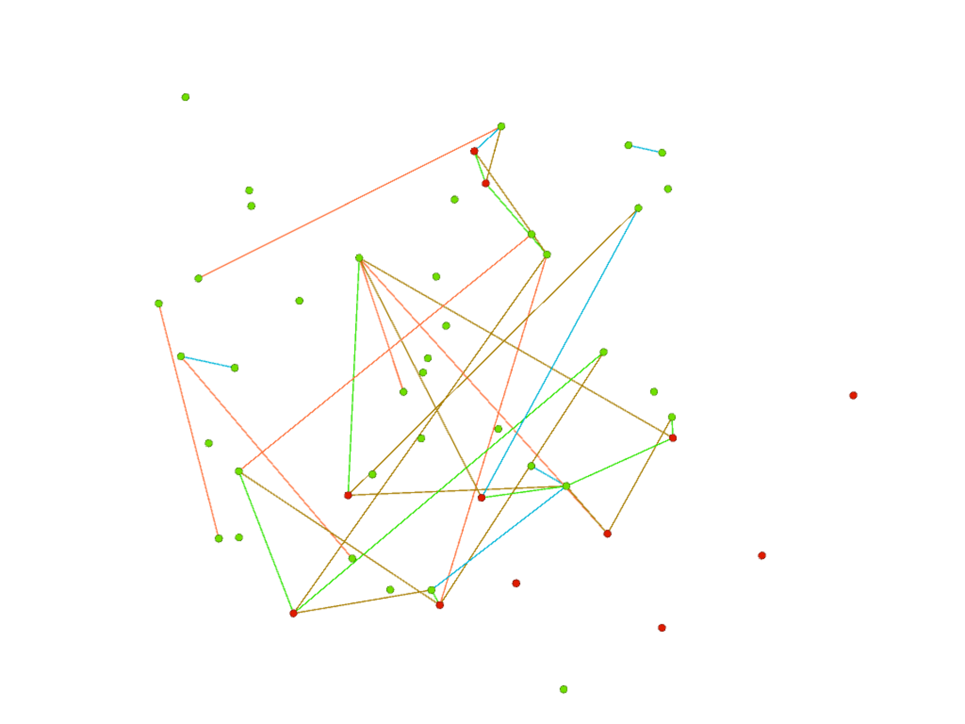
Result with FGNN

Matched edges.



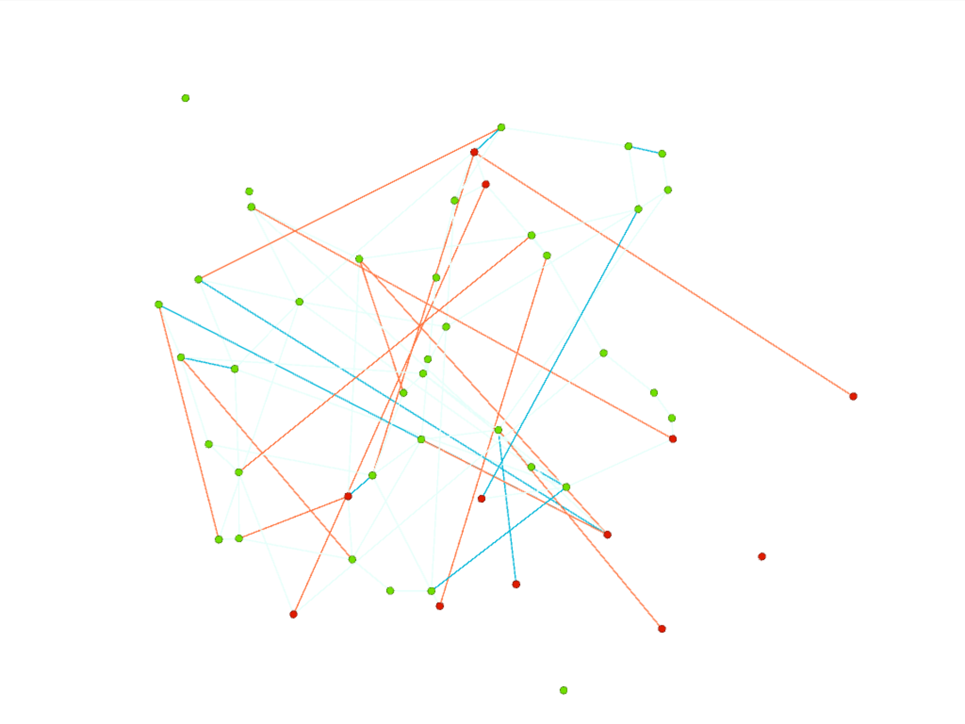
Result with FGNN

Mismatched edges.



Result with FGNN

Green vertices are well paired vertices. Red vertices are errors.



Motivation

Optimization $\quad \text{Overlap}(\pi) = \sum_{i \le j} G_{ij} G'_{\pi(i)\pi(j)}$

QAP $\quad G \neq G'$ $\quad$ difficult NP-hard.

difficult to approx

ex. $G, \; G' = \text{cycle} \longrightarrow$ TSP

$G, \; G' = K_k \longrightarrow$ k-clique

Motivation

Optimization $\quad \text{Overlap}(\pi) = \sum_{i \le j} G_{ij} G'_{\pi(i)\pi(j)}$

QAP $\quad G \neq G'$ $\quad$ difficult NP-hard.

difficult to approx

ex. $G,\ G' = \text{cycle} \longrightarrow \text{TSP}$

$G,\ G' = K_k \longrightarrow k\text{-clique}$

QAP for $G = G'$ $\iff$ Graph Isomorphism

quasipoly time $2^{O(\log n)^3}$

(Babai 2019)

Motivation

Optimization Overlap$(\pi) = \sum_{i \leq j} G_{ij} G'_{\pi(i)\pi(j)}$

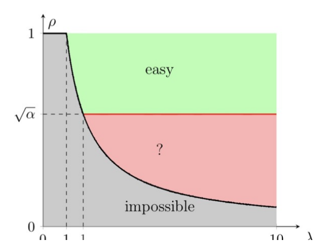
QAP $G \neq G'$ difficult NP-hard.
difficult to approx

ex. G, G' = cycle ⟶ TSP
G, $G' = K_k$ ⟶ k-clique

QAP for $G = G'$ ⟺ Graph Isomorphism
quasipoly time $2^{O(\log n)^3}$
(Babai 2019)

What happens on random instances? $G(n, p)$

$\mathrm{Cor}\left(G_{uv}, G'_{\tilde{\pi}(u)\tilde{\pi}(v)}\right) = \rho$



(Mao
Wu
Xu
Yu 2022)

Motivation

New Pb. $\mathrm{Cor}(G_{uv}, G'_{\tilde{\pi}(u)\tilde{\pi}(v)}) = \rho \rightarrow$ Recover $\tilde{\pi}$

(planted problem)

graphs are not connected.

Graph Isomorphism

(MWXY'22)

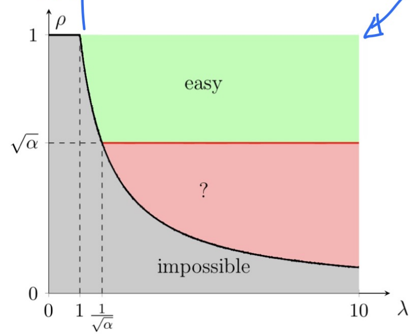


Figure 2: The phase diagram for exact recovery in the logarithmic degree regime, where $nq = \lambda \log n$ for a fixed constant $\lambda > 0$. The impossible and easy regime are given by $\rho < \min\{1, 1/\lambda\}$ and $\rho > \max\{\sqrt{\alpha}, 1/\lambda\}$, respectively. No polynomial-time algorithm is known to achieve exact recovery in the red regime.

Motivation

New Pb. $\mathrm{Cor}(G_{uv}\, G'_{\tilde\pi(u)\tilde\pi(v)}) = \rho \to$ Recover $\tilde\pi$

(planted problem)

graphs are not connected.

