Proceedings of Machine Learning Research vol 99:1–30, 2019 32nd Annual Conference on Learning Theory

Planting trees in graphs, and finding them back

Laurent Massoulié LAURENT.MASSOULIE@INRIA.FR
Inria, École Normale Supérieure, PSL Research University, Microsoft Research-Inria Joint Centre

Ludovic Stephan LUDOVIC.STEPHAN@INRIA.FR
Inria, École Normale Supérieure, PSL Research University

Don Towsley TOWSLEY@CS.UMASS.EDU
University of Massachusetts Amherst

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

In this paper we study the two inference problems of detection and reconstruction in the context of planted structures in sparse Erdős-Rényi random graphs $\mathcal{G}(n, \lambda/n)$ with fixed average degree $\lambda > 0$. Motivated by a problem of communication security, we focus on the case where the planted structure consists in the addition of a tree graph.

In the case of planted line graphs, we establish the following phase diagram for detection and reconstruction. In a low density region where the average degree λ of the original graph is below some critical value $\lambda_c = 1$, both detection and reconstruction go from impossible to easy as the line length K crosses some critical value $K^* = \ln(n)/\ln(1/\lambda)$, where n is the number of nodes in the graph. In a high density region where $\lambda > \lambda_c$, detection goes from impossible to easy as K goes from $o(\sqrt{n})$ to $\omega(\sqrt{n})$. In contrast, reconstruction remains impossible so long as $K = o(n)$.

We then consider planted D-ary trees of varying depth h and $2 \leq D \leq O(1)$. For these we identify a low-density region $\lambda < \lambda_D$, where λ_D is the threshold for emergence of the D-core in Erdős-Rényi random graphs $\mathcal{G}(n, \lambda/n)$ for which the following holds. There is a threshold $h* = g(D)\ln(\ln(n))$ with the following properties. Detection goes from impossible to feasible as h crosses $h*$. Interestingly, we show that only partial reconstruction is feasible at best for $h \geq h*$. We conjecture a similar picture to hold for D-ary trees as for lines in the high-density region $\lambda > \lambda_D$, but confirm only the following part of this picture: Detection is easy for D-ary trees of size $\omega(\sqrt{n})$, while at best only partial reconstruction is feasible for D-ary trees of any size $o(n)$.

These results provide a clear contrast with the corresponding picture for detection and reconstruction of *low rank* planted structures, such as dense subgraphs and block communities. In the examples we study, there is i) an absence of hard phases for both detection and reconstruction, and ii) a discrepancy between detection and reconstruction, the latter being impossible for a wide range of parameters where detection is easy. The latter property does not hold for previously studied low rank planted structures.

1. Introduction

This paper is concerned with the detection of additional structures planted in a graph initially without structure (such as an Erdős-Rényi graph) and, in case such a structure is detected, with the reconstruction of the corresponding structure. We focus on planted structures that consist in a superimposed graph, and more specifically on superimposed trees.

A first motivation for this focus stems from the following application scenario. Assume that the original graph without planted structure represents normal communications among agents, while the superimposed graph represents communications among a subset of *attackers* who, when active, connect directly among themselves to coordinate their activity. Detection then amounts to estimating whether an attack occurs, while reconstruction amounts to identifying the attackers in case of an attack.

A second motivation is theoretical: previous work reviewed in Section 2 has shown that detection and reconstruction of planted structures in graphs displays rich and intriguing behaviour, with phases where the task is either impossible, computationally hard, or easy. It is important to understand what causes such phases, and whether phases for detection always coincide with the corresponding phase for reconstruction. Our present study sheds light on these questions, by showing that in the cases of planted tree structures we consider, no hard phase occurs, while feasibility phases of detection and reconstruction differ widely. In contrast, the latter property does not hold for previously studied low rank planted structures.

More specifically, our contributions are as follows. In the particular case of planted line graphs, we determine the complete phase diagram for detection and reconstruction: In a low density region where the average degree λ of the original graph is below some critical value λ_c, both detection and reconstruction go from impossible to easy as the line length K crosses some critical value $K^* = f(\lambda)\ln(n)$, where n is the number of nodes in the graph. In a high density region where $\lambda > \lambda_c$, detection goes from impossible to easy as K goes from $o(\sqrt{n})$ to $\omega(\sqrt{n})$. In contrast, reconstruction remains impossible so long as $K = o(n)$.

We then consider the case of D-ary trees for fixed $D > 1$, of height h. For these our results provide a similar picture with significant differences. Specifically, there exists a limit height $h_* = \ln\ln(D) + O(1)$ such that detection is impossible if $h < h_* - \ln(h_*)$, and easy for $h > h_* + \Omega(1)$. In that latter case, non-trivial reconstruction is feasible, but it must fail on a non-vanishing fraction of the K attack nodes. In a high-density region $\lambda > \lambda_D$, we have again that detection is easy for $K = \omega(\sqrt{n})$, and that reconstruction must fail at least on a fraction of nodes.

The paper is organized as follows. We review related work in Section 2. We describe our model and main results in Section 3. The proofs for planted lines and planted D-ary trees are in Sections 5 and 6 respectively, with detailed proofs of auxiliary results in the Appendix.

2. Related work

Planted clique detection and reconstruction has been the object of many works, see e.g. Dekel et al. (2014), Deshpande and Montanari (2015), Barak et al. (2016) for recent results and surveys. A central result in that context is that detection appears hard (i.e. no algorithm is known to succeed at detection in polynomial time) for cliques of size $o(\sqrt{n})$ planted in $G(n, 1/2)$. IT thresholds for planted dense subgraph detection are developed in Verzelen and Arias-Castro (2015).

Computational hardness of planted clique is used in reduction arguments to show that other planted structure detection problems are hard, eg sparse PCA Berthet and Rigollet (2013), and dense

subgraph detection Hajek et al. (2015). The latter also displays IT-impossible phases, hard phases and easy phases. A systematic development of such reductions between problems with planted structure is initiated in Brennan et al. (2018).

Community detection and reconstruction has also been thoroughly studied, the seminal article Decelle et al. (2011) introducing several conjectures on feasibility of detection and reconstruction for the stochastic block model. Almost all conjectures in Decelle et al. (2011) have been verified in subsequent works, in particular Mossel et al. (2015), Massoulié (2013), Mossel et al. (2013), Abbe and Sandon (2016).

Presence of specific subgraphs in random graphs has been thoroughly studied, see e.g. Janson et al. (2011). We leverage the corresponding techniques in our study of low density regions, for which detection feasibility corresponds to absence of copies of the planted graph structure in the original random graph.

Most planted structures considered so far were typically of "low rank" (e.g. planted dense graph's expected adjacency matrix is, up to diagonal terms, a rank one perturbation); in contrast, adjacency matrices of trees and lines are not close to a low rank matrix. One notable exception is the planted Hamiltonian cycle reconstruction addressed in Bagaria et al. (2018).

3. Model and main results

A total population of n agents interconnects according to one of the following two modalities. Under the null hypothesis H_0 the interconnection does not display any specific structure. We assume that the corresponding graph G is an Erdős-Rényi $\mathcal{G}(n,p)$ graph, with edge probability $p \in [0,1]$ taken equal to λ/n for some fixed $\lambda > 0$. We thus focus on sparse random graphs with average degree $O(1)$. Under the alternative hypothesis H_1, the graph G is the union of a base graph G_0 distributed according to $\mathcal{G}(n,p)$, with another graph G' connecting a distinguished subset $\mathcal{K}$ of nodes. Specifically, for a fixed graph Γ on node set $[K]$ with edge set $\mathcal{E}$, and an injective map $\sigma : [K] \to [n]$ chosen uniformly at random and independently of G_0, G' consists of the nodes $\mathcal{K} = \{\sigma(i), i \in [K]\}$ and edges $\{(\sigma(i),\sigma(j)), (i,j) \in \mathcal{E}\}$.

We shall mostly focus on tree graphs Γ, and more specifically on D-ary trees, i.e. trees with a distinguished root, or depth-0 node, and for each $\ell \in [h-1]$, D^ℓ depth-ℓ nodes being connected to one parent at depth $\ell - 1$ and D children at depth $\ell + 1$. The two exteme cases are a line graph for $D = 1$ and a star for $D = K - 1$.

We are interested in answering, on the basis of an observed graph G, the following questions:

Q1 (Detection): For a given planted graph shape Γ (e.g. line, star, D-ary tree,...), under what parameter regimes specified by λ and K is there a test that distinguishes H_0 from H_1 with error probabilities of both kinds going to zero as $n \to \infty$? This is an information-theoretic property characterized by the likelihood ratio $\frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)}$, where $\mathbb{P}_i$ denotes the distribution of G uner H_i, $i = \{0,1\}$. Indeed by the Neyman-Pearson lemma, among tests with given probability of correctly deciding H_1, there is one which minimizes probability of erroneously rejecting H_0 which decides H_1 if and only if the likelihood ratio $L(G) := \frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)}$ is larger than some threshold τ. We can ask the same question as Q1 when we restrict ourselves to tests that can be implemented in polynomial time. This then corresponds to a computational property.

