Asymmetric tree correlation testing for graph alignment

Jakob Maier, Laurent Massoulié

Asymmetric tree correlation testing for graph alignment

Jakob Maier
INRIA, DI/ENS, PSL
Paris, France
jakob.maier@inria.fr

Laurent Massoulié
INRIA, DI/ENS, PSL, MSR-INRIA joint centre
Paris, France
laurent.massoulie@inria.fr

***Abstract*—We consider the partial graph alignment problem on two correlated sparse Erdős–Rényi graphs with differing edge or node densities. Exploiting that these graphs are locally tree-like, we come to consider a hypothesis testing problem on correlated Galton-Watson trees. To solve this problem, we give several equivalent conditions for the existence of likelihood-ratio tests with vanishing type-I-error and significant power. We then show that these same conditions enable the partial graph alignment algorithm *MPAlign* to succeed.**
This paper generalizes recent results from Ganassali L., Massoulié L. and Lelarge M. to the asymmetric edge and node density case. This extension allows for greater applicability of the results and resolves a special case of the subgraph isomorphism problem.

***Index Terms*—graph alignment, sparse Erdős–Rényi graphs, tree correlation testing**

I. Introduction

Given two correlated graphs without node labels, the graph alignment problem (or graph matching problem) consists of finding an *optimal* mapping between the two graph's node sets. This problem has received considerable attention in the last decade due to its various applications, including de-anonymization of social network data [1], and comparing protein-protein interaction graphs of different species [2].

There are different ways of formalizing the graph alignment problem and we focus on the *planted* version where the two correlated graphs are issued from a generative random graph model with a latent true node matching. The simplest and most well-studied random graphs are so-called Erdős–Rényi graphs with parameters $n \in \mathbb{N}$ and $p_n \in [0,1]$: To sample $G_0 \sim \mathrm{ER}(n, p_n)$, one fixes n nodes and draws an edge between each pair of nodes independently with probability p_n. Two correlated Erdős–Rényi graphs are typically obtained by subsampling twice independently from G_0.

This model has first been studied in [3] and recent results sharply quantify the information theoretic limits for recovering the latent node matching. In [4], the authors establish exact information theoretic threshoIds for partial alignment if $p_n = n^{-\alpha+o(1)}$. In the case where $np_n \to \infty$, [5] presents a polynomial-time algorithm which asymptotically achieves full alignment.

This work was partially supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

This leaves the question open about a *polynomial time algorithm* for the *sparse case* $p_n = \lambda/n$, i.e. when the graphs have constant average degree $\lambda \in \mathbb{R}$. The works of Ganassali et.al. have examined this regime, especially [6] which this paper aims to generalize for asymmetrically correlated graphs.

A. Correlated graph model

To specify what we mean by asymmetrically correlated graphs and their latent matching, we fix parameters $s, s' \in (0,1]$ and sample an Erdős–Rényi graph $G_0 \sim \mathrm{ER}(n, \lambda/n)$ whose nodes $V(G_0)$ receive labels in $[n] := \{1, ..., n\}$. Now we sample two subgraphs of G_0, using two distinct procedures:

1) *Varying edge densities*: For every edge in G_0, independently delete it with probability $1-s$ to obtain G. Then, restarting with G_0, repeat the deletion process independently with probability $1-s'$ to obtain $\tilde{G}$.
2) *Differring node numbers:* For every node in G_0 independently, delete it and its adjacent edges with probability $1-s$ to obtain G. Then, restarting with G_0, repeat the process independently with probability $1-s'$ to get $\tilde{G}$.

After either subsampling procedure, let σ^* be a random permutation of $[n]$ which induces a map $\sigma^* : V^* \to [n]$ where $V^* := V(\tilde{G})$ denotes the set of true labels of the second graph. This σ^* is used to shuffle the nodes in $\tilde{G}$ to obtain a randomly labeled graph G'. We call the couple (G, G') *correlated Erdős–Rényi graphs (with varying edge densities / with differing node numbers).*

B. Goal: One-sided partial graph alignment

Given correlated Erdős–Rényi graphs (G, G'), the goal is to recover $\sigma^* : V^* \to [n]$ partially, i.e. retrieve a significant percentage of node correspondances. As pointed out in [6], we cannot retreive the full permutation σ^*, one reason being isolated nodes in G and G' that cannot be correctly matched.

To define partial alignment, we need the notions of overlap and error fraction. For a subset $\hat{V} \subset V(G)$ of the nodes of G and a matching estimator $\hat{\sigma} : \hat{V} \to V(G')$, we define

$$\mathrm{ov}(\sigma^*, \hat{\sigma}) := \frac{1}{|V^*|} \sum_{i \in \hat{V} \cap V^*} \mathbb{1}_{\hat{\sigma}(i) = \sigma^*(i)}, \text{ and}$$

$$\mathrm{err}(\sigma^*, \hat{\sigma}) := \frac{1}{|V(G)|} \sum_{i \in \hat{V}} \mathbb{1}_{i \notin V^* \text{ or } \hat{\sigma}(i) \neq \sigma^*(i)}.$$

A high overlap $\mathrm{ov}(\sigma^*, \hat{\sigma})$ indicates a powerful matching procedure but ignores the number of falsely matched nodes. Therefore, it is essential to control the error fraction $\mathrm{err}(\sigma^*, \hat{\sigma})$ which lets us define *good* estimators as follows:

Definition 1. A sequence of alignment estimators $\hat{\sigma} = (\hat{\sigma}_n)_n$ achieves one-sided partial alignment if there is $\varepsilon > 0$ such that

$$\mathbb{P}\left(\mathrm{ov}(\sigma^*, \hat{\sigma}) \geq \varepsilon\right) \xrightarrow{n\to\infty} 1, \text{ and}$$
$$\mathbb{P}\left(\mathrm{err}(\sigma^*, \hat{\sigma}) = o(1)\right) \xrightarrow{n\to\infty} 1.$$

Our goal in the following is to find an algorithm and sufficient conditions on λ, s, s' to enable one-sided partial alignment.