Graph Isomorphism

(MWXY '22)

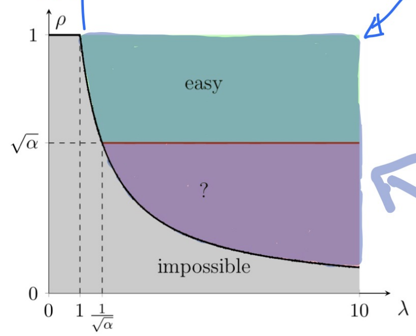


Figure 2: The phase diagram for exact recovery in the logarithmic degree regime, where $nq = \lambda \log n$ for a fixed constant $\lambda > 0$. The impossible and easy regime are given by $\rho < \min\{1, 1/\lambda\}$ and $\rho > \max\{\sqrt{\alpha}, 1/\lambda\}$, respectively. No polynomial-time algorithm is known to achieve exact recovery in the red regime.

exact reconstruction: MAP = QAP.

$\tilde\pi = \pi^* = \operatorname{argmax}$ Overlap.

QAP in the impossible regime?

Motivation

New Pb. $\mathrm{Cor}(G_{uv}\, G'_{\tilde{\pi}(u)\tilde{\pi}(v)}) = \rho \rightarrow$ Recover $\tilde{\pi}$
(planted problem)

graphs are not connected.

Graph Isomorphism

(MWXY '22)

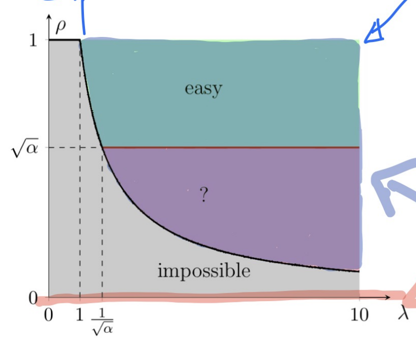


Figure 2: The phase diagram for exact recovery in the logarithmic degree regime, where $nq = \lambda \log n$ for a fixed constant $\lambda > 0$. The impossible and easy regime are given by $\rho < \min\{1, 1/\lambda\}$ and $\rho > \max\{\sqrt{\alpha}, 1/\lambda\}$, respectively. No polynomial-time algorithm is known to achieve exact recovery in the red regime.

exact reconstruction: MAP = QAP.
$\tilde{\pi} = \pi^* = \mathrm{argmax}$ Overlap.

QAP in the impossible regime?

(Ding, Du, Gong '22) PTAS for QAP with $\rho = 0$
(Poly-time approx scheme)

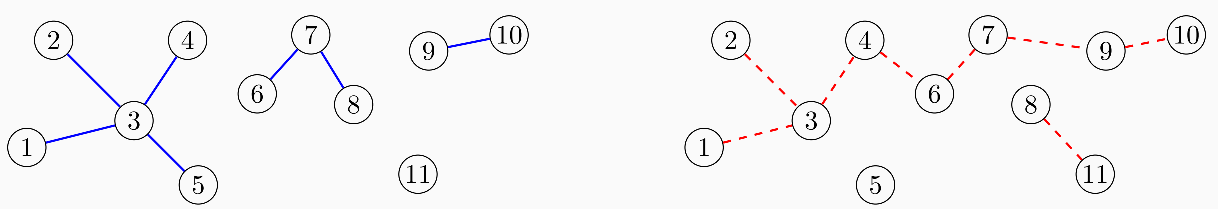
Theoretical results

Correlated Erdős-Rényi (1)

Two graphs $\mathcal{G}$ and $\mathcal{G}'$ with the same set of nodes $[n]$ and with respectively blue and red edges. The blue and red edges are obtained by sampling uniformly at random :

- with probability $\lambda s/n$ to get two-colored edges ;
- with probability $\lambda(1-s)/n$ to get a blue (monochromatic) edge ;
- with probability $\lambda(1-s)/n$ to get a red (monochromatic) edge ;
- with probability $1-\lambda(2-s)/n$ to get a non-edge,

where $\lambda > 0$ and $s \in [0,1]$ are fixed parameters when n will tend to infinity.

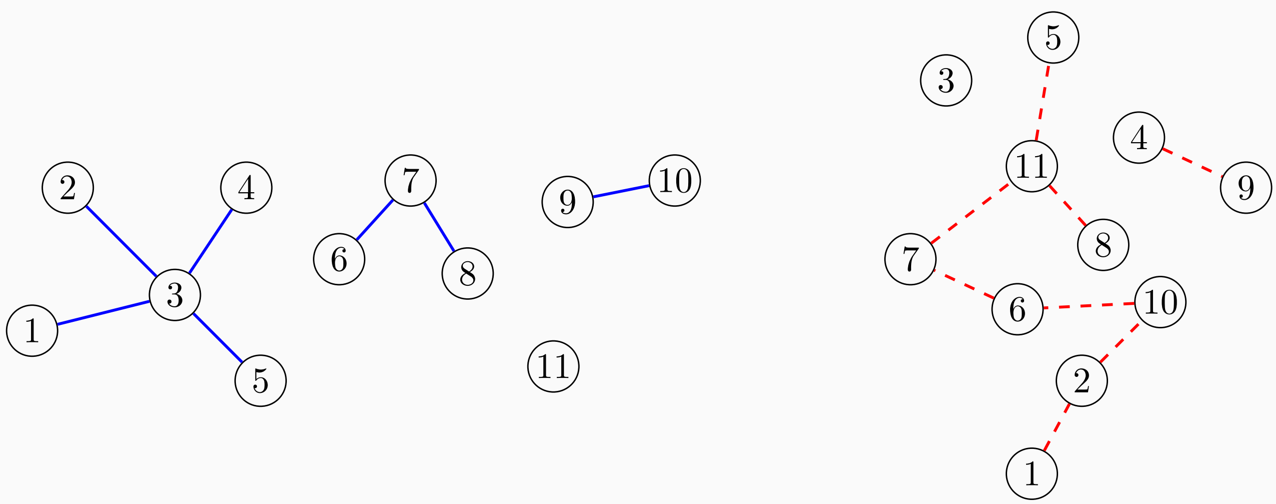


Correlated Erdős-Rényi (2)

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

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

The goal is to estimate the latent vertex correspondence π^*.



Exact recovery of π^*

Dependence of joint distribution in $e(G \wedge G')$:

$$\mathbb{P}(\mathcal{G} = G, \mathcal{G}' = G') \propto \left[\frac{s(n - \lambda(2-s))}{\lambda(1-s)^2}\right]^{e(G \wedge G')}.$$

MAP estimator of π^* is the solution of the QAP problem :

$$\arg\max_{\Pi} \langle A, \Pi B \Pi^T \rangle$$

Exact recovery of π^*

Dependence of joint distribution in $e(G \wedge G')$:

$$\mathbb{P}(\mathcal{G} = G, \mathcal{G}' = G') \propto \left[\frac{s(n - \lambda(2 - s))}{\lambda(1 - s)^2}\right]^{e(G \wedge G')} .$$

MAP estimator of π^* is the solution of the QAP problem :

$$\arg\max_{\Pi} \langle A, \Pi B \Pi^T \rangle$$

Only feasible when $\lambda = \Omega(\log n)$ Cullina and Kiyavash (2017), Mao et al. (2022) Ding et al. (2022), Ding and Du (2022).

Exact recovery of π^*

Dependence of joint distribution in $e(G \wedge G')$:

$$\mathbb{P}(\mathcal{G} = G, \mathcal{G}' = G') \propto \left[\frac{s(n - \lambda(2 - s))}{\lambda(1 - s)^2}\right]^{e(G \wedge G')}.$$

MAP estimator of π^* is the solution of the QAP problem :

$$\arg\max_{\Pi} \langle A, \Pi B \Pi^T \rangle$$

Only feasible when $\lambda = \Omega(\log n)$ Cullina and Kiyavash (2017), Mao et al. (2022) Ding et al. (2022), Ding and Du (2022).

In the rest of this talk, we consider the sparse regime where the average degree λ is not scaling with n.