Q2 (Reconstruction): Can one reconstruct the planted structure G', or at least a subset of its constituent nodes? Several metrics of reconstruction accuracy are possible. We shall focus on the

following **overlap** metric, which we now define for estimation procedures that produce a set $\hat{\mathcal{K}}$ of K nodes in $[n]$, aimed to estimate at best the actual set $\mathcal{K}$ of K nodes involved in the attack.

Definition 1 *The* **overlap** *of a set $\hat{\mathcal{K}}$ estimating the actual ground truth $\mathcal{K}$ is by definition the expected size of their intersection, i.e.*

$$\mathrm{ov}(\hat{\mathcal{K}}) := \sum_{i\in[n]} \mathbb{P}(i \in \hat{\mathcal{K}} \cap \mathcal{K}).$$

We say that a particular reconstruction $\hat{\mathcal{K}}$ of size K **fails** *if $\mathrm{ov}(\hat{\mathcal{K}}) = o(K)$,* **succeeds** *if $\mathrm{ov}(\hat{\mathcal{K}}) = K(1-o(1))$, and* **partially succeeds** *if $\mathrm{ov}(\hat{\mathcal{K}}) = cK(1-o(1))$ for some $c \in (0,1)$.*

Reconstruction (respectively, partial reconstruction) is then deemed **feasible** if there exists an estimator $\hat{\mathcal{K}}$ that is successful (respectively, partially successful). These properties are of an information-theoretic nature. Indeed the best possible overlap is achieved by the so-called Maximum a Posteriori (MAP) estimation procedure, and these properties are therefore determined by the overlap of the MAP estimator. One can, as for detection, consider a computational version of reconstruction: reconstruction (respectively, partial reconstruction) is **easy** when it can be achieved by an estimator $\hat{\mathcal{K}}$ that is efficiently computable.

Before stating our results for planted lines and D-ary trees, we first consider planted star graphs, for which a simpler picture holds:

Theorem 2 *For any fixed $\lambda > 0$, a planted star of size $K = \ln(n)/\ln(\ln(n))[1-\omega(1/\ln(\ln(n)))]$ is not detectable, while both detection and reconstruction of a planted star of size $K = \ln(n)/\ln(\ln(n))[1+\omega(1/\ln(\ln(n)))]$ are easy.*

$K < \ln(n)/\ln(1/\lambda)$	$\ln(n)/\ln(1/\lambda) < K \ll n/\ln(n)$
Detection & reconstruction IT impossible	Detection & reconstruction easy

(a) Subcritical regime : $\lambda < 1$

$K \ll \sqrt{n}$	$\sqrt{n} \ll K \ll n$
Detection & reconstruction IT impossible	Detection easy reconstruction IT impossible

(b) Supercritical regime : $\lambda > 1$

Table 1: Summary of results for planted line graph

The result for line graphs, summarized in Table 1, is

Theorem 3 (Line graphs) *In the low-density region $\lambda < \lambda_c = 1$, detection and reconstruction are impossible if $K = \ln(n)/\ln(1/\lambda) - \omega(\ln(\ln(n)))$, while both detection and reconstruction are easy if $K = \ln(n)/\ln(1/\lambda) + \omega(1)$ and $K = o(n/\ln(n))$.*

In the high-density region $\lambda > \lambda_c = 1$, detection and reconstruction are impossible if $K = o(\sqrt{n})$, detection is easy if $K = \omega(\sqrt{n})$, while reconstruction is impossible for $K = o(n)$.

$h < \ln\ln(n)/\ln(D)$	$\ln\ln(n)/\ln(D) < h < \ln(n)$
Detection & reconstruction IT impossible	Detection easy complete reconstruction impossible

(c) Subcritical regime : $\lambda < \lambda_D$

$h < \ln(n)/2$	$\ln(n)/2 < h < \ln(n)$
Detection unknown complete reconstruction impossible	Detection easy complete reconstruction impossible

(d) Supercritical regime : $\lambda > \lambda_D$

Table 2: Summary of results for planted D-ary tree

For D-ary trees, the results are similar. However the critical parameter λ_D defined in (11) is the threshold for emergence of the D-core (see Moore and Mertens (2011)), and only partial reconstruction is possible in the subcritical regime $\lambda < \lambda_D$. We consider D-ary trees Γ of depth h with corresponding size $K = \frac{D^{h+1}-1}{D-1}$; the main results (in terms of h) are summarized in Table 2.

Theorem 4 (D-ary trees) *In the low-density region $\lambda < \lambda_D$, there exist two parameters $\underline{h}$ and $\bar{h}$ such that the following holds.*

$\bar{h} = \ln\ln(n)/\ln(D) + \Theta(1)$, and $\underline{h} = \bar{h} - 1$ for almost all λ.

When $h \leq \underline{h} - O(\ln(\underline{h}))$, both detection and reconstruction are impossible with high probability.

Detection is easy whenever $h \geq \bar{h} + O(1)$.

For any $\lambda > 0$, hence in both low-density and high-density regions, detection is easy whenever $K = \omega(\sqrt{n})$ while complete reconstruction is impossible for $K = o(n)$.

4. Preliminary results

We now state three results that hold for arbitrary planted structures, and that will be used extensively. The first is a characterization of the likelihood ratio $\frac{\mathbb{P}_1}{\mathbb{P}_0}$:

Lemma 5 *The likelihood ratio $L(G) = \frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)}$ is given by $L(G) = \frac{X_\Gamma}{\mathbb{E}_0(X_\Gamma)}$, where X_Γ denotes the number of copies of Γ in G.*

The second gives a generic detection process that succeeds for K large enough, and all planted graph structures Γ that are connected.

Theorem 6 *Assume that $\lambda > 0$, $K = \omega(\sqrt{n})$, and the hidden graph is any connected subgraph on K nodes, not necessarily a line. Then the total variation distance $|\mathbb{P}_1 - \mathbb{P}_0|_{var}$ between $\mathbb{P}_0$ and $\mathbb{P}_1$ goes to 1 as $n \to \infty$.*

Let A_i, $i \in \{1, 2, 3\}$ denote the number of size i-connected components in G, $\hat{\lambda} = (nA_3)/(A_1A_2)$, and $\hat{k} = n - e^{\hat{\lambda}}A_1$. The test that decides H_1 if $\hat{k} \geq t_n := \sqrt{K\sqrt{n}}$, and H_0 otherwise is polynomial-time computable and distinguishes with high probability graphs sampled from $\mathbb{P}_1$ or $\mathbb{P}_0$.

Remark 7 *When λ is known, a simpler test based on the number of edges in the graph also succeeds. The test in Theorem 6 still applies even when λ is unknown. The proof further implies that under $\mathbb{P}_1$, G can be distinguished from $\mathcal{G}(n, \lambda'/n)$ for any λ' not necessarily equal to λ.*

Finally, it is important to note that, as evidenced in Banks et al. (2018), impossibility of detection does not imply immediately that of reconstruction. Fortunately, in our setting, the following result will imply the latter as soon as the former is proved :

Theorem 8 *Assume that* $K = o(\sqrt{n})$ *that* $\mathbb{E}_0(X_\Gamma) = \omega(1)$ *and that* $\mathbb{E}_0\left(L^2\right) = 1 + o(1)$. *Then, for every estimator* $\hat{\mathcal{K}}$ *of the planted set* $\mathcal{K}$, *we have*

$$\mathrm{ov}(\hat{\mathcal{K}}) = o(K),$$

that is, reconstruction fails as well.

5. Proof strategy for planted paths

We say that the ordered set $\{i_1, \dots, i_K\}$ of K distinct nodes in $[n]$ is a K-path in G if and only if the edges $(i_\ell, i_{\ell+1})$ are present in G for all $\ell = 1, \dots, K-1$. The previous Lemma 5 yields, in the case where Γ is the line graph, the following result, whose proof is in the appendix:

Lemma 9 *For planted* K*-path, the likelihood ratio reads*

$$L(G) := \frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)} = \frac{1}{n(n-1)\cdots(n-K+1)} |\{K\text{-paths in } G\}| \left(\frac{\lambda}{n}\right)^{-K+1}. \tag{1}$$

Moreover one has

$$\mathbb{E}_0(L^2) = \mathbb{E}_0(x^S), \tag{2}$$

where $x = n/\lambda$, *and* S *is a random variable counting the number of edges common to the* K*-path* $(1-2-\cdots-K)$ *and a random* K*-path* π *chosen uniformly at random among the* $n(n-1)\cdots(n-K+1)$ *possible ones on node set* $[n]$.