C. Notations

• For a graph G, one of its nodes $i \in V(G)$ and a distance $d \in \mathbb{N}$, we denote the depth-d neighborhood of i in G by

$$\mathcal{N}_G(i, d) := \{j \in V(G) \,:\, \mathrm{dist}_G(i, j) \leq d\}.$$

• Recall that the Poisson-distribution $\mathrm{Poi}(\mu)$ is defined for $k = 0, 1, 2, ...$ via

$$\pi_\mu(k) := \mathbb{P}(X = k) = e^{-\mu}\frac{\mu^k}{k!} \quad \text{for } X \sim \mathrm{Poi}(\mu).$$

• A Galton Watson tree t with $\mathrm{Poi}(\mu)$-offspring is iteratively defined to be a root node with $X_0 \sim \mathrm{Poi}(\mu)$ children $i_1, ..., i_{X_0}$, which all independently have $X_{i_j} \sim \mathrm{Poi}(\mu)$ children and so forth. We write $t \sim \mathrm{GW}_\mu$ and denote by $\mathrm{GW}_\mu(\tau)$ the likelihood of a fixed tree τ being of law GW_μ. Furthermore, we use $\mathrm{Ext}(\mathrm{GW}_\mu)$ to denote the event of extincion, i.e. that a tree of law GW_μ has finite depth.

• Let $\mathcal{X}_d$ be the set of all unlabeled rooted trees with depth at most d. Rigorously, one may define $t \in \mathcal{X}_d$ recursively as a tuple $(k_\tau)_{\tau\in\mathcal{X}_{d-1}} \in \mathbb{N}^{\mathcal{X}_{d-1}}$ where each coordinate k_τ counts how many child-trees of t's root are equal to τ.

• Let τ be a rooted tree. For a node i of τ, let $c_\tau(i)$ be the number of its children. Let $\mathcal{V}_d(\tau)$ denote the vertices up to depth d and write $\mathcal{L}_d(\tau) = \mathcal{V}_d(\tau) \setminus \mathcal{V}_{d-1}(\tau)$ for the d^{th} generation of τ.

II. From sparse graphs to tree testing.

The key idea in the sparse setting is that *sparse Erdős–Rényi graphs are locally tree-like*. The following lemma, which combines lemmata 2.2 and 2.4 in [8], makes this idea rigorous.

Lemma 1. *Let $G \sim \mathrm{ER}(n, \lambda/n)$ and $d := \lfloor c\log(n)\rfloor$ for some c fulfilling $c\,\log(\lambda) < 1/2$. Then there exists $\varepsilon > 0$ such that for all nodes i of G, one has*

$$\mathbb{P}(\mathcal{N}_G(i, d) \textit{ contains cycles (i.e. is not a tree)}\ ) = \mathcal{O}(n^{-\varepsilon}).$$

Furthermore, if we compare the law of $\mathcal{N}_{G,d}(i)$ with GW_λ truncated to depth d, their total variation distance is $\mathcal{O}(n^{-\varepsilon})$.

Determining whether two nodes $i \in V(G)$ and $j \in V(G')$ should be matched therefore translates to whether the trees $\mathcal{N}_G(i, d)$ and $\mathcal{N}_{G'}(j, d)$ are correlated. We will see that this translates to a tree correlation testing problem with the following hypotheses:

a) Independent model $\mathbb{P}_0$: Let τ, τ' be independent $\mathrm{GW}_{\lambda s}$ and $\mathrm{GW}_{\lambda s'}$ trees. Obtain t, t' by randomly relabeling both trees. This defines $\mathbb{P}_0 \sim (t, t') \sim \mathrm{GW}_{\lambda s} \times \mathrm{GW}_{\lambda s'}$.

b) Correlated model $\mathbb{P}_1$: We introduce a model for correlated trees by viewing them as augmentations of their intersection tree τ^*:

Definition 2. For $\lambda > 0$, $s, s' \in (0, 1)$ and a tree $\tau = (V, E)$ define its (λ, s, s')-augmentation $A_{\lambda,s,s'}(\tau)$ as follows:

1) For each $i \in V$ draw $c^+(i) \sim \mathrm{Poi}(\lambda s(1 - s')))$ independently and attach $c^+(i)$ additional children to i.
2) Attach an independent $\mathrm{GW}_{\lambda s}$ tree to each new child added in the first step.

To sample from the distribution $\mathbb{P}_1$, one starts with an intersection tree $\tau^* \sim \mathrm{GW}_{\lambda ss'}$ and augment it twice independently with different parameters: first with (λ, s, s'), then with (λ, s', s). This yields $\tau = A_{\lambda,s,s'}(\tau^*)$ and $\tau' = A_{\lambda,s',s}(\tau^*)$. Relabeling τ and τ', we obtain t, t' and denote their joint distribution as $\mathbb{P}_1$.

For $i = 0, 1$ and random trees $(t, t') \sim \mathbb{P}_i$, we write (t_d, t'_d) for their depth-d-truncated versions and denote by $\mathbb{P}_{i,d}$ the law of (t_d, t'_d). Given a pair of trees (τ, τ'), we denote by $\mathbb{P}_{i,d}(\tau, \tau')$ the likelihood of them being sampled from $\mathbb{P}_i$ up to depth d.