Planted graph alignment

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

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

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

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

Planted graph alignment

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

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

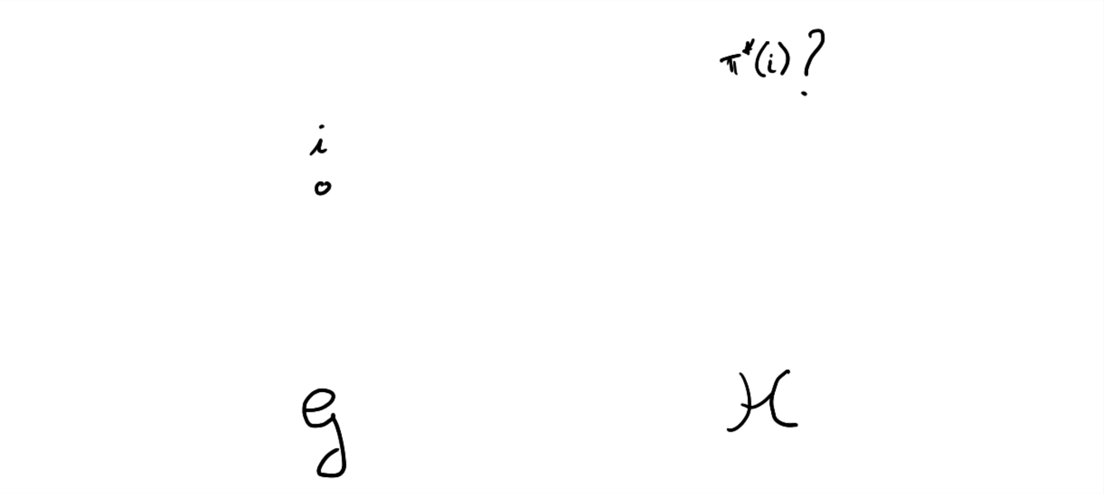
Note that the estimator $\hat{\sigma}$ only consists in a partial matching. The *error fraction* of $\hat{\sigma}$ with the unknown permutation π^* is defined as

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

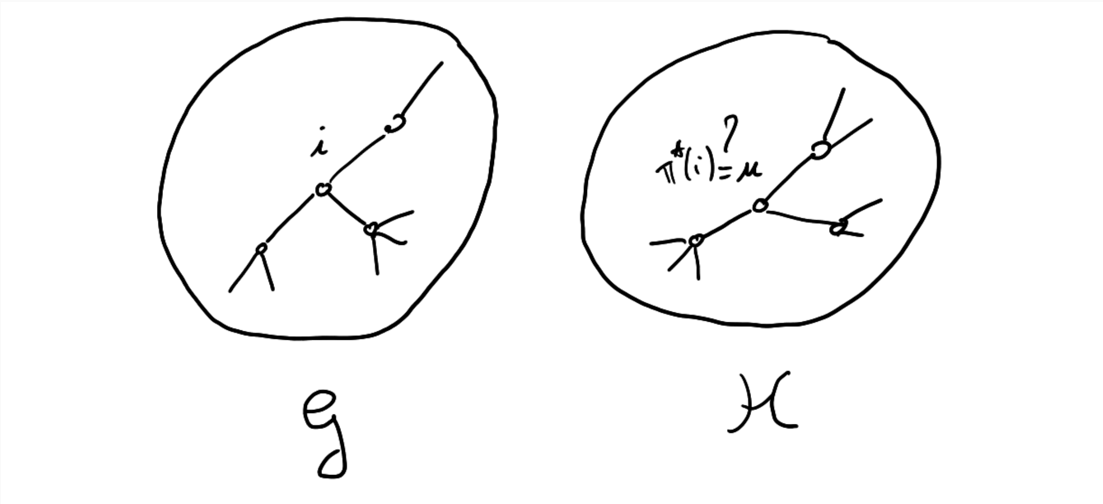
A sequence of injective estimators $\{\hat{\sigma}_n\}_n$ is said to achieve

- *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 $\mathbb{P}(\mathrm{err}(\pi^*, \hat{\sigma}) = o(1)) \underset{n \to \infty}{\longrightarrow} 1$.