5.1. Impossibility of detection

We have the following

Theorem 10 *Assume that* $\lambda > 1$ *and* $K = o(\sqrt{n})$, *or alternatively that* $\lambda < 1$ *and* $K = \ln(n)/\ln(1/\lambda) - \omega(\ln(\ln(n)))$. *Then the total variation distance* $|\mathbb{P}_1 - \mathbb{P}_0|_{var}$ *between* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *goes to zero as* $n \to \infty$. *Thus for any arbitrary test* $T(G) \in \{0, 1\}$, $\mathbb{P}_1(T(G) = 1) - \mathbb{P}_0(T(G) = 1) \to 0$ *as* $n \to \infty$.

By a standard argument, the variation distance $|\mathbb{P}_1 - \mathbb{P}_0|_{var}$ is upper-bounded by $\sqrt{\mathbb{E}_0(L^2) - 1}$, and thus the Theorem is a direct consequence of the following

Lemma 11 *Assume that* $\lambda > 1$ *and* $K = o(\sqrt{n})$, *or alternatively that* $\lambda < 1$ *and* $K = \ln(n)/\ln(1/\lambda) - \omega(\ln(\ln(n)))$. *Then* $\lim_{n\to\infty} \mathbb{E}_0(L^2) = 1$.

The proof of Lemma 11 (details in the Appendix) is based on an analysis of expression (2). Set $Z_t = 1$ if edge (I_t, I_{t+1}) is part of path $(1 \cdots K)$, $Z_t = 0$ if it is not part of that path, but $I_{t+1} \in [K]$, and finally $Z_t = -1$ if $I_{t+1} \notin [K]$, so that

$$\mathbb{E}_0(L^2) = \mathbb{E}_0(x^{\sum_{t=1}^{K-1} Z_t^+}) \tag{3}$$

In order to upper-bound this expression, a key step is the following Lemma, which exhibits a tractable upper bound involving a Markov chain:

Lemma 12 *Let $n' := n-K$. The Markov chain $\{Z'_t\}_{t\geq 1}$ taking values in $\{-1,0,1\}$ with transition probability matrix*

$$P := \begin{pmatrix} 1-K/n' & K/n' & 0 \\ 1-K/n' & (K-2)/n' & 2/n' \\ 1-K/n' & (K-1)/n' & 1/n' \end{pmatrix} \tag{4}$$

can be constructed jointly with process $\{Z_t\}_{t\geq 1}$ so that, for all $m \geq 1$, one has

$$\mathbb{E}_0(x^{\sum_{t=1}^m Z_t^+}) \leq \mathbb{E}_0(x^{\sum_{t=1}^m Z_t'^+}). \tag{5}$$

Its proof is in the appendix, together with the analysis of the right-hand side of (5). The latter relies on spectral analysis of a matrix derived from P in (4), which leverages perturbation arguments as $K/n \to 0$. It concludes the proof of Lemma 11 by showing that $\mathbb{E}_0(L^2) = 1 + o(1)$ under the Lemma's assumptions.

5.2. Easiness of detection and reconstruction, sparse case

Assume $\lambda < 1$ and $K = \ln(n)/\ln(1/\lambda) + \omega(1)$. Detection is then easy: under $\mathbb{P}_0$, the expected number of K-paths in the graph is $o(1)$. A test which decides $\mathbb{P}_1$ if there is a K-path and $\mathbb{P}_0$ otherwise thus discriminates the two hypotheses with high probability. Presence of a K-path can moreover be determined in polynomial time by running depth-first searches from each node in G.

For reconstruction, we need the following

Lemma 13 *For $\lambda < 1$, $K = \ln(n)/\ln(1/\lambda) + \omega(1)$ and $K = o(n/\ln(n))$, let C be the connected component of the graph containing the longest path. Apply $\sqrt{K}$ times a* peeling operation *to C, which consists in removing all degree one nodes, to obtain set C'. Under $\mathbb{P}_1$, set C' and its intersection with the planted path both have with high probability size $K \pm o(K)$.*

The Lemma readily implies a polynomial-time algorithm for reconstruction that achieves overlap $K - o(K)$: set C' can be obtained in polynomial time. By adding / removing $o(K)$ nodes to it one obtains a set of size K with overlap $K - o(K)$.

5.3. Impossibility of reconstruction, dense case

We assume $\lambda > 1$ and $K = \omega(\sqrt{n})$. We have seen that with high probability, observation of G allows to determine whether or not an attack has taken place. We now assume that an attack has indeed happened. We have the following result, showing the impossibility of efficient planted structure reconstruction:

Theorem 14 *Given $\lambda > 1$, $K = \omega(\sqrt{n})$, $K = o(n)$, and a realization G of the graph under $\mathbb{P}_1$, any estimator $\hat{\mathcal{K}}$ of the ground truth achieves negligible overlap, i.e. $\mathrm{ov}(\hat{\mathcal{K}}) = o(K)$.*

Its proof structure is as follows. Fix an arbitrary integer $\tau \geq 1$. We shall establish that necessarily

$$\mathrm{ov}(\mathcal{K}) \leq K/(\tau+1) + o(K). \tag{6}$$

Fix

$$L = C\ln(n) \text{ for some suitable constant } C,\ D \gg L \text{ and } D^2 \ll \frac{n}{\ln(n)}. \tag{7}$$

Condition on the event the attack path is precisely $k_1, \dots, k_K =: k_1^K$. Chop the attack path into $K/(L+D)$ contiguous segments, each of length $M := L + D$.

Consider the ℓ-th segment $\{k_{(\ell-1)M+1}, \dots, k_{\ell M}\}$. We shall construct, for some $I(\ell) \in [(\ell-1)M+1, (\ell-1)M+L]$, τ random paths of edges in the graph G of the form $k_{I(\ell)}, I_2(t,\ell), I_3(t,\ell), \dots, I_D(t,\ell), k_{I(\ell)+D}$ for $t \in [\tau]$ such that the nodes $I_2(t,\ell), \dots, I_D(t,\ell)$ are all distinct, none of them belongs to the attack path, and such that the paths $(k_1, \dots, k_K) =: k_1^K$ and $k_1^{I(\ell)}, I_2^D(t,\ell), k_{I(\ell)+D}^K$ are statistically indistinguishable. More precisely, we have the following:

Lemma 15 *There is a construction, for any $\ell \in [K/M]$, of τ random paths*

$$k_{I(\ell)}, I_2(t,\ell), I_3(t,\ell), \dots, I_D(t,\ell), k_{I(\ell)+D},\ t \in [\tau],$$

such that for any $i \in [(\ell-1)M+1, (\ell-1)M+L]$, any τ disjoint ordered sets of $D-1$ distinct nodes $i_2^D(t)$, $t \in [\tau]$ in $[n] \setminus k_1^K$, we have

$$d_{var}(\mathbb{P}_1(G \in \cdot | \mathcal{K} = k_1^K, I(\ell) = i, (I_2^D(t,\ell))_t = (i_2^D(t))_t), \mathbb{P}_0(G \in \cdot | k_1^K \in G, (k_i, i_2^D(t), k_{i+D})_t \in G)) = \epsilon = o(1). \tag{8}$$

This construction moreover verifies the following property. There is an event $\mathcal{E}$ such that $\mathbb{P}_1(\mathcal{E}) = 1 - o(1)$, and such that, denoting $|(\cup_{t\in[\tau]} I_2^D(t,\ell)) \cap (\cup_{t\in[\tau]} I_2^D(t,\ell'))|$ the number of common points between the node sets $\cup_{t\in[\tau]} I_2^D(t,\ell)$ and $\cup_{t\in[\tau]} I_2^D(t,\ell')$, one has:

$$\forall \ell \neq \ell' \in [K/M], \quad \mathbb{E}_1\left(|(\cup_{t\in[\tau]} I_2^D(t,\ell)) \cap (\cup_{t\in[\tau]} I_2^D(t,\ell'))| \mathbf{1}_{\mathcal{E}}\right) = O\left(\frac{D^2}{n}\right). \tag{9}$$

The Lemma's proof idea is as follows. The τ non-overlapping alternative path segments, that we refer to as a τ-path, are obtained by selecting uniformly at random one such τ-path among all present in the graph. Then (8) is established by showing that the number of τ-paths concentrates. In turn, this concentration is established by bounding the variance of the number of τ-paths. This is done using the Markov chain bounding technique used in Lemma 12. The second part of the Lemma, (9), requires further concentration results on the numbers of τ-paths, that follow from applying Janson's inequality Boucheron et al. (2013), p. 205, Theorem 6.31.