Note that under both models $\mathbb{P}_0$ and $\mathbb{P}_1$, the pair (t, t') has the same marginal distributions as the local neighboorhoods in correlated Erdős–Rényi graphs G and G'. Remarkably, this is true in the varying edge densities as well as the differing node numbers framework since deleting a node in a tree is equivalent to deleting the edge leading to that node.

Having translated the initial problem to a testing problem, it remains to find an equivalent notion for one-sided alignment:

Definition 3. An asymptotic one-sided test is a sequence of functions $\mathcal{T}_d : \mathcal{X}_d \times \mathcal{X}_d \to \{0, 1\}$ fulfilling

- $\mathbb{P}_{0,d}\left(\mathcal{T}_d(t_d, t'_d) = 1\right) \xrightarrow{d\to\infty} 0$ (vanishing type-I-error),
- $\liminf_{d\to\infty} \mathbb{P}_{1,d}(\mathcal{T}_d = 1) \geq \varepsilon > 0$ (significant power).

We will see that one-sided testability enables one-sided alignment.

III. Conditions for one-sided testability

The goal of this section is to point out conditions which are equivalent to one-sided testability, more easily understandable and which yield an efficiently computable test for $\mathbb{P}_0$ vs. $\mathbb{P}_1$.

A natural quantity to examine in our testing problem is the likelihood ratio at depth d for a pair of trees (t, t'):

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

Note that L_d enables a simple notation for the Kullback-Leibler divergence at depth d:

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

Using Jensen's inequality and the convexity of $x \mapsto x\log(x)$, one can easily show that the sequence $(\mathrm{KL}_d)_d$ is increasing and therefore admits a limit $\mathrm{KL}_\infty \in [0, \infty]$.

With these notations, we can state our main theorem:

Theorem 1. *In the above hypothesis test setting, the following are equivalent:*

(i) There exists a one-sided test to decide $\mathbb{P}_0$ vs. $\mathbb{P}_1$,
(ii) There exists $a_d \to \infty$ such that $\mathbb{P}_{0,d}(L_d > a_d) \to 0$ and $\liminf_d \mathbb{P}_{1,d}(L_d > a_d) > 0$.
(iii) $\lambda s s' > 1$ and $KL_\infty = \infty$,
(iv) $\lambda s s' > 1$ and there is $C = C(\lambda, s, s') > 0$ such that

$$\mathbb{P}_1\left(\liminf_d (\lambda s s')^{-d} \log(L_d) \geq C\right) \geq 1 - \mathbb{P}(\mathrm{Ext}(\mathrm{GW}_{\lambda s s'}))$$

These equivalences are the result we were looking for: Point (ii) motivates the use of likelihood ratio tests while (iv) provides the necessary sequence $a_d = \exp(C(\lambda s s')^d)$. Additionally, future work may leverage (iii) to derive exact conditions on λ, s and s' for one-sided testability.

For the remainder of this section, we prove Thorem 1. This requires some technical results first.

A. Properties of the likelihoood ratio

A remarkable property of L_d in our setting is the following *recursive likelihood ratio formula* which will be useful for efficient computation:

Proposition 1. *Let $t, t' \in \mathcal{X}_d$ and denote their root children as $t_1, ..., t_c$ and $t'_1, ..., t'_{c'}$ respectively. Using $\mathfrak{S}_c$ to denote the set of permutations of $[c]$, we can express $(L_d)_d$ recursively:*

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

where

$$\psi(k, c, c') = \exp(\lambda s s') \frac{(1-s')^{c-k}(1-s)^{c'-k}}{\lambda^k \, k!(c-k)!(c'-k)!}.$$

A proof for this proposition can be found in the appendix. Taking conditional expectations in the recursive likelihood ratio formula and using independence of (c, c') from the pairs (t_i, t'_i), we obtain the following corollary. For details, one can effortlessly translate the proof of proposition 2.1 in [6].

Corollary 1. *The sequence $(L_d)_d$ is a martingale with respect to $\mathbb{P}_0$ and since $L_d \geq 0$ for all d, this martingale almost surely converges to a limit L_∞.*

Using this result, we can take the recursive likelihood ratio formula and let $d \to \infty$, then take $\mathbb{E}_0$ to obtain

$$\begin{aligned}
\mathbb{E}_0[L_\infty] &= \mathbb{E}_0\left[\sum_{k=0}^{c \wedge c'} \psi(k, c, c') \sum_{\substack{\sigma \in \mathfrak{S}_c, \\ \sigma' \in \mathfrak{S}_{c'}}} \prod_{i=1}^{k} \mathbb{E}_0\left[L_\infty \mid c, c'\right]\right] \\
&= \mathbb{E}_0\left[\sum_{k=0}^{\infty} \mathbb{1}_{k \leq c \wedge c'} \psi(k, c, c') c! \, c'! \, \mathbb{E}_0[L_\infty]^k\right] \\
&= \sum_{k=0}^{\infty} \pi_{\lambda s s'}(k) \mathbb{E}_0[L_\infty]^k
\end{aligned}$$

In other words, $\mathbb{E}_0[L_\infty]$ is a fixed point of the probability generating function of $\mathrm{Poi}(\lambda s s')$.

B. Proof of Theorem 1

The first equivalence can be shown quickly:

Proof of $(i) \iff (ii)$. We start by noting that $(ii) \implies (i)$ directly follows from the choice $\mathcal{T}_d = \mathbb{1}_{L_d > a_d}$. The reverse direction naturally uses the Neyman-Pearson lemma, but it is not immediate due to some technicalities. In particular, one needs to carefully handle the probabiliy of $\{L_d = a_d\}$. Due to lack of space, we refer to Step 3 in the proof of Theorem 1 in [6] for a complete argument. □

For the remaining points, we will show the circular implications $(iv) \implies (i) \implies (iii) \implies (iv)$.