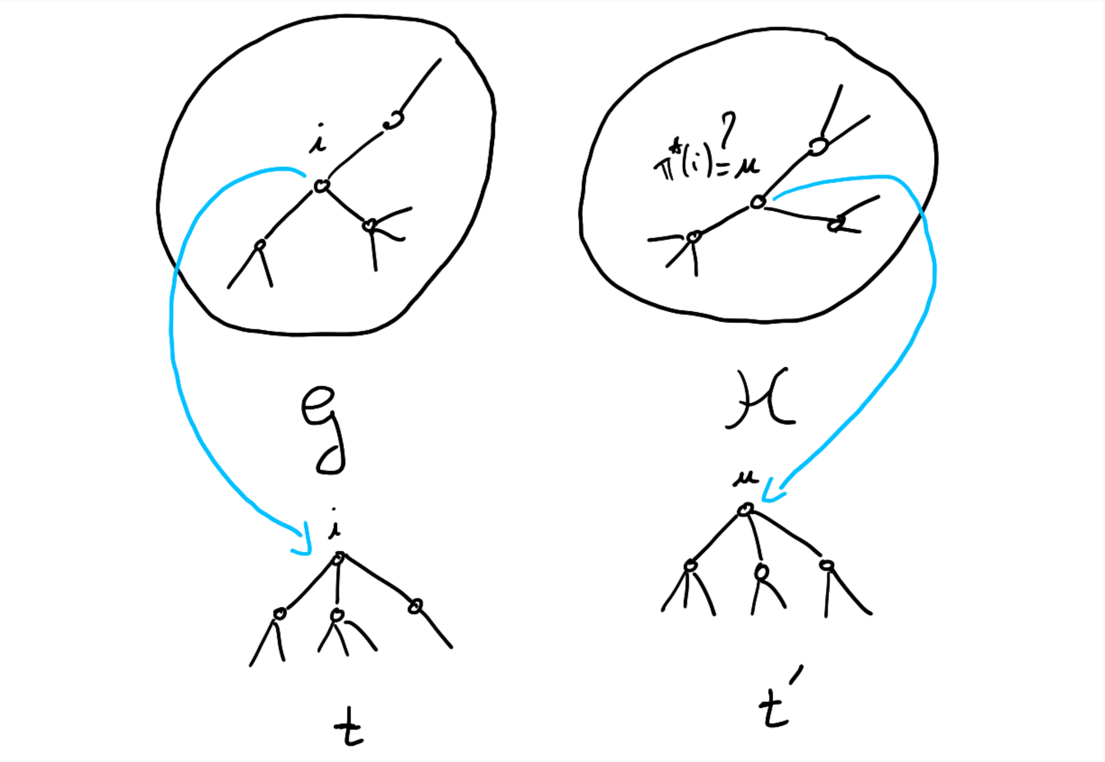
A (local) algorithm



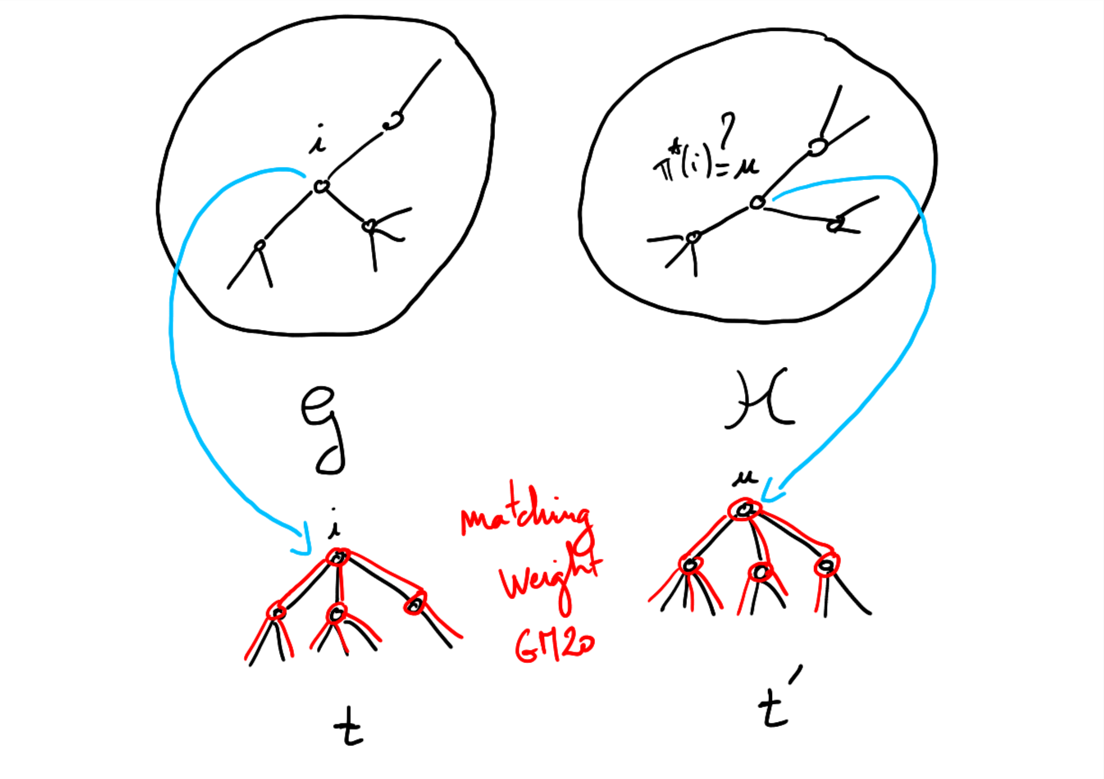
A (local) algorithm



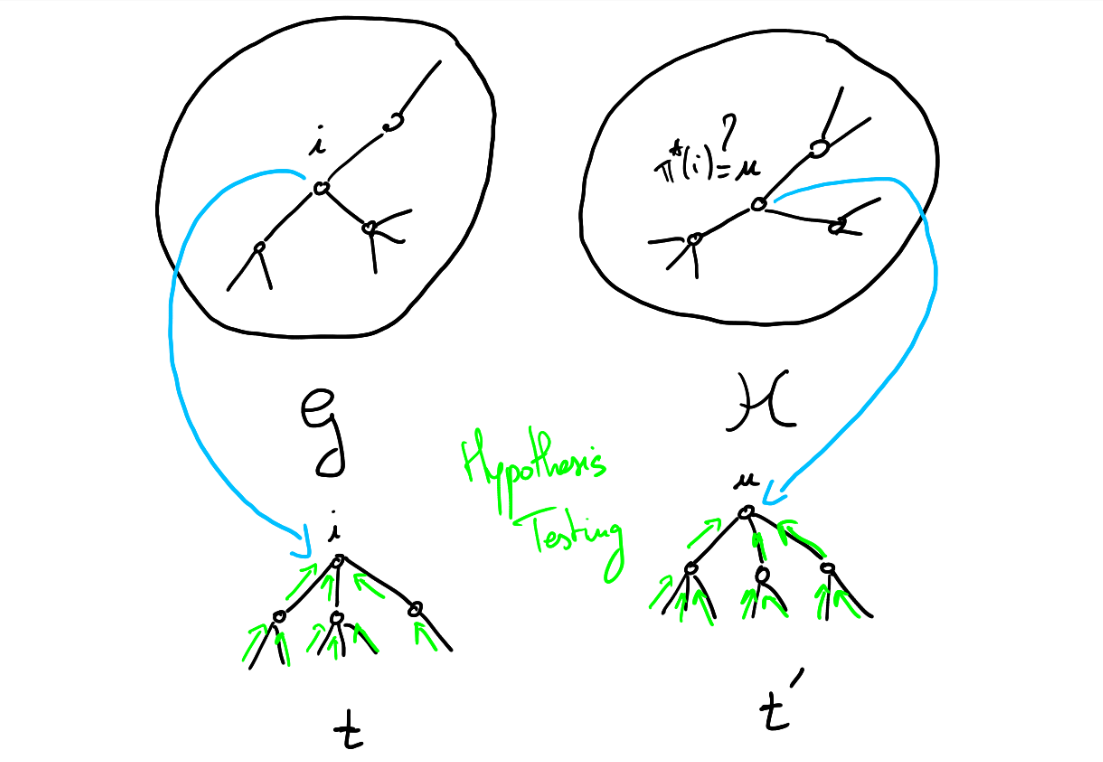
A (local) algorithm



A (local) algorithm



A (local) algorithm



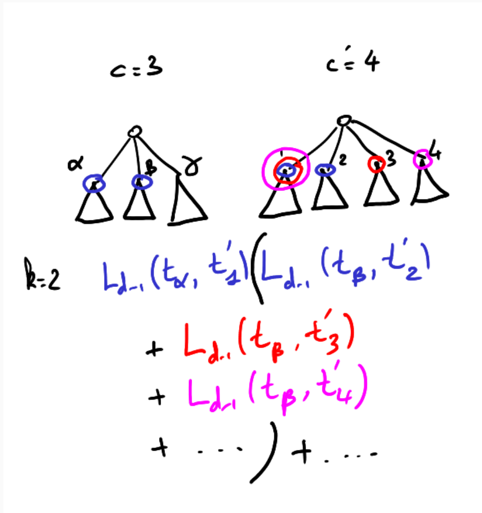
From graphs to trees

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 $\mathbb{P}_{0,d}$?

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 $[\ell]$.

Correlation detection in trees

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* $\Omega\left(\exp\left(\Omega(1) \times (\lambda s)^d\right)\right)$.