The proof idea of Theorem 14 (detailed in the appendix) is then as follows. The τ-paths of Lemma 15 provide τ alternative K-paths to the actual planted path. These are "lures" for the optimal MAP reconstruction algorithm, that must return on average as many points of each of these lure paths as of the planted path. Since all these $\tau+1$ paths have intersection of negligible size, the overlap achieved by MAP must necessarily be at most $K/(\tau+1)$.

6. Proof strategy for planted D-ary trees

We assume here that Γ is a complete D-ary tree of size K and depth h, with $D > 1$ a fixed constant.

Under $\mathbb{P}_0$, the neighbourhood of a given vertex in G is close to a Galton-Watson process with offspring law $\mathrm{Poi}(\lambda)$. The probability of the existence of an infinite D-ary subtree in this process is the largest non-negative root $p_*(D,\lambda)$ of the equation

$$p = \psi_D(\lambda p), \tag{10}$$

where

$$\psi_D(\mu) := \mathbb{P}(\mathrm{Poi}(\mu) \geq D),\ \mu \geq 0.$$

The behavior of the random graph differs based on whether the above probability is zero or not. We define the *critical* threshold λ_D as

$$\lambda_D = \sup\left\{\lambda > 0 \;\middle|\; p_*(D,\lambda) = 0\right\} \tag{11}$$

In the following, we focus on *subcritical* λ, that is whenever $\lambda < \lambda_D$.

6.1. Study of the Galton-Watson process

Let (T, o) be a rooted Galton-Watson tree with offspring law $\mathrm{Poi}(\lambda)$, with $\lambda < \lambda_D$. The following Theorem characterizes the distribution of the maximum height of a D-ary tree rooted in o.

Theorem 16 *Let (T, o) be a Galton-Watson tree as above, and $n > 0$. Let p_h be the probability that a D-ary tree of height h rooted in o is contained in T. Then, for almost all λ, there exists h_* such that*

$$p_{h_*+1} = o\left(\frac{1}{n}\right) \tag{12}$$

$$p_{h_*} = \Omega(n^{-c}) \text{ for some } c < 1 \tag{13}$$

Moreover, as $n \to \infty$ one has $h_* = \frac{\ln\ln(n)}{\ln(D)} + O(1)$.

Thus h_* depends on λ only through terms of lower (constant) order. The Theorem's proof, detailed in the appendix, relies on the following

Lemma 17 *The sequence p_h satisfies the recurrence relation*

$$p_1 = 1$$
$$p_{h+1} = \psi_D(\lambda p_h) \text{ for all } h \geq 1.$$

Necessarily $0 \leq p_{h+1} \leq p_h$ for all h (since a tree of height $h+1$ contains a tree of height h), and therefore by continuity of ψ_D, p_h converges as $h \to \infty$ to the largest fixed point of (10). By definition of λ_D, the only solution of this equation is $p_\infty = 0$, and thus

$$\lim_{h\to\infty} p_h = 0 \tag{14}$$

Now, $\psi_D(x) \sim \frac{x^D}{D!}$ as $x \to 0$, which implies that for h large enough, $p_{h+1} \simeq C\, p_h^D$, and thus $p_h \simeq C\, \varepsilon^{D^h}$ for some small $\varepsilon > 0$. A more rigorous version of this argument, as well as its use in the proof of Theorem 16, is presented in the Appendix.

6.2. Coupling and application to planted trees

Following the insights from the previous section, we define the two thresholds $\overline{h}$ and $\underline{h}$ by :

$$\overline{h} = \inf\left\{h > 0 \;\middle|\; p_h < \frac{1}{n}\right\}, \qquad \underline{h} = \sup\left\{h > 0 \;\middle|\; p_h > \frac{\ln(n)}{n}\right\}.$$

Theorem 16 implies that $\overline{h} \sim \frac{\ln\ln(n)}{\ln(D)}$, and that for almost all λ, $\overline{h} = \underline{h} + 1$, and otherwise $\overline{h} = \underline{h} + 2$. Also, $p_{\overline{h}} = o(\frac{1}{n})$ and $p_{\underline{h}} = \Omega(n^{-c})$ for some $c > 1$. The following Theorem connects the study from section 6.1 to our planted tree problem:

Theorem 18 *Let G be a graph drawn according to $\mathbb{P}_0$, and $h > 0$. Then with high probability:*
1. For $h \leq \underline{h}$, there are $\omega(1)$ D-ary trees of height h in G.
2. For $h \geq \overline{h} + C$, where C is a large enough constant, there are no D-trees of height h in G.

The second part of this theorem yields an easy detection algorithm whenever $h \geq \overline{h} + \Omega(1)$.

Corollary 19 *Assume that Γ is a complete D-ary tree of height h, with $h \geq \overline{h} + \Omega(1)$. Then w.h.p under $\mathbb{P}_0$, $X_\Gamma = 0$, and therefore the test $T(G) = 1$ iff $X_\Gamma > 0$ discriminates between H_0 and H_1 correctly with high probability.*

The two statements of Theorem 18 are a consequence of the following coupling lemma, whose proof, as well as the full proof of the theorem, is deferred to the appendix :

Lemma 20 *For a graph G and a vertex v in G, denote by $(G, v)_\ell$ the ℓ-neighbourhood of v in G. Similarly, let $(T, o)_\ell$ be the ℓ-neighbourhood of o in the Galton-Watson process described above.*

Then, under $\mathbb{P}_0$, assuming that $\ell = o(\log(n))$, the total distance variation between the law of $(G, v)_\ell$ and that of $(T, o)_\ell$ goes to 0 as a negative power of n when $n \to \infty$.

Furthermore, for $\lambda' > \lambda$, and (T', o') a GW process with parameter λ', then, provided the ℓ-neighbourhood of v is cycle-free, there exists a coupling between $(G, v)_\ell$ and $(T', o')_\ell$ such that $(G, v)_\ell \subseteq (T', o')_\ell$ with probability 1.

There is therefore a sharp cutoff in the probability of presence of tree of height h in G, and we have already seen in Corollary 19 that it can be leveraged to obtain a detection algorithm when $h \leq \underline{h}$. It remains however to study two aspects of the problem: reconstruction for $h \geq \overline{h}$, as well as the possibility (or lack thereof) of detection when $h \leq \underline{h}$.

6.3. Likelihood ratio and detection for $h \leq \underline{h}$

We conjecture, as is the case when $D = 1$, that when $h = \underline{h} - \omega(1)$, then the total variation distance $|\mathbb{P}_1 - \mathbb{P}_0|_{\text{var}}$ goes to 0 when $n \to \infty$. However, the Markov chain bounds used for lines cannot be easily adapted to the current setting, and we only prove this result for $h \leq \underline{h} - \Omega(\ln \ln \ln(n))$:

Theorem 21 *Assume that Γ is a D-ary tree of height h, with $D > 1$ and*

$$h \leq \underline{h} - \frac{\ln(\underline{h})}{\ln(D)} + \frac{\ln\left(1 - \frac{1}{D}\right)}{\ln(D)}.$$

Then, the total variation distance $|\mathbb{P}_1 - \mathbb{P}_0|_{\text{var}}$ goes to zero as $n \to \infty$. Thus, for any test $T(G) \in \{0, 1\}$, $\mathbb{P}_1(T(G) = 1) - \mathbb{P}_0(T(G) = 1) \to 0$ as $n \to \infty$.

As before this is deduced from the following Lemma, shown in the Appendix:

Lemma 22 *Under the same assumptions as Theorem 21, $\mathbb{E}_0(L^2) \to 1$ as $n \to \infty$.*

We believe the following stronger version of the Theorem to hold:

Conjecture 23 *The result of Theorem 21 holds true for all $h \leq \underline{h}$.*

If true, this conjecture would complete the bottom left part of the phase diagram for D-ary tree, with a sharp threshold between undetectability and detection/reconstruction.

6.4. Reconstruction for large h

When $\lambda < \lambda_D$ and $h \geq \overline{h}$, we have shown that under $\mathbb{P}_0$ there is w.h.p no copy of Γ in G. One could therefore expect to be able to reconstruct Γ with overlap $1 - o(1)$; however, this is not the case :

Theorem 24 *Given $\lambda > 0$, $h \geq \overline{h}$ such that $K = o(n)$, and a realization G of the graph under $\mathbb{P}_1$, the overlap achieved by any estimator $\hat{\mathcal{K}}$ of the attack is bounded above, i.e $\mathrm{ov}(\hat{\mathcal{K}}) \leq (1-\delta)K$ for some $\delta > 0$.*

The proof is based on the fact that when $D > 1$, the leaves make up a positive proportion of Γ, and they are hard to reconstruct with high precision. On the other hand, since there is no copy of Γ in G w.h.p, one can still reasonably expect to achieve a partial reconstuction. This is the heuristic behind our second conjecture :

Conjecture 25 *For all $h \geq \overline{h}$, there exists a $\delta > 0$ and an estimator (possibly random) $\hat{\mathcal{K}}$ such that w.h.p $\mathrm{ov}(\hat{\mathcal{K}}) \geq \delta K$.*

References

E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. In *NIPS'16*, 2016.