Proof of $(iv) \implies (i)$. Point (iv) suggests the test sequence

$$\mathcal{T}_d = 1 \quad :\iff \quad L_d \geq \exp\left(C\left(\lambda s s'\right)^d\right).$$

Markov's inequality paired with $\lambda s s' > 1$ and $\mathbb{E}_0[L_d] = 1$ gives

$$\mathbb{P}_0(\mathcal{T}_d = 1) \leq \frac{\mathbb{E}_0[L_d]}{\exp\left(C\left(\lambda s s'\right)^d\right)} \xrightarrow{d \to \infty} 0.$$

To show significant power, we apply Fatou's lemma to obtain

$$\begin{aligned}
\liminf_{d \to \infty} \mathbb{P}_1(\mathcal{T}_d = 1) &= \liminf_{d \to \infty} \mathbb{E}_1(\mathbb{1}_{\mathcal{T}_d = 1}) \geq \mathbb{E}_1(\liminf_{d \to \infty} \mathbb{1}_{\mathcal{T}_d = 1}) \\
&= \mathbb{P}_1\left(\liminf_{d \to \infty} \{\mathcal{T}_d = 1\}\right) \\
&= \mathbb{P}_1\left(\exists D \, \forall d \geq D : L_d \geq \exp\left(C\left(\lambda s s'\right)^d\right)\right) \\
&= \mathbb{P}_1\left(\liminf_{d \to \infty} \mu^{-d} \log(L_d) \geq C\right) > 0.
\end{aligned}$$

Thus, $(\mathcal{T}_d)_d$ is asymptotically one-sided. □

The next step requires the technical properties of the likelihood ratio:

Proof of $(i) \implies (iii)$. Assuming (i) to be true, we refer to Step 5 in the proof of Theorem 1 in [6] for a demonstration of $\mathrm{KL}_\infty = \infty$ which is independent of s and s'.

In order to show $\lambda s s' > 1$, we assume the contrary, i.e. $\lambda s s' \leq 1$. A standard result from Galton-Watson tree theory (c.f. [10]) states that in this case, 1 is the only fixed point in $[0, 1]$ of the probability generating funciton of $\mathrm{Poi}(\lambda s s')$. Since $\mathbb{E}_0[L_\infty]$ has the same fixed point property, we obtain $\mathbb{E}_0[L_\infty] = 1$. This gives us trivial convergence of the means $\mathbb{E}_0[L_d] = 1 \to 1 = \mathbb{E}_0[L_\infty]$, which we can combine with the almost sure convergence $L_d \to L_\infty$ and Scheffé's Lemma to obtain that L_d converges to L_∞ in L^1. This in turn yields that $(L_d)_d$ is a flat martingale and is therefore uniformly integrable. However, one-sided testability in the form of (ii) contradicts the definition of uniform integrability, since for all but finitely many d,

$$\mathbb{E}_0\left[|L_d| \mathbb{1}_{L_d > a_d}\right] = \mathbb{P}_1(L_d > a_d) \geq \varepsilon > 0$$

but we would expect the left term to go to 0. This contradiction lets us conclude $\lambda s s' > 1$. □

The final step $(iii) \implies (iv)$ is more involved and requires the following result on Galton-Watson trees from [10]:

Lemma 2. *Let $\tau \sim \mathrm{GW}_\mu$ and denote $w_d := |\mathcal{L}_d(\tau)|/\mu^d$. Then, $(w_d)_d$ is a positive martingale whose almost sure limit w satisfies*

$$\mathbb{P}\left(w > 0 \mid \tau \text{ survives }\right) = 1.$$

Proof of $(iii) \implies (iv)$. Let $(t, t') \sim \mathbb{P}_1$ and set $\tau^* = t \cap t'$ to be their true intersection tree. Call σ^* and σ'^* the injections of τ^* into t and t' respectively. Marginally, the intersection tree is distributed as $\tau^* \sim \mathrm{GW}_{\lambda ss'}$. Set $\mu := \lambda ss'$ and write $W_d := |\mathcal{L}_d(\tau^*)|$. By Lemma 2, the martingale $\mu^{-d} W_d$ converges almost surely towards a random variable w.

From now on, condition on the event $\{\tau^* \text{ survives }\}$ which is possible because $\mu > 1$. Lemma 2 consequently lets us assume $w > 0$. Using the result from section B of the Appendix, we have that for all $d, k \in \mathbb{N}$,

$$\begin{aligned} L_{d+k}(t, t') \geq & \prod_{i \in \mathcal{V}_{d-1}(\tau^*)} \psi\left(c_{\tau^*}(i),\, c_t(\sigma^*(i)),\, c_{t'}(\sigma'^*(i))\right) \\ & \times \prod_{j \in \mathcal{L}_d(\tau^*)} L_k(t_{\sigma^*(j)}, t'_{\sigma'^*(j)}) \\ =: & \; A_d \times B_{d,k} \end{aligned}$$

We examine A_d and $B_{d,k}$ individually starting with the latter. Taking the logarithm, we rewrite this product to a sum:

$$\frac{\log\left(B_{d,k}\right)}{|\mathcal{L}_d(\tau^*)|} = \underbrace{\frac{1}{|\mathcal{L}_d(\tau^*)|} \sum_{i \in \mathcal{L}_d(\tau^*)} \log\left(L_k(t_{\sigma^*(i)}, t'_{\sigma'^*(i)})\right)}_{=:a_d}.$$

