Sharp threshold for alignment of graph databases with Gaussian weights

Luca Ganassali

Mathematical and Scientific Machine Learning, August 16-19th, 2021

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

ÉCOLE NORMALE SUPÉRIEURE
RESEARCH UNIVERSITY PARIS

Graph/matrix 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'?*

Graph/matrix 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

$$\max_{\Pi} \langle G, \Pi G' \Pi^\top \rangle,$$

where Π runs over all permutation matrices.

Graph/matrix 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

$$\max_{\Pi} \langle G, \Pi G' \Pi^\top \rangle,$$

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

Planted Alignment with gaussian weights

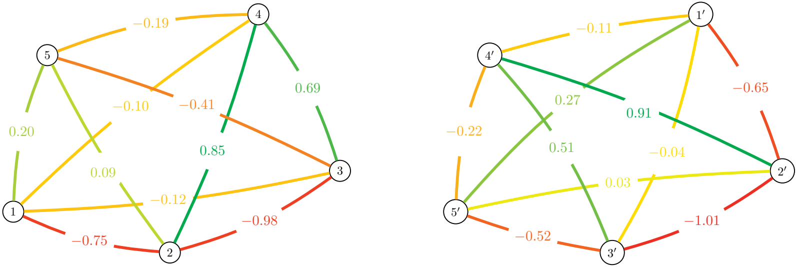
Correlated Wigner model:

- Draw the planted permutation π^* uniformly at random in $\mathcal{S}_n$.
- $(A_{i,j}, B_{\pi^*(i),\pi^*(j)})_{1\leq i<j\leq n}$ are i.i.d. $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho \in [0,1]$.

In other words:

$$B = \rho \cdot {\Pi^*}^T A \Pi^* + \sqrt{1-\rho^2} \cdot H,$$

where H is an independent copy of A, and $\Pi^*_{i,j} = \mathbf{1}_{j=\pi^*(i)}$.



Planted Alignment with gaussian weights

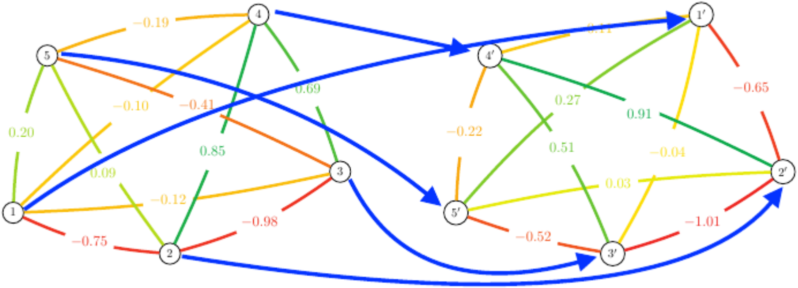
Correlated Wigner model:

- Draw the planted permutation π^* uniformly at random in $\mathcal{S}_n$.
- $(A_{i,j}, B_{\pi^*(i),\pi^*(j)})_{1\leq i<j\leq n}$ are i.i.d. $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho \in [0,1]$.

In other words:

$$B = \rho \cdot {\Pi^*}^T A \Pi^* + \sqrt{1-\rho^2} \cdot H,$$

where H is an independent copy of A, and $\Pi^*_{i,j} = \mathbf{1}_{j=\pi^*(i)}$.



MAP estimation

$$
\begin{aligned}
p_{\pi^*|A,B}(\pi|a,b) &\propto p_{\pi^*,A,B}(\pi,a,b) \\
&\propto \exp\left(-\frac{1}{2(1-\rho^2)}\sum_{1\le i<j\le n}\left(B_{\pi(i),\pi(j)}-\rho A_{i,j}\right)^2\right),
\end{aligned}
$$

where $\propto$ indicates equality up to some factors that do not depend on σ. The MAP estimator is given by

$$\hat{\pi}_{\mathrm{MAP}} := \arg\max_{\pi} p_{\pi^*|A,B}(\pi|A,B) = \arg\max\langle A, \Pi B \Pi^T\rangle.$$

Main result: sharp threshold for exact alignment

Theorem (Achievability part)

If for n large enough

$$\rho^2 \geq \frac{(4+\varepsilon)\log n}{n} \tag{1}$$

for some $\varepsilon > 0$, then there is an estimator (namely, the MAP estimator) $\hat{\pi}$ of π given A, B such that $\hat{\pi} = \pi^*$ with probability $1 - o(1)$.

Theorem (Converse part)

Conversely, if

$$\rho^2 \leq \frac{4\log n - \log\log n - \omega(1)}{n} \tag{2}$$

then any estimator $\hat{\pi}$ of π given A, B verifies $\hat{\pi} = \pi^*$ with probability $o(1)$.

Proof outline

- Achievability result: analysis of the MAP estimator

$$\hat{\pi}_{\mathrm{MAP}} = \arg\min_{\pi} \mathcal{L}(\pi, A, B),$$

with

$$\mathcal{L}(\pi, A, B) := \sum_{1 \le i < j \le n} \left(B_{\pi(i),\pi(j)} - \rho A_{i,j} \right)^2 .$$

We show that $\hat{\pi}_{\mathrm{MAP}} = \pi^*$ with high probability whenever $\rho^2 \geq \frac{(4+\varepsilon)\log n}{n}$.
First moment method fails because of correlation.