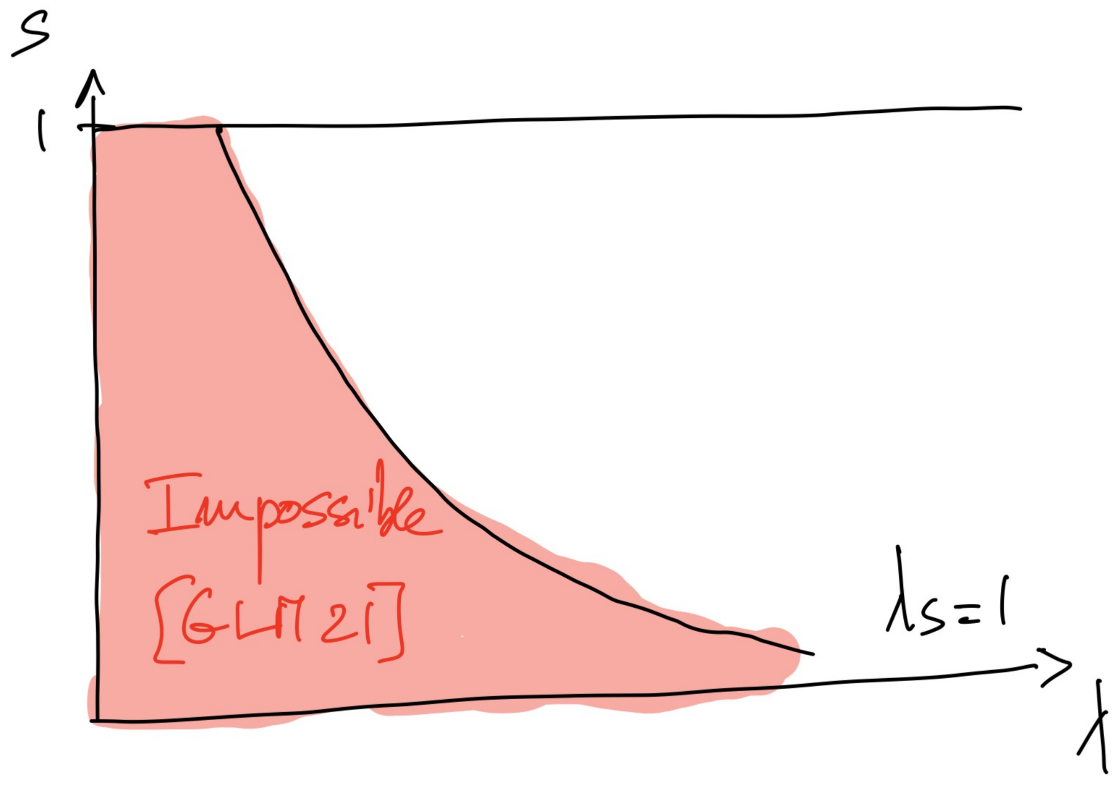
One-sided partial graph alignement

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

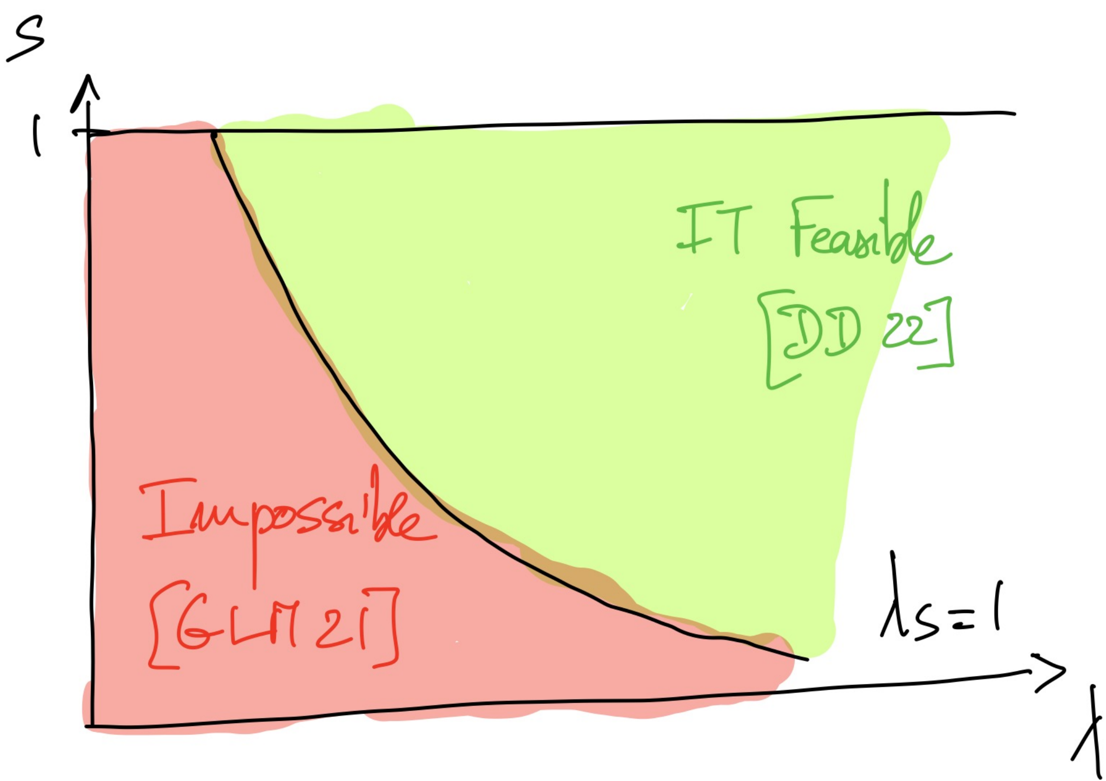
- *Partial recovery* if there exists some $\epsilon > 0$ such that $\mathbb{P}(\mathrm{ov}(\pi^*, \hat{\sigma}) > \epsilon) \underset{n\to\infty}{\longrightarrow} 1$,
- *One-sided partial recovery* if it achieves partial recovery and $\mathbb{P}(\mathrm{err}(\pi^*, \hat{\sigma}) = o(1)) \underset{n\to\infty}{\longrightarrow}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

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 algorithm.

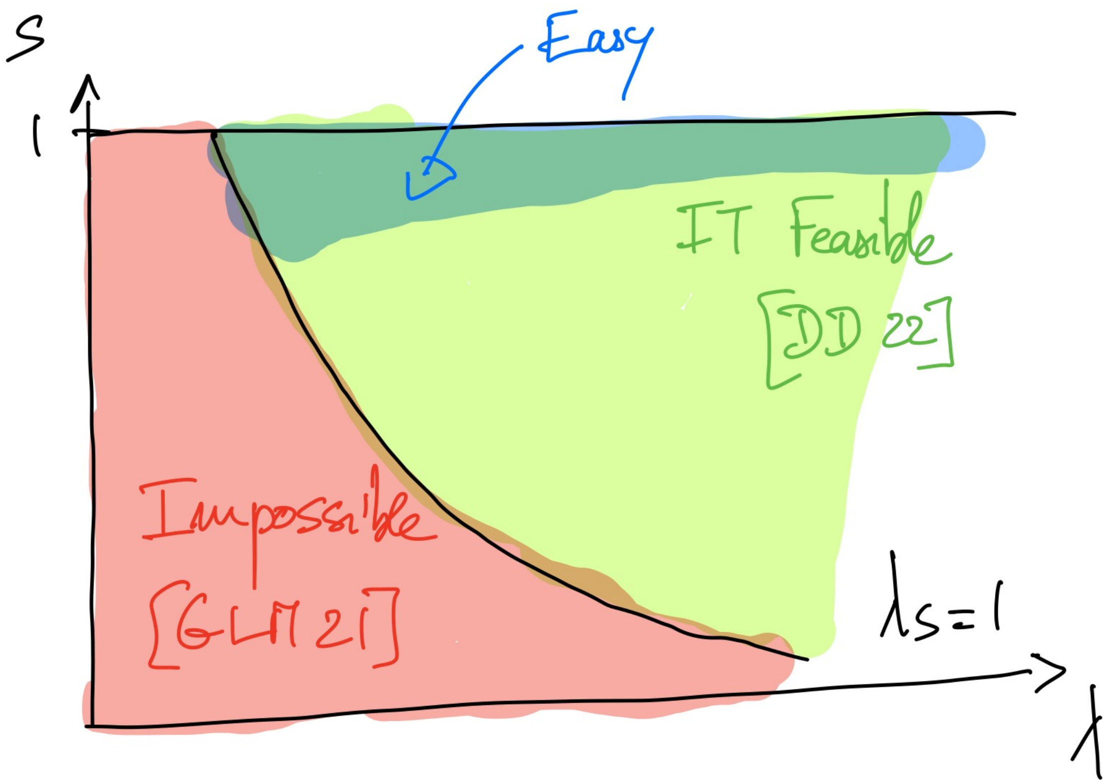
Results in the regime with constant mean degree