For $i \neq j$, the random pairs $(t_{\sigma^*(i)}, t'_{\sigma'^*(i)})$ and $(t_{\sigma^*(j)}, t'_{\sigma'^*(j)})$ are independent, they follow the law $\mathbb{P}_1$ and the logarithms of their likelihoods are in $L^1(\mathbb{P}_1)$. The law of large numbers is therefore applicable, yielding $a_d \xrightarrow{d\to\infty} \mathbb{E}_1[\log L_k]$ almost surely. Consequently, on an event of probability 1, there exists a sequence $0 < \varepsilon_d \to 0$ such that for all d,

$$a_d \geq \mathbb{E}_1[\log(L_k)]| - \varepsilon_d \tag{1}$$

Similarly, since $|\mathcal{L}_d(\tau^*)|\mu^{-d} \to w$ almost surely, there also exists a sequence $0 < \delta_d \to 0$ such that

$$|\mathcal{L}_d(\tau^*)| \geq w\mu^d - \delta_d \mu^d \quad \text{for all } d. \tag{2}$$

Realizing $\mathbb{E}_1[\log(L_k)] = \mathrm{KL}_k$ and combining (1) and (2) gives

$$\begin{aligned} B_{d,k} &= \exp\left\{|\mathcal{L}_d(\tau^*)|a_d\right\} \geq \exp\left\{|\mathcal{L}_d(\tau^*)|\left(\mathrm{KL}_k - \varepsilon_d\right)\right\} \\ &\geq \exp\left\{\left(w\,\mu^d - \delta_d\mu^d\right)\left(\mathrm{KL}_k - \varepsilon_d\right)\right\} \\ &\geq \exp\left\{\mathrm{KL}_k\, w\,\mu^d - \mu^d\left(\delta_d \mathrm{KL}_k + \varepsilon_d w\right)\right\} \end{aligned}$$

Let k be large enough so that $\mathrm{KL}_k > 1$ and fix any $\omega \in \Omega$ for which all sequences converge. Choosing d sufficiently large, one has $\delta_d(\omega) \leq w(\omega)/4$ and $\varepsilon_d(\omega) \leq \mathrm{KL}_k/4$, yielding

$$\mu^d \delta_d(\omega)\mathrm{KL}_k + \mu^d \varepsilon_d(\omega) w(\omega) \leq ½\ \mathrm{KL}_k w(\omega)\mu^d.$$

This implies that on an event $\mathcal{B}$ of probability 1 and for d large enough,

$$B_{d,k} \geq \exp\left\{½\ \mathrm{KL}_k w \mu^d\right\} \tag{3}$$

Turning to A_d now, we start by introducing the notation

$$F_i := \log\left(\psi\left(c_{\tau^*}(i),\, c_t(\sigma^*(i)),\, c_{t'}(\sigma'^*(i))\right)\right)$$

As before, we take the logarithm to rewrite

$$\log(A_d) = \sum_{i \in \mathcal{V}_{d-1}(\tau^*)} F_i = \sum_{g=0}^{d-1} \sum_{i \in \mathcal{L}_g(\tau^*)} F_i.$$

Through basic computations, one can show that all F_i have the same distribution, finite variance and nonpositive mean. Letting $F \overset{(d)}{=} F_i$ and writing $\mathcal{L}_g$ as a shorthand for $\mathcal{L}_g(\tau^*)$, we set

$$\hat{S}_d := \sum_{g=0}^{d-1} \sum_{i \in \mathcal{L}_g} F_i - \mathbb{E}[F_i] = \log(A_d) - \mathbb{E}[F] \sum_{g=0}^{d-1} |\mathcal{L}_g|$$

Define sigma-fields $\mathcal{F}_d := \sigma\left(F_i : i \in \mathcal{L}_g, g \leq d-1\right)$ to get

$$\begin{aligned} \mathbb{E}\left[\hat{S}_d^2\right] =& \mathbb{E}\left[\mathbb{E}\left[\hat{S}_d^2 \mid \mathcal{F}_{d-1}\right]\right] \\ =& \mathbb{E}\left[\hat{S}_{d-1}^2\right] + \mathbb{E}\left[\sum_{i \in \mathcal{L}_{d-1}} \mathbb{E}\left[\left(F_i - \mathbb{E}[F_i]\right)^2\right]\right] \\ & + 2\mathbb{E}\left[\hat{S}_{d-1} \sum_{i \in \mathcal{L}_{d-1}} \mathbb{E}[F_i - \mathbb{E}[F_i]]\right] \\ =& \mathbb{E}\left[\hat{S}_{d-1}^2\right] + \mathrm{Var}(F)\, \underbrace{\mathbb{E}[|\mathcal{L}_{d-1}|]}_{=\mu^{d-1}}. \end{aligned}$$

Using this result recursively, we obtain

$$\mathbb{E}\left[\hat{S}_d^2\right] = \sum_{g=1}^{d} \mathrm{Var}(F)\, \mu^{g-1} = \mathrm{Var}(F)\, \frac{\mu^d - 1}{\mu - 1}. \tag{4}$$

This allows us to establish

$$\mathbb{P}\left(|\hat{S}_d| \geq d\mu^{\frac{d}{2}}\right) \leq \frac{\mathbb{E}\left[\hat{S}_d^2\right]}{d^2\mu^d} \leq \frac{1}{d^2}\, \mathrm{Var}(F)\, \frac{1 - \mu^{-d}}{\mu - 1} \leq \frac{C}{d^2}$$

where we have used Markov's inequality, (4), $\mathrm{Var}(F) < \infty$, and $1 - \mu^{-d} \leq 1$. The constant C depends on μ and $\mathrm{Var}(F)$ but not on d. Consequently, $\sum_{d=1}^{\infty} \mathbb{P}(|\hat{S}_d| \geq d\mu^{\frac{d}{2}}) < \infty$ and by Borel-Cantelli there almost surely exists $M \in \mathbb{N}$ such that

$$\left|\ \log(A_d) - \mathbb{E}[F] \sum_{g=0}^{d-1} |\mathcal{L}_g|\ \right| < d\mu^{\frac{d}{2}} \quad \text{for all } d \geq M.$$