Vivek Kumar Bagaria, Jian Ding, David Tse, Yihong Wu, and Jiaming Xu. Hidden hamiltonian cycle recovery via linear programming. *CoRR*, abs/1804.05436, 2018. URL http://arxiv.org/abs/1804.05436.

Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Trans. Information Theory*, 64(7):4872–4894, 2018. doi: 10.1109/TIT.2018.2810020. URL https://doi.org/10.1109/TIT.2018.2810020.

Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, 2016.

A. D. Barbour and Louis H. Y. Chen, editors. *An introduction to Stein's method*, volume 4 of *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*. Singapore University Press, Singapore, 2005.

Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *CoRR*, abs/1304.0828, 2013. URL http://arxiv.org/abs/1304.0828.

Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. ISBN 0-387-94846-5. doi: 10.1007/978-1-4612-0653-8. URL http://dx.doi.org/10.1007/978-1-4612-0653-8.

B. Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001. ISBN 9780521797221. URL https://books.google.fr/books?id=o9WecWgilzYC.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1347–1357, 2015. doi: 10.1109/FOCS.2015.86. URL https://doi.org/10.1109/FOCS.2015.86.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255. URL https://books.google.fr/books?id=koNqWRluhP0C.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 48–166. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/brennan18a.html.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106 (1–19), Dec 2011. doi: 10.1103/PhysRevE.84.066106.

Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *Combinatorics, Probability and Computing*, 23(1):29–49, 2014. doi: 10.1017/S096354831300045X.

Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, August 2015.

B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. In *Proceedings COLT 2015*, pages 899—928, June 2015.

S. Janson, T. Luczak, and A. Rucinski. *Random Graphs*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2011.

Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 1966.

L. Massoulié. Community detection thresholds and the weak ramanujan property. *arXiv:1109.3318. The conference version appeared in Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2013.

Cristopher Moore and Stephan Mertens. *The Nature of Computation*. Oxford University Press, Inc., New York, NY, USA, 2011. ISBN 0199233217, 9780199233212.

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arxiv:1311.4115*, 2013.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015. ISSN 0178-8051.

Nicolas Verzelen and Ery Arias-Castro. Community detection in sparse random networks. *Ann. Appl. Probab.*, 25(6):3465–3510, 12 2015. doi: 10.1214/14-AAP1080. URL https://doi.org/10.1214/14-AAP1080.

Appendix A. Proof of preliminary results

A.1. Proof of Lemma 5

Let $\Gamma_1, \dots, \Gamma_m$ be the copies of Γ in K_n the complete graph on $[n]$, where, denoting $\mathrm{Aut}(\Gamma)$ the automorphism group of Γ, $m = \binom{n}{K} \frac{K!}{|\mathrm{Aut}(\Gamma)|}$. Then, by Bayes' formula, letting $e(G)$ denote the number of edges in graph G, one has for any graph g:

$$\begin{aligned}\mathbb{P}_1(G=g) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_0(G=g \,|\, \Gamma_i \in G) &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\Gamma_i \in g} \left(\frac{\lambda}{n}\right)^{e(g)-e(\Gamma_i)} \left(1-\frac{\lambda}{n}\right)^{\binom{n}{2}-e(g)} \\ &= \frac{1}{m}\left(\frac{\lambda}{n}\right)^{-e(\Gamma)} \sum_{i=1}^{m} \mathbf{1}_{\Gamma_i \in g}\, \mathbb{P}_0(G=g) \\ &= \frac{X_\Gamma}{\mathbb{E}_0[X_\Gamma]} \mathbb{P}_0(G),\end{aligned}$$

which completes the proof of Lemma 5.

A.2. Proof of Theorem 2

We first prove that planted stars of size $K = \ln(n)/\ln(\ln(n))[1-\omega(1/\ln(\ln(n)))]$ are undetectable. The number X of K-stars verifies

$$\mathbb{E}_0(X) = n\binom{n-1}{K}\left(\frac{\lambda}{n}\right)^K \sim n\frac{\lambda^K}{K!}.$$

We will have undetectability if $\mathbb{E}_0(L^2) \sim 1$, or equivalently by symmetry arguments, if

$$\mathbb{E}_0(X|\Gamma_1 \in G) \sim \mathbb{E}_0(X),$$

where Γ_1 is an arbitrary K-star, e.g. that made of edges $(i, K+1)$, $i \in [K]$. We decompose $\mathbb{E}_0(X|\Gamma_1 \in G)$ into three terms M_1, M_2 and M_3, the expected numbers of K-stars centered respectively: at node $K+1$, at some node $i \in [K]$, and finally at some node $i \in [n] \setminus [K+1]$. Since M_3 is upper-bounded by $\mathbb{E}_0(X)$, it suffices to show that M_1 and M_2 are $o(\mathbb{E}_0(X))$. One has:

$$\begin{aligned} M_2 &= K\left(\tbinom{n-2}{K-1}\left(\tfrac{\lambda}{n}\right)^{K-1} + \tbinom{n-2}{K}\left(\tfrac{\lambda}{n}\right)^{K-1}\right) \\ &\le \tfrac{2K^2}{n}\mathbb{E}_0(X) \\ &\ll \mathbb{E}_0(X).\end{aligned}$$

Also,

$$\begin{aligned} M_1 &= \textstyle\sum_{\ell=0}^{K} \binom{K}{\ell}\binom{n-K-1}{K-\ell}\left(\frac{\lambda}{n}\right)^{K-\ell} \\ &\le \textstyle\sum_{\ell=0}^{K} \binom{K}{\ell}\frac{\lambda^\ell}{\ell!} \\ &\le (1+\lambda)^K.\end{aligned}$$

The desired result $M_1 \ll \mathbb{E}_0(X)$ will follow if

$$\ln(n) + K\ln(\lambda) - \ln(K!) - K\ln(1+\lambda) \to +\infty.$$

The terms in K are of order at most $\ln(n)/\ln(\ln(n))$. By Stirling's formula, this will therefore hold provided $\ln(n) - K\ln(K) = \omega(\ln(n)/\ln(\ln(n)))$. By assumption,

$$K\ln(K) \leq \frac{\ln(n)}{\ln(\ln(n))}(1-\omega(1/\ln(\ln(n))))\ln(\ln(n)) = \ln(n) - \omega(\ln(n)/\ln(\ln(n))),$$

hence the undetectability result.

Similarly for detectability, the assumption that $K = \ln(n)/\ln(\ln(n))[1+\omega(1/\ln(\ln(n)))]$ entails that

$$\ln(\mathbb{E}_0(X)) = K\ln(\lambda) + \ln(n) - \ln(K!) = -\omega(1).$$

Thus a test which decides H_1 if and only if there is a node in G with degree at least K succeeds with high probability. Moreover, with high probability, only the centre of the planted star has degree at least K. The reconstruction method which consists in choosing, besides the highest degree node, K of its neighbours chosen uniformly at random, achieves an overlap of $K-o(K)$: indeed, conditional on the planted star's centre having initially Y neighbors in the original graph, the expected number of nodes in the reconstructed set will be

$$1 + \frac{K^2}{Y+K} \geq 1 + K(1-Y/K) = 1 + K - Y.$$

Its expectation is lower-bounded by $K+1-\lambda$, and is thus $K-o(K)$.

A.3. Proof of Theorem 6

Let k denote the size of the hidden connected component, with $k=0$ under $\mathbb{P}_0$ and $k=K$ under $\mathbb{P}_1$. Let A_1 count the number of isolated nodes in G, A_2 the number of connected pairs (i,j) that form an isolated component, and A_3 the number of triplets (i,j,k) that form a connected component.

These quantities satisfy with high probability

$$A_1 = e^{-\lambda}(n-k)+O(\sqrt{n}),\ A_2 = \frac{(n-k)^2}{2}\frac{\lambda}{n}e^{-2\lambda}+O(\sqrt{n}),\ A_3 = \frac{(n-k)^3}{2}\frac{\lambda^2}{n^2}e^{-3\lambda}+O(\sqrt{n}). \tag{15}$$

Indeed, only the $n-k$ nodes that are not part of the hidden connected graph can contribute to counts of connected components of size 1, 2 or 3. (15) then follows from evaluation of the expectation and variance of these quantities.