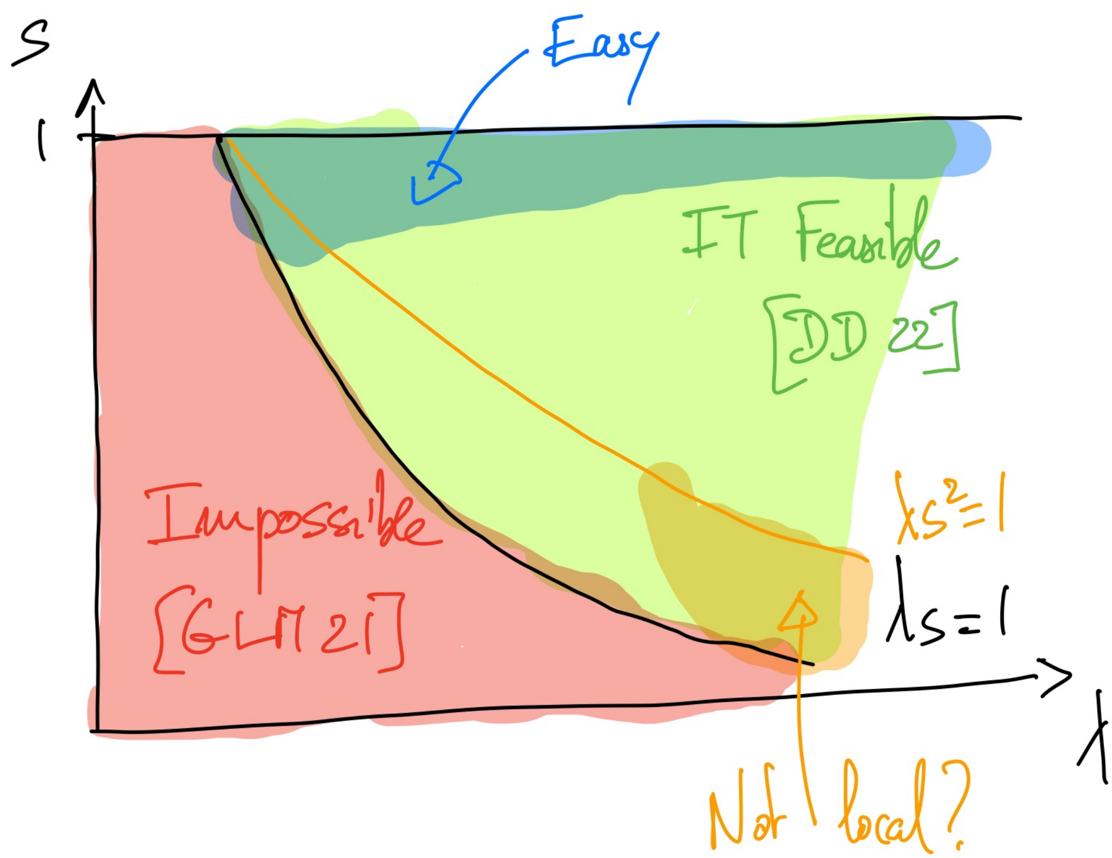
Results in the regime with constant mean degree



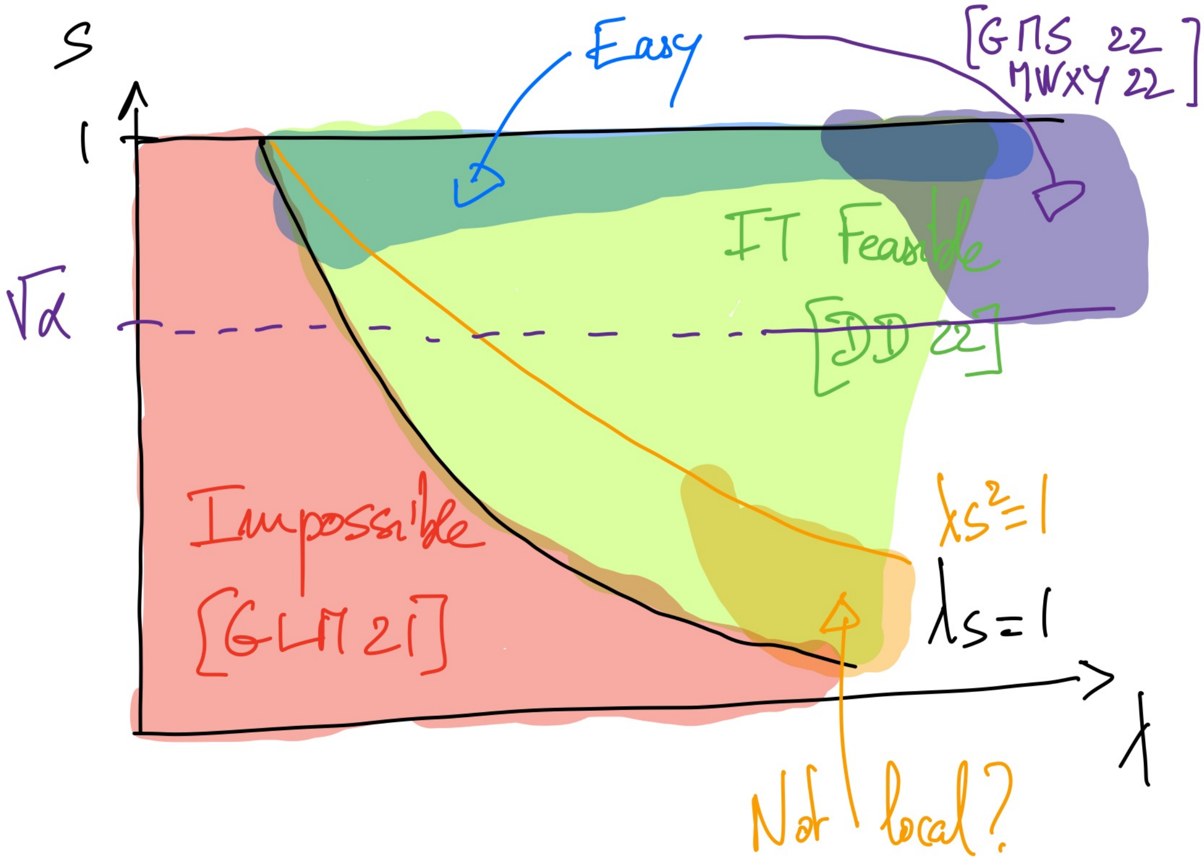
Results in the regime with constant mean degree



Results in the regime with constant mean degree

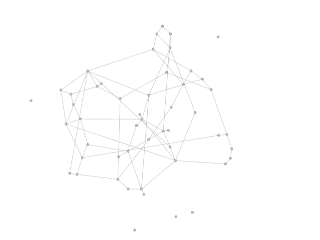


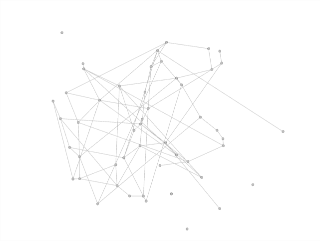
Results in the regime with constant mean degree



A learning algorithm

Learning the graph alignment problem with Siamese FGNNs





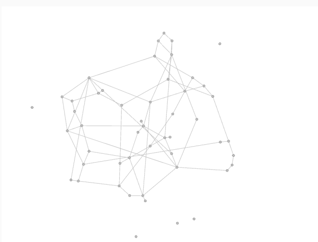
$$G_1 \in \{0,1\}^{n^2} \xrightarrow{\text{FGNN}} E_1 \in \mathbb{R}^{n \times b}$$

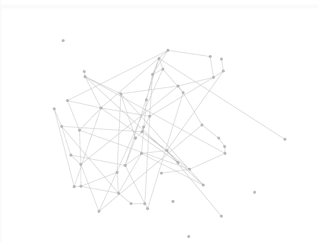
$$G_2 \in \{0,1\}^{n^2} \xrightarrow{\text{FGNN}} E_2 \in \mathbb{R}^{n \times b}$$

$$E_1 E_2^T \in \mathbb{R}^{n^2}$$

- The same FGNN is used for both graphs.

Learning the graph alignment problem with Siamese FGNNs





$$G_1 \in \{0,1\}^{n^2} \xrightarrow{\text{FGNN}} E_1 \in \mathbb{R}^{n \times b}$$

$$G_2 \in \{0,1\}^{n^2} \xrightarrow{\text{FGNN}} E_2 \in \mathbb{R}^{n \times b}$$

$$\to E_1 E_2^T \in \mathbb{R}^{n^2}$$

- The same FGNN is used for both graphs.
- We added a LAP solver to get a permutation from $E_1 E_2^T$.

Challenges in designing a learning algorithm for graphs

For a permutation $\sigma \in \mathcal{S}_n$, for $G \in \mathbb{F}^{n \times n}$($\mathbb{F} = \mathbb{R}^p$ feature space), we define : $(\sigma \star G)_{\sigma(i_1),\sigma(i_2)} = G_{i_1,i_2}$.

G_1, G_2 are isomorphic iff $G_1 = \sigma \star G_2$.