Combining this with $|\mathcal{L}_q| \geq w\mu^q - \delta_q\mu^q$ from above, we get

$$\begin{aligned} \log(A_d) &\geq \mathbb{E}[F]\, w \sum_{g=0}^{d-1} \mu^q \;-\; \mathbb{E}[F] \sum_{g=0}^{d-1} \delta_q \mu^q - d\mu^{\frac{d}{2}} \\ &\geq \mathbb{E}[F]\, w \frac{\mu^d}{\mu - 1} - d\mu^{\frac{d}{2}} \quad \text{for } d \geq M. \end{aligned}$$

In the last step, we have used $\mathbb{E}[F] \leq 0$ to omit the second term. Summarizing our arguments about A_d, there exists an event $\mathcal{A}$ of probability 1 such that for any $\omega \in \mathcal{A}$ we find a constant $M'(\omega)$ such that for all $d \geq M'(\omega)$,

(i) $d\mu^{\frac{d}{2}} \leq w(\omega)\mu^d$, and
(ii) $\log(A_d)(\omega) \geq \left(\mathbb{E}[F](\mu-1)^{-1}-1\right) w(\omega)\mu^d$.

To conclude the proof, let $C' := \mathbb{E}[F](\mu-1)^{-1}-1$ and choose k large enough so that $\frac{1}{2}\,\mathrm{KL}_k > |C'|$. Combining (ii) with (3) implies that on the event $\mathcal{A}\cap\mathcal{B}$, for large enough d,

$$\log\left(L_{d+k}(t,t')\right) \geq \log(A_d) + \log(B_{d,k}) \geq w\mu^d(C' + \frac{1}{2}\mathrm{KL}_k)$$

Since this holds true on $\{\tau^* \text{ survives}\}\cap\mathcal{A}\cap\mathcal{B}$, we conclude

$$\mathbb{P}_1\left(\liminf_{d\to\infty} \mu^{-(d+k)}\log(L_{d+k}) \geq \mu^{-k}(C' + \tfrac{1}{2}\,\mathrm{KL}_k)\right) \geq \mathbb{P}\left(\{\tau^* \text{ survives}\}\cap\mathcal{A}\cap\mathcal{B}\right).$$

Setting $C := \mu^{-k}(C' + \frac{1}{2}\,\mathrm{KL}_k) > 0$ and noting

$$\mathbb{P}\left(\mathrm{Ext}^c(\tau^*)\cap\mathcal{A}\cap\mathcal{B}\right) = 1 - \mathbb{P}(\mathrm{Ext}(\mathrm{GW}_{\lambda ss'}))$$

this yields the desired result. Finally note that C can be computed from μ, $\mathbb{E}[F]$, KL_k, and n, which all solely depend on λ, s and s'. □

IV. Graph alignment algorithm

As a final step, we present the algorithm MPAlign from [6] which is the bridge between tree correlation testing and sparse graph alignment. Thanks to the recursive likelihood ratio formula from Proposition 1, we can calculate L_d in polynomial time as part of the algorithm.

Contrary to the original description of the algorithm, we refrain from introducing a specific notation for particular tree subgraphs: For two nodes $i, i' \in V(G)$, we talk about the *tree rooted at i and pointing away from i'* to describe the subgraph $\mathcal{N}_{G\setminus\{i,i'\}}(i,d)$. In other words, this is the subtree of G after removing the edge $\{i,i'\}$, which is rooted at i and restricted to depth d .

Algorithm 1 MPAlign (message-passing for graph alignment)

Input: Two graphs G, G', a depth d and a threshold β.
Output: A set of pairs $\mathcal{M}$ representing the matched nodes.

- $\mathcal{M} \leftarrow \emptyset$;
- For all pairs of nodes $i \in V(G)$ and $j \in V(G')$, use the recursive likelihood ratio formula to compute L_d for all pairs of trees rooted at i and j that are pointing away from their neighbors.

for $(i,j) \in V(G)\times V(G')$ **do**
  if neither $\mathcal{N}_G(i,d)$ nor $\mathcal{N}_{G'}(j,d)$ contain cycles and there exist triples of neighbors $i_1,i_2,i_3 \in \mathcal{N}_G(i,1)$ as well as $j_1,j_2,j_3 \in \mathcal{N}_{G'}(j,1)$ such that $L_{d-1} > \beta$ for all pairs of trees rooted at i_t, j_t, $t \in \{1,2,3\}$ and pointing away from i, j **then**
  - $\mathcal{M} \leftarrow \mathcal{M} \cup \{i,j\}$

  end if
end for
return $\mathcal{M}$

The reason why triples of neighbors are considered is the so-called *dangling tree trick*. It assures that in the case where i and j should not be matched, there exists at least one pair (i_t, j_t) whose corresponding trees are disjoint in the sense of the correct matching. This pair should therefore easily be identified as a sample from $\mathbb{P}_0$.

Remark 1. As detailed in [6], this algorithm's runtime is polynomial in n. This is why we talk about an *efficient* algorithm despite the fact that the polynomial's degree is rather high, which restricts the applicability to larger graphs.

We will now give a theoretical guarantee for Algorithm 1 to achieve one-sided partial alignment. The proof the following theorem can be adapted one-to-one from Section 6 in [6], by noting that $G\cup G' \sim \mathrm{ER}(n, \lambda(s+s'-ss')/n)$ and $G\cap G' \sim \mathrm{ER}(n, \lambda ss'/n)$. We emphasize that this result heavily relies on Lemma 1 from Section II.