Set $\hat{\lambda} = (nA_3)/(A_1A_2)$. By (15), $\hat{\lambda} = \lambda + O(n^{-1/2})$. Now form $\hat{k} = n - e^{\hat{\lambda}}A_1$. Again by (15), $\hat{k} = n-(1-O(n^{-1/2}))(n-k)+O(\sqrt{n}) = k+O(\sqrt{n})$. Our test then decides H_1 if $\hat{k} \geq t_n$ and H_0 otherwise where t_n is such that $\sqrt{n} \ll t_n \ll K$, which is indeed satisfied for $t_n = \sqrt{K\sqrt{n}}$. This ensures that the test discriminates correctly between the two hypotheses with high probability. Necessarily then, the variation distance $|\mathbb{P}_0 - \mathbb{P}_1|_{var}$ goes to 1 as $n \to \infty$.

A.4. Proof of Theorem 8

We first begin by a simple lemma, using the concentration of X_Γ :

Lemma 26 *Let $\mathcal{I}_\Gamma$ be the proportion of pairs copies of Γ in G whose intersection is nonempty :*

$$\mathcal{I}_\Gamma = \frac{1}{X_\Gamma^2}\sum_{\Gamma',\Gamma''\in G} \mathbf{1}_{\Gamma'\cap\Gamma''\neq\emptyset},$$

where Γ' and Γ'' range over all copies of Γ in G.

Then $\mathbb{E}_0(\mathcal{I}_\Gamma) = o(1)$.

Proof (of Lemma 26). As in the proof of Lemma 5, let $\Gamma_1, \dots, \Gamma_m$ be the copies of Γ in K_n, and let $X_i = \mathbf{1}_{\Gamma_i \in G}$. Write

$$\mathbb{E}_0\left(X_\Gamma^2\right) = \sum_{i,j} \mathbb{E}_0\left(X_i X_j\right) = \mathbb{E}' + \mathbb{E}'', \tag{16}$$

where $\mathbb{E}'$ is the sum over Γ_i, Γ_j having disjoint vertex sets.

We can easily compute $\mathbb{E}'$:

$$\mathbb{E}' = \binom{n}{K}\binom{n-K}{K}\left(\frac{K!}{|\mathrm{Aut}(\Gamma)|}\right)^2 p^{2K-2} \sim \frac{n^{2K}p^{2K-2}}{|\mathrm{Aut}(\Gamma)|^2} \sim \mathbb{E}_0\left(X_\Gamma\right)^2$$

Since $\mathbb{E}_0\left(L^2\right) = 1 + o(1)$, it follows that

$$\frac{\mathbb{E}''}{\mathbb{E}_0\left(X_\Gamma^2\right)} = o(1). \tag{17}$$

Now, it is straightforward to see that

$$\sum_{\Gamma', \Gamma'' \in G} \mathbf{1}_{\Gamma' \cap \Gamma'' \neq \emptyset} = \sum_{\Gamma_i \cap \Gamma_j \neq \emptyset} X_i X_j.$$

Recall that $L = X_\Gamma / \mathbb{E}_0(X_\Gamma)$; we can decompose $\mathcal{I}_\Gamma$ as follows :

$$\begin{aligned}\mathcal{I}_\Gamma &= \mathcal{I}_\Gamma \mathbf{1}_{L^2>1/2} + \mathcal{I}_\Gamma \mathbf{1}_{L^2<1/2} \\ &= \frac{\sum_{\Gamma_i \cap \Gamma_j \neq \emptyset} X_i X_j}{\mathbb{E}_0\left(X_\Gamma^2\right)} \cdot \frac{1}{L^2} \cdot \mathbf{1}_{L^2>1/2} + \mathcal{I}_\Gamma \mathbf{1}_{L<1/\sqrt{2}}\end{aligned}$$

We can now bound each term separately. The first one is straightforward since $1/L^2 < 2$ whenever the indicator variable is nonzero ; for the second one, notice that $I_\Gamma \leq 1$ and thus

$$\begin{aligned}\mathbb{E}_0(\mathcal{I}_\Gamma) &\leq \frac{\mathbb{E}''}{\mathbb{E}_0\left(X_\Gamma^2\right)} \cdot 2 + \mathbb{P}_0\left(L < \frac{1}{\sqrt{2}}\right) \\ &= \frac{\mathbb{E}''}{\mathbb{E}_0\left(X_\Gamma^2\right)} \cdot 2 + o(1),\end{aligned}$$

having used the Bienaymé-Chebychev inequality to bound the second term.

Using (17) then completes the proof. ∎

We can now move on to the proof of Theorem 8 ; we first transform the expression of $\mathrm{ov}(\hat{\mathcal{K}})$ to better suit our needs :

$$\begin{aligned}\mathrm{ov}(\hat{\mathcal{K}}) &= \sum_G \sum_{\mathcal{K}} \mathbb{P}_1(G, \mathcal{K}) \left|\hat{\mathcal{K}} \cap \mathcal{K}\right| \\ &= \sum_G \mathbb{P}_1(G) \sum_{\mathcal{K}} \mathbb{P}_1(\mathcal{K} \,|\, G) \left|\hat{\mathcal{K}} \cap \mathcal{K}\right|\end{aligned}$$

where $\mathcal{K}$ ranges over all K-subsets of $[n]$ and G over all graphs on n vertices.

The second sum can be transformed as in the proof of Lemma 5 into :

$$\begin{aligned}\mathrm{ov}(\hat{\mathcal{K}}) &= \sum_G \mathbb{P}_1(G) \sum_{\Gamma' \in G} \frac{|\hat{\mathcal{K}} \cap \Gamma'|}{X_\Gamma} \\ &= \sum_G \mathbb{P}_0(G) \sum_{\Gamma' \in G} \frac{|\hat{\mathcal{K}} \cap \Gamma'|}{X_\Gamma} + o(K),\end{aligned}$$

since the conditions in Theorem 8 imply that $|\mathbb{P}_1 - \mathbb{P}_0|_{\mathrm{var}} = o(1)$ (see the remark after Theorem 10). The sum now ranges over all copies of Γ in G.

This can now be expressed as an expectation :

$$\begin{aligned}\mathrm{ov}(\hat{\mathcal{K}}) &= \mathbb{E}_0 \left[\sum_{\Gamma' \in G} \frac{|\hat{\mathcal{K}} \cap \Gamma'|}{X_\Gamma} \right] + o(K) \\ &= \sum_{i \in [n]} \mathbb{E}_0 \left[\mathbf{1}_{i \in \hat{\mathcal{K}}} \sum_{\Gamma' \in G} \frac{\mathbf{1}_{i \in \Gamma'}}{X_\Gamma} \right] + o(K).\end{aligned}$$

We can now finally use Lemma 26 : indeed,

$$\begin{aligned}\left(\sum_{\Gamma' \in G} \frac{\mathbf{1}_{i \in \Gamma'}}{X_\Gamma} \right)^2 &= \frac{1}{X_\Gamma^2} \sum_{\Gamma', \Gamma'' \in G} \mathbf{1}_{i \in \Gamma'} \mathbf{1}_{i \in \Gamma''} \\ &\leq \frac{1}{X_\Gamma^2} \sum_{\Gamma', \Gamma'' \in G} \mathbf{1}_{\Gamma' \cap \Gamma'' \neq \emptyset} \\ &= \mathcal{I}_\Gamma.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{ov}(\hat{\mathcal{K}}) &\leq \sum_{i \in [n]} \mathbb{E}_0 \left[\mathbf{1}_{i \in \hat{\mathcal{K}}} \sqrt{\mathcal{I}_\Gamma} \right] + o(K) \\ &= K \mathbb{E}_0 \left[\sqrt{\mathcal{I}_\Gamma} \right] + o(K) \\ &= o(K),\end{aligned}$$

using Jensen's inequality as well as Lemma 26. This completes the proof of Theorem 17.

Appendix B. Detailed proofs for planted paths

B.1. Proof of Lemma 9

Expression (1) follows directly from Lemma 5. In the display below, by $\sum_{(i_1 \cdots i_K)}$ we mean summation over all the $n(n-1)\cdots(n-K+1)$ oriented paths $(i_1, \ldots, i_K)$ of length K over nodes in $[n]$. Write:

$$\begin{aligned}\mathbb{E}_0(L^2) &= \textstyle\sum_{(i_1 \cdots i_K)} \sum_{(j_1 \cdots j_K)} \left(\frac{(n/\lambda)^{K-1}}{n \cdots (n-K+1)} \right)^2 \mathbb{P}_0(\text{paths } (i_1 \cdots i_K) \text{ and } (j_1 \cdots j_K) \text{ present in } G) \\ &= \textstyle\sum_{(i_1 \cdots i_K)} \left(\frac{(n/\lambda)^{2(K-1)}}{n \cdots (n-K+1)} \right) \mathbb{P}_0(\text{paths } (i_1 \cdots i_K) \text{ and } (1 \cdots K) \text{ present in } G) \\ &= \left(\tfrac{n}{\lambda} \right)^{K-1} \mathbb{P}_0(\text{ path } \pi = (I_1 \cdots I_K) \text{ present in } G | \text{ path } (1 \cdots K) \text{ present in } G),\end{aligned}$$

where $\pi = (I_1 \cdots I_K)$ is a candidate path chosen uniformly at random from the $n(n-1)\dots(n-K)$ possible length-K paths. In the above we used symmetry to consider a single path $(1 \cdots K)$ instead of all paths $(j_1 \cdots j_K)$.