Challenges in designing a learning algorithm for graphs

For a permutation $\sigma \in \mathcal{S}_n$, for $G \in \mathbb{F}^{n \times n}$($\mathbb{F} = \mathbb{R}^p$ feature space), we define : $(\sigma \star G)_{\sigma(i_1),\sigma(i_2)} = G_{i_1,i_2}$.

G_1, G_2 are isomorphic iff $G_1 = \sigma \star G_2$.

Equivariance :
A function $f : \mathbb{F}^{n^2} \to \mathbb{F}^n$ is said to be equivariant if $f(\sigma \star G) = \sigma \star f(G)$.

For the graph alignment problem, we used an equivariant GNN from $\{0,1\}^{n \times n}$ to $\mathbb{F}^n$.

Challenges in designing a learning algorithm for graphs

For a permutation $\sigma \in \mathcal{S}_n$, for $G \in \mathbb{F}^{n\times n}$($\mathbb{F} = \mathbb{R}^p$ feature space), we define : $(\sigma \star G)_{\sigma(i_1),\sigma(i_2)} = G_{i_1,i_2}$.

G_1, G_2 are isomorphic iff $G_1 = \sigma \star G_2$.

Equivariance :
A function $f : \mathbb{F}^{n^2} \to \mathbb{F}^n$ is said to be equivariant if $f(\sigma \star G) = \sigma \star f(G)$.

For the graph alignment problem, we used an equivariant GNN from $\{0,1\}^{n\times n}$ to $\mathbb{F}^n$.

Expressiveness : Azizian and Lelarge (2020)
FGNN has the best power of approximation among all architectures working with tensors of order 2 (MGNN or LGNN).

Better expressive power with FGNN

(Maron et al., 2019) to propose the folklore graph layer (FGL) :

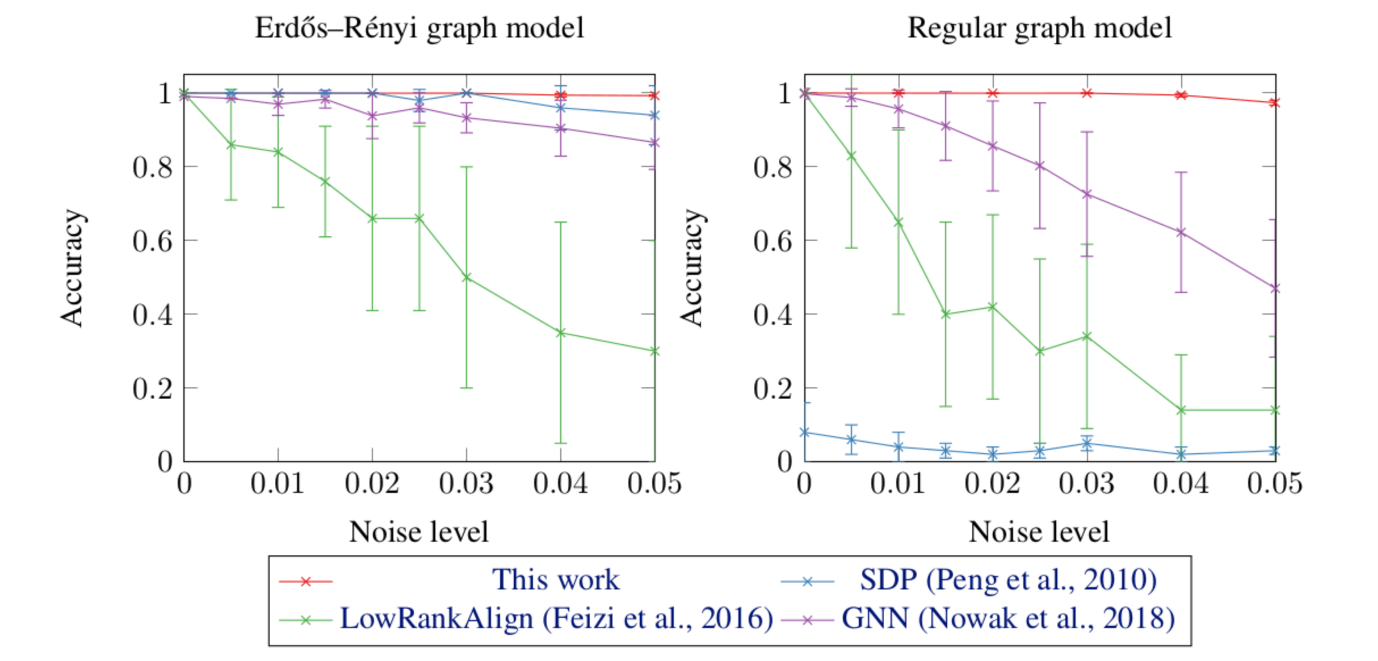
$$h_{i\to j}^{\ell+1} = f_0\left(h_{i\to j}^{\ell}, \sum_{k\in V} f_1\left(h_{i\to k}^{\ell}\right) f_2\left(h_{k\to j}^{\ell}\right)\right),$$

where f_0, f_1 and f_2 are learnable functions.

For FGNNs, messages are associated with pairs of vertices as opposed to MGNN where messages are associated with vertices.

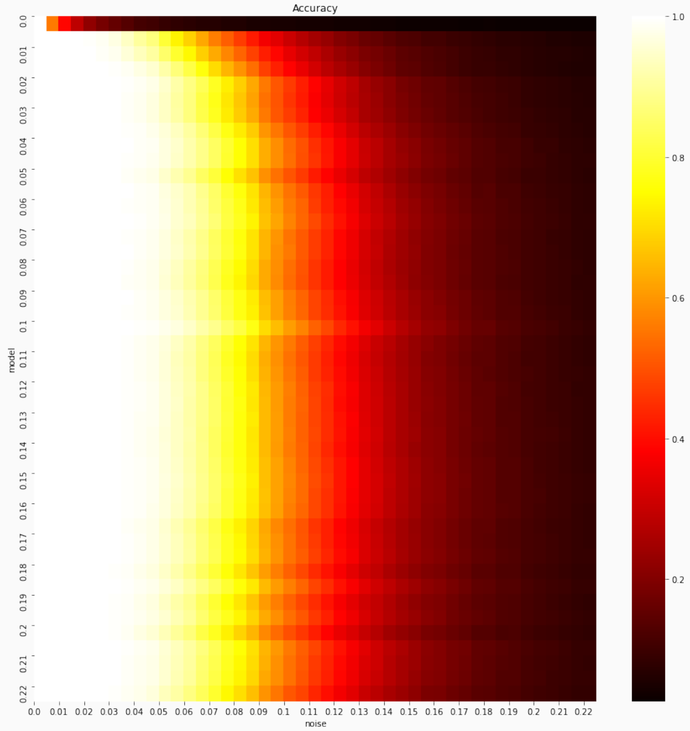
FGNN : a FGNN is the composition of FGLs and a final invariant/equivariant reduction layer from $\mathbb{F}^{n^2}$ to $\mathbb{F}/\mathbb{F}^n$.