Proof outline

- Achievability result: analysis of the MAP estimator

$$\hat{\pi}_{\mathrm{MAP}} = \arg\min_{\pi} \mathcal{L}(\pi, A, B),$$

with

$$\mathcal{L}(\pi, A, B) := \sum_{1 \le i < j \le n} \left(B_{\pi(i),\pi(j)} - \rho A_{i,j} \right)^2 .$$

We show that $\hat{\pi}_{\mathrm{MAP}} = \pi^*$ with high probability whenever $\rho^2 \ge \frac{(4+\varepsilon)\log n}{n}$.
First moment method fails because of correlation.

- Converse result: we show that when $\rho^2 \le \frac{4\log n - \log\log n - \omega(1)}{n}$, w.h.p. there exists a perturbation of π^* (namely $\pi^* \circ \tau$ for some transposition τ) s.t. $\mathcal{L}(\pi^* \circ \tau, A, B) < \mathcal{L}(\pi^*, A, B)$.

Linear Assignment problem: $\pi^* \sim U(\mathcal{S}_N)$ and u, v are random vectors such that $(u_i, v_{\pi^*(i)})_{1 \le i \le n}$ are i.i.d. $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho \in [0, 1]$.

Comparison with vector alignment

Linear Assignment problem: $\pi^* \sim U(\mathcal{S}_N)$ and u, v are random vectors such that $(u_i, v_{\pi^*(i)})_{1 \le i \le n}$ are i.i.d. $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho \in [0, 1]$.

MAP estimator:

$$\arg\max_{\pi} \langle u, \Pi v \rangle.$$

Comparison with vector alignment

Linear Assignment problem: $\pi^* \sim U(\mathcal{S}_N)$ and u, v are random vectors such that $(u_i, v_{\pi^*(i)})_{1 \le i \le n}$ are i.i.d. $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho \in [0, 1]$.

MAP estimator:

$$\arg\max_{\pi} \langle u, \Pi v \rangle.$$

(Dai-Cullina-Kiyavash '19): sharp threshold for exact recovery at $-\frac{1}{2}\log(1-\rho^2) \gtrsim 2\log N$, for $N = n(n-1)/2$, i.e. at

$$1 - \rho^2 \lesssim n^{-8}.$$

Comparison with vector alignment

Linear Assignment problem: $\pi^* \sim U(\mathcal{S}_N)$ and u, v are random vectors such that $(u_i, v_{\pi^*(i)})_{1 \le i \le n}$ are i.i.d. $\mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho \in [0,1]$.

MAP estimator:

$$\arg\max_{\pi} \langle u, \Pi v \rangle.$$

(Dai-Cullina-Kiyavash '19): sharp threshold for exact recovery at $-\frac{1}{2}\log(1-\rho^2) \gtrsim 2\log N$, for $N = n(n-1)/2$, i.e. at

$$1-\rho^2 \lesssim n^{-8}.$$

$\to$ vector alignment (resp. LAP) is a very bad relaxation of matrix alignment (resp. QAP).

State-of-the art algorithms for (almost) exact recovery

- Degree profiles (Ding-Ma-Wu-Xu 18'), spectral method (Fan-Mao-Wu-Xu 19') with time complexity $\mathcal{O}(n^3)$ requires

$$\sqrt{1-\rho^2} \leq \mathcal{O}\left(\log^{-1} n\right).$$

- A simpler spectral method with complexity $\mathcal{O}(n^2)$ (G-Massoulié-Lelarge 19') requires

$$\sqrt{1-\rho^2} \leq \mathcal{O}\left(n^{-7/6}\right).$$

State-of-the art algorithms for (almost) exact recovery

- Degree profiles (Ding-Ma-Wu-Xu 18'), spectral method (Fan-Mao-Wu-Xu 19') with time complexity $\mathcal{O}(n^3)$ requires

$$\sqrt{1-\rho^2} \leq \mathcal{O}\left(\log^{-1} n\right).$$

- A simpler spectral method with complexity $\mathcal{O}(n^2)$ (G-Massoulié-Lelarge 19') requires

$$\sqrt{1-\rho^2} \leq \mathcal{O}\left(n^{-7/6}\right).$$

In any case, ρ needs to tend to 1 : very far from the informational threshold at $\frac{n\rho^2}{\log n} \sim 4$

State-of-the art algorithms for (almost) exact recovery

- Degree profiles (Ding-Ma-Wu-Xu 18'), spectral method (Fan-Mao-Wu-Xu 19') with time complexity $\mathcal{O}(n^3)$ requires

$$\sqrt{1-\rho^2} \leq \mathcal{O}\left(\log^{-1} n\right).$$

- A simpler spectral method with complexity $\mathcal{O}(n^2)$ (G-Massoulié-Lelarge 19') requires

$$\sqrt{1-\rho^2} \leq \mathcal{O}\left(n^{-7/6}\right).$$

In any case, ρ needs to tend to 1 : very far from the informational threshold at $\frac{n\rho^2}{\log n} \sim 4$

$\longrightarrow$ *hard phase* conjectured to be really wide for this reconstruction problem.

Thank you!