Theorem 2. *Let (G,G') be correlated Erdős–Rényi graphs with either varying edge densities or differing node numbers. Further assume that the parameters λ, s and s' fulfill one of the conditions in Theorem 1. Set $\hat{n} := \min\{|V(G)|, |V(G')|\}$ and define $d = \lfloor c\log(\hat{n})\rfloor$ as well as $\beta = \exp(\hat{n}^\gamma)$ for some c, γ with $c\log(\lambda(s+s'-ss')) < 1/4$ and $\gamma \in (0, c\log(\lambda ss'))$. Then there exists $\varepsilon > 0$ such that with high probability, the output $\mathcal{M}$ from Algorithm 1 fulfills*

$$\mathrm{ov}_n(\mathcal{M}) = \frac{1}{|V^*|}\sum_{i\in V^*} \mathbb{1}_{(i,\sigma^*(i))\in\mathcal{M}} \geq \varepsilon \quad \textit{for large } n,$$

and

$$\mathrm{err}_n(\mathcal{M}) := \frac{1}{|V(G)|}\sum_{i\in V(G)} \mathbb{1}_{\{\exists j\neq\sigma^*(i)\,:\,(i,j)\in\mathcal{M}\}} \xrightarrow{n\to\infty} 0.$$

In other words, MPAlign achieves one-sided partial alignment.

Remark 2. This theorem is formulated to cover differing node numbers alongside varying edge densities. In the latter case, notations simplify since $n = \hat{n}$ and $V^* = [n]$.

Remark 3 (Subgraph isomorphism problem). Choosing $s = 1$ and $s' < 1$ in the differing node numbers setting, the second graph G' is a strict subgraph of G. Assuming the conditions in Theorem 2 to hold, one can partially recover the unknown injection $V(G_2) \hookrightarrow V(G_1)$ using Algorithm 1. This automatically yields a partial solution to the subgraph isomorphism problem on sparse Erdős–Rényi graphs!

V. Conclusion and outlook

This document provides insights into asymmetric tree correlation testing and has proposed a first alignment algorithm of sparse graphs with differing node numbers. The addition of asymmetry generalizes the results in [6] and makes them more applicable. Future work aims at finding more precise conditions on λ, s and s' for feasibility of tree correlation testing, following the line of work in [7]. Other open questions revolve around accelerating MPAlign for better scalability and understanding whether sparse graph alignment can be possible when tree correlation testing is not.

Acknowledgment

The authors thank Luca Ganassali for helpful discussions and the two reviewers for their valuable feedback.

REFERENCES

[1] Narayanan, Arvind, and Shmatikov, Vitaly, "Robust de-anonymization of large sparse datasets," 2008 IEEE Symposium on Security and Privacy (2008), pp. 111–125.

[2] Aladağ, Ahmet E, and Erten, Cesim, "SPINAL: scalable protein interaction network alignment," Bioinformatics, vol. 29, no. 7 (2013), pp. 917–924.

[3] Pedarsani, Pedram, and Grossglauser, Matthias, " On the privacy of anonymized networks," Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (2011), pp. 1235–1243.

[4] Ding, Jian, and Hang, Du, "Matching recovery threshold for correlated random graphs," arXiv preprint, https://arxiv.org/abs/2205.14650 (2022).

[5] Mao, Cheng, Yihong Wu, Jiaming Xu, and Sophie H. Yu., "Random graph matching at Otter's threshold via counting chandeliers", arXiv preprint, https://arxiv.org/abs/2209.12313 (2022).

[6] Ganassali, Luca, Massoulié, Laurent, and Lelarge, Marc, "Correlation detection in trees for planted graph alignment," arXiv preprint, https://arxiv.org/abs/2107.07623 (2022).

[7] Ganassali, Luca, Massoulié, Laurent, and Semerjian, Guilhem, "Statistical limits of correlation detection in trees," arXiv preprint, https://arxiv.org/abs/209.13723 (2022).

[8] Ganassali, Luca, and Massoulié, Laurent, "From tree matching to sparse graph alignment," Conference on Learning Theory (2020), pp. 1633–1665.

[9] Ganassali, Luca, Massoulié, Laurent, and Lelarge, Marc, "Impossibility of partial recovery in the graph alignment problem," Conference on Learning Theory, PMLR (2021), pp. 2080–2102.

[10] Abraham, Romain, and Delmas, Jean-François, "An introduction to Galton-Watson trees and their local limits," arXiv preprint, https://arxiv.org/abs/1506.05571 (2015).

APPENDIX

A. The recursive likelihood ratio formula

We present a proof of the essential part of Proposition 1:

Let $(t,t') \in \mathcal{X}_d^2$ and denote their root nodes as $\rho(t))$ and $\rho(t')$. Define $c := \deg(\rho(t))$ and $c' := \deg(\rho(t'))$ as well as the event $D := \{\deg(\rho(\tau)) = c, \deg(\rho(\tau')) = c'\}$. Compute

$$\begin{aligned}
\mathbb{P}_{1,d}(t,t') &= \mathbb{P}_{1,d}\left((\tau,\tau') = (t,t'), D\right) \\
&\overset{(a)}{=\!=} \sum_{k=0}^{c\wedge c'} \mathbb{P}_{1,d}\left((\tau,\tau') = (t,t'),\, D,\, \deg(\rho(\tau\cap\tau')) = k\right) \\
&= \sum_{k=0}^{c\wedge c'} \mathbb{P}_{1,d}\left(D,\, \deg(\rho(\tau\cap\tau')) = k\right)\, \boldsymbol{\xi} \\
&\overset{(b)}{=\!=} \sum_{k=0}^{c\wedge c'} \pi_{\lambda s s'}(k)\pi_{\lambda s(1-s')}(c-k)\pi_{\lambda s'(1-s)}(c'-k)\, \boldsymbol{\xi}
\end{aligned}$$

where

$$\boldsymbol{\xi} = \mathbb{P}_{1,d}\left((\tau,\tau') = (t,t') \mid D,\, \deg(\rho(\tau\cap\tau')) = k\right)$$