Note that conditionally on the event that path $(1 \cdots K)$ be present in G and on the path π, the probability that path π is also present in G is given by $(\lambda/n)^{K-1-S}$, where S is the number of edges in common between the two paths π and $(1 \cdots K)$. This yields expression (2).

B.2. Proof of Lemma 12

Let $\mathcal{F}_t = \sigma(I_1, \dots I_t)$. Recall that $n' = n - K$. It is easily verified that we have the following inequalities for all $t = 2, \dots, K-1$:

$$\mathbb{P}(Z_t = 1 | \mathcal{F}_t) \leq \begin{cases} \frac{1}{n'} & \text{if } Z_{t-1} = 1, \\ \frac{2}{n'} & \text{if } Z_{t-1} = 0, \\ 0 & \text{if } Z_{t-1} = -1. \end{cases}$$

Similarly we have

$$\mathbb{P}(Z_t \geq 0 | \mathcal{F}_t) \leq \frac{K}{n'}.$$

Moreover it is easily seen that $\mathbb{P}(Z_1 = 1) \leq (K/n')(2/n')$, and $\mathbb{P}(Z_1 \geq 0) \leq K/n'$.

As in Lemma 12, we introduce the Markov chain $\{Z'_t\}_{t \geq 1}$ on state space $\{-1, 0, 1\}$ specified by the initial distribution $\mathbb{P}(Z'_1 = 1) = (K/n')(2/n')$, $\mathbb{P}(Z'_1 \geq 0) = K/n'$ and by the transition probability matrix P in (4), that we recall for convenience:

$$P = \begin{pmatrix} 1 - K/n' & K/n' & 0 \\ 1 - K/n' & (K-2)/n' & 2/n' \\ 1 - K/n' & (K-1)/n' & 1/n' \end{pmatrix}$$

The previous inequalities ensure that we can construct by induction over t a coupled version of the two processes $\{Z_t\}$ and $\{Z'_t\}$ such that $Z_1 \leq Z'_1$, and for $t \geq 1$, if $Z'_t = -1$ then $Z_t = -1$, and furthermore we have the following implications:

$$\begin{array}{ll} Z_t = -1 & \Rightarrow Z_{t+1} \leq Z'_{t+1}, \\ Z_t = Z'_t & \Rightarrow Z_{t+1} \leq Z'_{t+1}, \\ (Z_t, Z'_t) = (1, 0) & \Rightarrow Z_{t+1} \leq Z'_{t+1}. \end{array}$$

Thus the only situation when we can have $Z_{t+1} > Z'_{t+1}$ is when $(Z_t, Z'_t) = (0, 1)$. That is to say, for each time $t+1$ when process Z hits 1 while chain Z' does not, then at time t chain Z' hits 1 while process Z does not.

Because of this, the number of times t at which process Z hits 1 is upper-bounded by the number of times t at which chain Z' does. Thus (5) holds, concluding the proof of Lemma 12.

B.3. Proof of Lemma 11

By (5) and (3), $\mathbb{E}_0(L^2)$ is upper bounded by

$$\mathbb{E}_0(L^2) \leq \mathbb{E}_0 x^{\sum_{s=1}^{K-1} Z'^+_s}. \tag{18}$$

To evaluate this term, introduce the row vector $F(t) := \{f_z(t)\}_{z\in\{-1,0,1\}}$ where $f_z(t) := \mathbb{E}_0 x^{\sum_{s=1}^t Z'^+_s} \mathbf{1}_{Z'_t=z}$. We then have

$$F(1) = (\mathbb{P}(Z'_1 = -1), \mathbb{P}(Z'_1 = 0), x\mathbb{P}(Z'_1 = 1)) = (1 - K/n', K/n'(1 - 2/n'), x(K/n')(2/n')), \tag{19}$$

together with the recurrence relation

$$F(t + 1) = F(t)M, \tag{20}$$

where

$$M = \begin{pmatrix} 1 - K/n' & K/n' & 0 \\ 1 - K/n' & K/n' - 2/n' & x2/n' \\ 1 - K/n' & K/n' - 1/n' & x/n' \end{pmatrix}$$

Recall now that $x = n/\lambda$ and $n' = n - K$, so that x/n' is asymptotic to $1/\lambda$. Thus the above matrix M reads

$$M = M_0 + (K/n)M_1,$$

where

$$M_0 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 2/\lambda \\ 1 & 0 & 1/\lambda \end{pmatrix}, \tag{21}$$

and the entries of matrix M_1 are $O(1)$. Note that M_0 admits eigenvalues $0, 1/\lambda, 1$ with respective left eigenvectors

$$\begin{array}{ll} u_0 & := (1, 1, -2), \\ u_{1/\lambda} & := (-\lambda/(\lambda - 1), 0, 1), \\ u_1 & := (1, 0, 0). \end{array}$$

We shall denote (μ_r, v_r) the (eigenvalue,eigenvector) pair of M obtained by perturbation of the eigenpair (r, u_r) of M_0, with $r \in \{0, 1/\lambda, 1\}$. By the Bauer-Fike theorem (see Bhatia (1997), Theorem VI.25.1), $|\mu_r - r| = O(K/n)$ for all r.

Moreover Eq. (1.16), p. 67 in Kato (1966) implies that a normed left (resp., right) eigenvector of M associated to an eigenvalue μ_r of M differs in norm from a normed left (resp., right) eigenvector of M_0 associated to eigenvalue r by $O(K/n)$. We can thus chose $v_r = u_r + O(K/n)$.

Let the decomposition of vector $F(1)$ in the basis provided by the eigenvectors $\{v_r\}$ be given by:

$$F(1) = \sum_{r\in\{0,1/\lambda,1\}} \alpha_r v_r.$$

Denote by e the all-ones 3×1 column vector. The upper bound (18) on $\mathbb{E}_0(L^2)$ then gives

$$\begin{array}{rcl} \mathbb{E}_0(L^2) & \leq & F(K - 1)e \\ & = & F(1)M^{K-2}e \\ & = & \sum_{r\in\{0,1/\lambda,1\}} \alpha_r v_r \mu_r^{K-2} e. \end{array} \tag{22}$$

By our choice of eigenvectors v_r such that $|v_r - u_r| = O(K/n)$, and the fact that

$$F(1) = (1 + O(K/n))u_1 + O(K/n)u_{1/\lambda} + O(K/n)u_0,$$

corresponding weights α_r verify $\alpha_1 = 1 + O(K/n)$, $\alpha_{1/\lambda} = O(K/n)$, $\alpha_0 = O(K/n)$.

In the case where $\lambda > 1$ and $K = o(\sqrt{n})$, (22) yields

$$\mathbb{E}_0(L^2) \leq o(1) + (1+o(1))\mu_1^{K-2} = (1+o(1))(1+O(K/n))^{K-2} \leq (1+o(1))e^{O(K^2/n)}.$$

The assumption that $K = o(\sqrt{n})$ then allows to conclude.

For $\lambda < 1$ and $K = \ln(n)/\ln(1/\lambda) - \omega(\ln(\ln(n)))$, (22) yields

$$\mathbb{E}_0(L^2) \leq (1+o(1))\,(1+O(K/n))^{K-2} + O(K/n)\,(1/\lambda + O(K/n))^{K-2}\,.$$

The first term is $1 + o(1))$ since $K^2/n = o(1)$. The second term's logarithm is equivalent to

$$\ln(K) - \ln(n) + (K-2)\ln(1/\lambda) \leq \ln(\ln(n)) - \ln(\ln(1/\lambda)) - \omega(\ln(\ln(n))),$$

and goes to $-\infty$ by assumption.

B.4. Proof of Lemma 13

We place ourselves under $\mathbb{P}_1$ and condition on the fact that the K-path planted in the original Erdős-Rényi graph G_0 is k_1^K. Denote for each $i \in [K]$ by C_i the connected component of node k_i in G_0. Denote by $\mathcal{E}_i$ the event that $C_i \cap \{\cup_{j\neq i} C_j\} \neq \emptyset$ and by $\mathcal{E}_i'$ the event that C_i contains a cycle.