Results on synthetic data



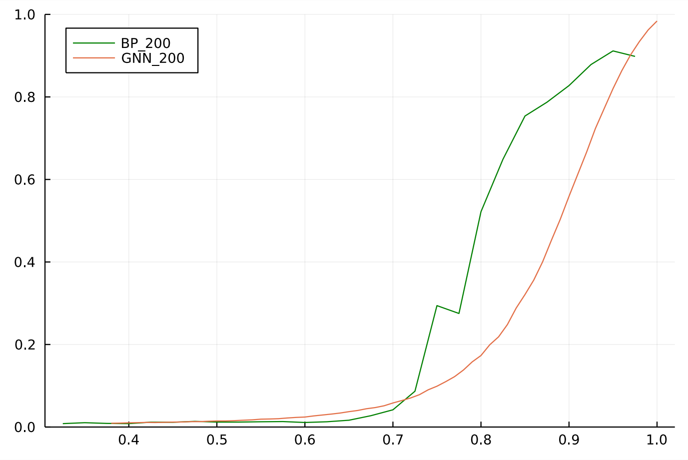
- Graphs : $n = 50$, density = 0.2
- Training set : 20000 samples
- Validation and Test sets : 1000 samples

Generalization for regular graphs



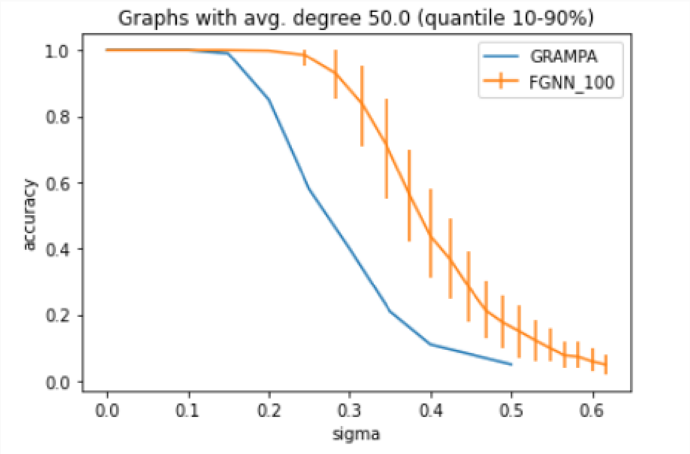
Each line corresponds to a model trained at a given noise level and shows its accuracy across all noise levels.

Comparison FGNN vs BPAlign



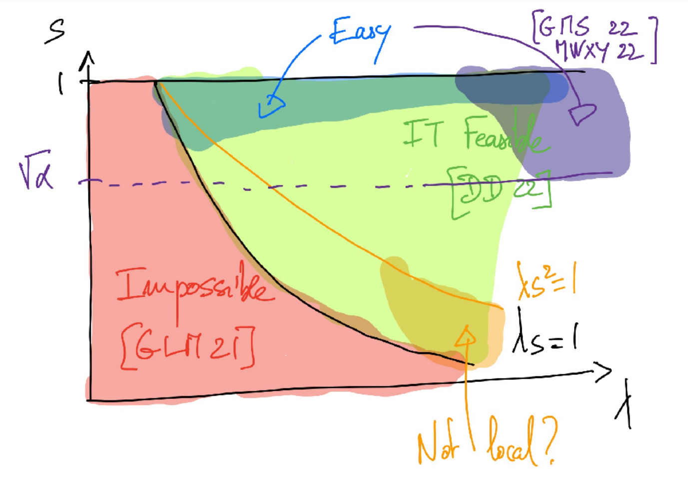
Overlap as a function of the correlation s for correlated E-R with average degree 3 (number of nodes : 200).

Comparison FGNN vs GRAMPA

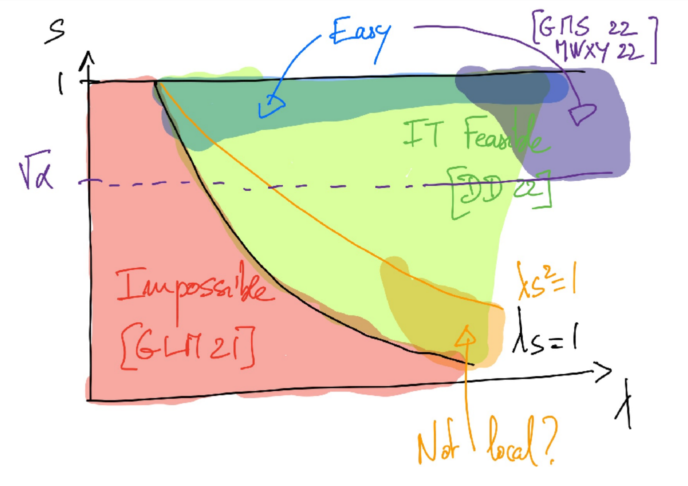


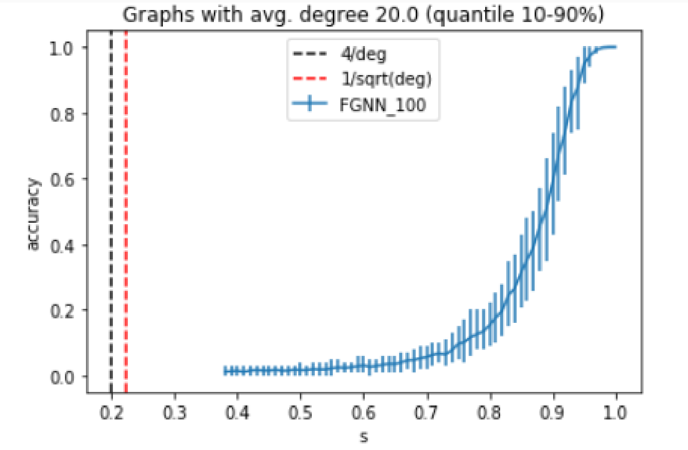
Overlap as a function of $\sigma = \sqrt{\frac{1-s}{1-\lambda/n}}$ for correlated E-R with average degree 50 (number of nodes : 100). Fan et al. (2019)

From theory to practice

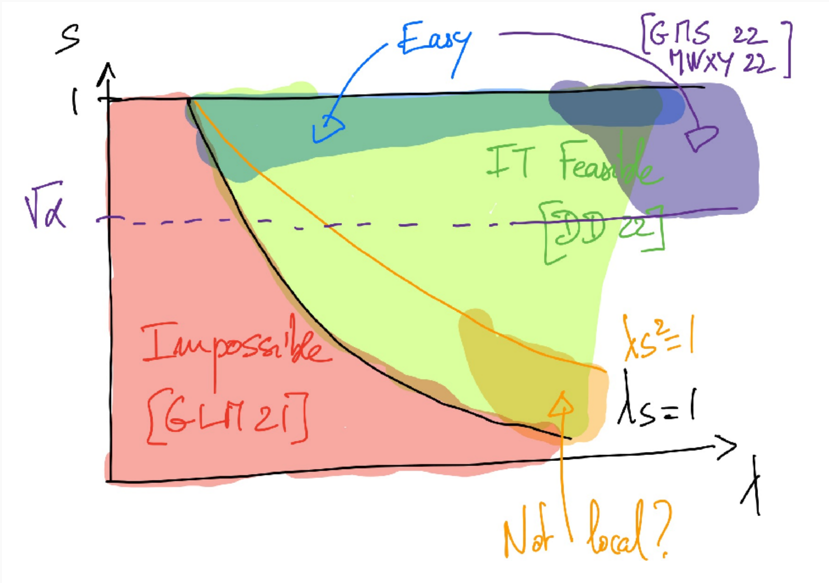


From theory to practice





Conclusion



The Bitter Lesson by Rich Sutton

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.

Thank You!

Références

W. Azizian and M. Lelarge. Expressive power of invariant and equivariant graph neural networks. *arXiv preprint arXiv :2006.15646*, 2020.

D. Cullina and N. Kiyavash. Exact alignment recovery for correlated Erdős-Rényi graphs, 2017.

J. Ding and H. Du. Matching recovery threshold for correlated random graphs. *arXiv preprint arXiv :2205.14650*, 2022.

J. Ding, H. Du, and S. Gong. A polynomial-time approximation scheme for the maximal overlap of two independent erd\h {o} sr\'enyi graphs. *arXiv preprint arXiv :2210.07823*, 2022.

Z. Fan, C. Mao, Y. Wu, and J. Xu. Spectral graph matching and regularized quadratic relaxations i : The gaussian model. *arXiv preprint arXiv :1907.08880*, 2019.

L. Ganassali, L. Massoulié, and M. Lelarge. Impossibility of partial recovery in the graph alignment problem. *arXiv preprint arXiv :2102.02685*, 2021.

L. Ganassali, L. Massoulié, and G. Semerjian. Statistical limits of correlation detection in trees. *arXiv preprint arXiv :2209.13723*, 2022.

C. Mao, Y. Wu, J. Xu, and S. H. Yu. Random graph matching at otter's threshold via counting chandeliers. *arXiv preprint arXiv :2209.12313*, 2022.

H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman. Provably powerful