Note that in (a) we have used that $\deg(\rho(\tau\cap\tau')) \le \deg(\rho(\tau)) \wedge \deg(\rho(\tau'))$, while (b) relies on the fact that τ and τ' are (λ, s, s') and (λ, s', s)-augmentations respectively. Denoting by $t_1,\dots,t_c$ the child trees of $\rho(t)$, we have

$$\begin{aligned}
\boldsymbol{\xi} = \sum_{\sigma,\sigma'} \frac{1}{c!\,c'!} &\left(\prod_{i=1}^{k} \mathbb{P}_{1,d-1}(t'_{\sigma(i)}, t'_{\sigma'(i)})\right) \\
&\times \left(\prod_{i=k+1}^{c} \mathrm{GW}_{\lambda s,d-1}(t_{\sigma(i)})\right)\left(\prod_{i=k+1}^{c'} \mathrm{GW}_{\lambda s',d-1}(t'_{\sigma'(i)})\right).
\end{aligned}$$

Under the null-hypothesis, the likelihood is computed much more easily:

$$\begin{aligned}
\mathbb{P}_{0,d}(t,t') &= \mathrm{GW}_{\lambda s,d}(t)\,\mathrm{GW}_{\lambda s',d}(t') \\
&= \pi_{\lambda s}(c)\left(\prod_{i=1}^{c}\mathrm{GW}_{\lambda s,d}(t_i)\right)\,\pi_{\lambda s'}(c')\left(\prod_{i=1}^{c'}\mathrm{GW}_{\lambda s',d}(t'_i)\right).
\end{aligned}$$

Putting things together, we obtain

$$\begin{aligned}
L_d(t,t') = &\sum_{k=0}^{c\wedge c'} \frac{\pi_{\lambda s s'}(k)\pi_{\lambda s(1-s')}(c-k)\pi_{\lambda s'(1-s)}(c'-k)}{\pi_{\lambda s}(c)\,\pi_{\lambda s'}(c')\,c!\,c'!} \\
&\times \sum_{\substack{\sigma\in\mathfrak{S}_c,\\ \sigma'\in\mathfrak{S}_{c'}}} \left(\prod_{i=1}^{k}\mathbb{P}_{1,d-1}(t'_{\sigma(i)}, t'_{\sigma'(i)})\right) \\
&\times \left(\prod_{i=1}^{k}\mathrm{GW}_{\lambda s}(t_{\sigma(i)})\right)^{-1}\left(\prod_{i=1}^{k}\mathrm{GW}_{\lambda s'}(t'_{\sigma'(i)})\right)^{-1} \\
= &\sum_{k=0}^{c\wedge c'} \psi(k,c,c') \sum_{\substack{\sigma\in\mathfrak{S}_c,\\ \sigma'\in\mathfrak{S}_{c'}}} \prod_{i=1}^{k} L_{d-1}(t_{\sigma(i)}, t'_{\sigma'(i)}).
\end{aligned}$$

where the term ψ implicitly defined in the last equation can be simplified to

$$\psi(k,c,c') = \exp(\lambda s s')\frac{(1-s')^{c-k}(1-s)^{c'-k}}{\lambda^k\, k!(c-k)!(c'-k)!}.$$

B. Lower bound on L_{d+k}

Applied iteratively, the recursive likelihood ratio formula yields the explicit formula

$$L_d(t,t') = \sum_{\tau\in\mathcal{X}_d} \sum_{\substack{\sigma\in S(\tau,t),\\ \sigma'\in S(\tau,t')}} \prod_{i\in\mathcal{V}_{d-1}(\tau)} \psi\left(c_\tau(i), c_t(\sigma(i)), c_{t'}(\sigma'(i))\right)$$

where $S(\tau,t)$ denotes the set of all injective mappings from τ to t. Applying this formula to $L_{d+k}(t,t')$ only keeping the summands such that $\tau = \tau^*$ up to depth d, we get the lower bound

$$\begin{aligned}
&L_{d+k}(t,t') \\
&\ge \sum_{\substack{\tau\in\mathcal{X}_{d+k}\\ \tau_d=\tau_d^*}} \sum_{\substack{\sigma\in S(\tau,t),\,\sigma=\sigma^* \text{ on } \mathcal{V}_d(\tau)\\ \sigma'\in S(\tau,t'),\,\sigma'=\sigma'^* \text{ on } \mathcal{V}_d(\tau')}} \\
&\quad\times \prod_{i\in\mathcal{V}_{d-1}(\tau^*)} \psi\left(c_{\tau^*}(i),\, c_t(\sigma^*(i)),\, c_{t'}(\sigma'^*(i))\right) \\
&\quad\times \prod_{j\in\mathcal{L}_d(\tau^*)}\ \prod_{m\in\mathcal{V}_{k-1}(\tau_j)} \\
&\qquad \psi\left(c_{\tau_j}(m),\, c_{t_{\sigma^*(j)}}(\sigma(m)),\, c_{t'_{\sigma'^*(j)}}(\sigma'(m))\right) \\
&= \left(\prod_{i\in\mathcal{V}_{d-1}(\tau^*)} \psi\left(c_{\tau^*}(i),\, c_t(\sigma^*(i)),\, c_{t'}(\sigma'^*(i))\right)\right) \\
&\quad\times \left(\prod_{j\in\mathcal{L}_d(\tau^*)} L_k(t_{\sigma^*(j)}, t'_{\sigma'^*(j)})\right).
\end{aligned}$$