A standard construction of connected components based on a random walk exploration implies the existence of a constant $c > 0$ such that for all $\ell \geq 0$,

$$\begin{array}{ll} \mathbb{P}(\mathcal{E}_i', |C_i| = \ell) & \leq \frac{\lambda\ell^2}{n}\mathbb{P}(|C_i| = \ell) \leq \frac{\lambda\ell^2}{n}e^{-c\ell}, \\ \mathbb{P}(\mathcal{E}_i, |C_i| = \ell) & \leq \frac{\ell K}{n}e^{-c\ell}, \\ \mathbb{P}(|C_i| \geq \ell) & \leq e^{-c\ell}. \end{array} \tag{23}$$

The first evaluation implies that with high probability, no C_i contains a cycle (i.e. no $\mathcal{E}_i'$ occurs) when $K = o(n)$. The second evaluation implies that the expected number of $i \in [K]$ such that $\mathcal{E}_i$ occurs and $|C_i| \geq \ell$ is upper bounded, for some constant $c' > 0$, by

$$\sum_{i\in[K]} \mathbb{P}(\mathcal{E}_i, |C_i| \geq \ell) \leq \frac{K^2}{n}e^{-c'\ell}.$$

If $K^2 = o(n)$, then this implies that with high probability, no $\mathcal{E}_i$ occurs. Thus with high probability, there is no cycle in the connected component C. Moreover, the third evaluation in (23) ensures that

$$\sum_{i\in K} \mathbb{P}(C_i \geq \sqrt{K}) \leq Ke^{-c\sqrt{K}} = o(1).$$

Thus the peeling process applied $\sqrt{K}$ times to C returns exactly the planted K-path, except for $\sqrt{K}$ nodes at each of its ends.

If on the other hand, $K^2 > o(n)$, we choose $\ell* = \theta\ln(n)$ and deduce from (23) that with probability $1 - O(n^{-2})$, say, there is no $i \in [K]$ such that both $\mathcal{E}_i$ and $|C_i| \geq \theta\ln(n)$ hold. The peeling process applied $\sqrt{K}$ times to C then returns the planted path, shortened by no more than $\sqrt{K}$ nodes at each end, plus parts of the neighborhoods C_i for which $\mathcal{E}_i$ occurs. The expected number of nodes returned that do not belong to the planted path is therefore no more than

$$K\mathbb{P}(\mathcal{E}_i)\ell^* = O(\frac{K^2}{n})\theta\ln(n).$$

This is $o(K)$ under the assumption that $K = o(n/\ln(n))$. The conclusion of the Lemma follows.

B.5. Proof of Theorem 14

We show that Lemma 15 implies (6). First, the optimal overlap is achieved by the **Maximum A Posteriori** (MAP) inference procedure, i.e. by putting in $\hat{\mathcal{K}}$ the K nodes with the highest probability, conditional on the observed graph G, of being in $\mathcal{K}$. The probability that node j belongs to $\mathcal{K}$ conditional on G is proportional to the number of K-paths in G to which j belongs. We denote by $\mathcal{K}^*$ the corresponding set.

Second, when under the alternative distribution $\mathbb{P}_2 := \mathbb{P}_0(G \in \cdot | k_1^K \in G, (k_i, i_2^D, k_{i+D}) \in G))$ in (8), the joint distribution of the numbers of K-paths going through the nodes k_1^K or through the nodes in k_1^i, i_2^D, k_{i+D}^K are statistically indistinguishable. Thus, letting N_ℓ (respectively N'_ℓ) denote the number of points of $k_{(\ell-1)M+1}^{\ell M}$ (respectively, $k_{(\ell-1)M+1}^{i}, i_2^D, k_{i+D}^{\ell M}$) that the MAP estimate selects, one has:

$$\mathbb{E}_2(N_\ell) = \mathbb{E}_2(N'_\ell).$$

Let also $N'_{t,\ell}$ denote the number of points that the MAP estimate selects in $k_{(\ell-1)M+1}^{I(t,\ell)}, I_2^D(t,\ell), k_{I(t,\ell)+D}^{\ell M}$. Since each of these variables is bounded by $M = L + D$, the variation distance bound (8) implies

$$\mathbb{E}_1(N_\ell) \leq \mathbb{E}_1(N'_{t,\ell}) + \epsilon M.$$

Summing these inequalities over $\ell \in [K/M]$ and $t \in [\tau]$ yields

$$\tau \sum_{\ell=1}^{K/M} \mathbb{E}_1(N_\ell) = \tau \ \mathrm{ov}(\mathcal{K}^*) \leq \sum_{t=1}^{\tau} \sum_{\ell=1}^{K/M} \mathbb{E}_1(N'_{t,\ell}) + \epsilon \tau K. \tag{24}$$

However, it holds that:

$$\sum_{i=1}^{K} \mathbf{1}_{k_i \in \mathcal{K}^*} + \sum_{j \in \cup_{t,\ell} I_2^D(t,\ell)} \mathbf{1}_{j \in \mathcal{K}^*} \leq K.$$

This entails (using e.g. Bonferroni's inequality):

$$\sum_{i=1}^{K} \mathbf{1}_{k_i \in \mathcal{K}^*} + \sum_{\ell=1}^{K/M} \sum_{r=2}^{D} \sum_{t \in [\tau]} \mathbf{1}_{I_r(t,\ell) \in \mathcal{K}^*} - \sum_{\ell \neq \ell', \ell, \ell' \in [K/M]} |(\cup_{t \in [\tau]} I_2^D(\ell)) \cap (\cup_{t \in [\tau]} I_2^D(\ell'))| \leq K.$$

Taking expectations and using the last statement (9) of the Lemma yields, separating evaluations on event $\mathcal{E}$ and on its complementary set $\overline{\mathcal{E}}$:

$$\mathrm{ov}(\mathcal{K}^*) + \left\{ \sum_{t \in [\tau]} \sum_{\ell=1}^{K/M} \mathbf{E}_1(N'_{t,\ell}) \right\} - \tau L(K/M) - (K/M)^2 O(D^2/n) - \tau K \mathbb{P}_1(\overline{\mathcal{E}}) \leq K.$$

Summed with the previous equation (24), this gives:

$$(\tau + 1)\, \mathrm{ov}(\mathcal{K}^*) \leq K + K\tau \left(\epsilon + (L/M) + (K/n)(D/M)^2 + \mathbb{P}_1(\overline{\mathcal{E}}) \right).$$

The announced result follows from $\epsilon \ll 1$, $L \ll D$, $K = o(n)$ and $\mathbb{P}_1(\overline{\mathcal{E}}) = o(1)$.

B.6. Proof of Lemma 15, Equation (8)

We let π_i denote the set of τ candidate paths $(k_i, i_2^D(t,\ell), k_{i+D})_{t\in[\tau]}$ of the graph, where for fixed ℓ, the $\{i_2^D(t,\ell)\}_{t\in[\tau]}$ are distinct and in $[n]\setminus k_1^K$. For $i \in [(\ell-1)M+1, \ell-1)M+L]$ these can all be used to construct the set of τ alternative paths in the ℓ-th segment of k_1^K. We denote by

$$\pi(\ell) = \cup_{i\in[(\ell-1)M+1,\ell-1)M+L]}\pi_i$$

the corresponding collection. Our construction simply amounts to choosing a set of τ paths (that we shall call for short a τ-path) uniformly at random from $\pi(\ell)$ in order to construct the alternative τ-path for the ℓ-th segment, and this independently for each segment.

Denote $Z_i = |\pi_i|$. Then

$$\mathbb{E}_1(Z_i) = (n-K)(n-K-1)\cdots(n-K-\tau(D-1)+1)\left(\frac{\lambda}{n}\right)^{\tau D} \sim \frac{1}{n^\tau}\lambda^{\tau D},$$

since we assumed in (7) that $D \sim C\ln(n)$. Also, by symmetry,

$$\begin{aligned}\mathbb{E}_1 Z_i^2 &= \textstyle\sum_{i_2^D(t), j_2^D(t)} \mathbb{P}_1(\forall t\in[\tau], (k_i, i_2^D(t), k_i+D)\in G, (k_i, j_2^D(t), k_{i+D})\in G)\\ &= \mathbb{E}_1(Z_i)\textstyle\sum_{j_2^D(t)}\mathbb{P}_1(\forall t\in[\tau], (k_i, j_2^D(t), k_{i+D})\in G | \forall t\in[\tau], (k_i, i_2^D(t), k_{i+D})\in G),\end{aligned}$$

where in the last expression we fixed an arbitrary choice $(i_2^D(t))_{t\in[\tau]}$. It follows that:

$$\mathbb{E}_1 Z_i^2 = (\mathbb{E}_1(Z_i))^2\mathbb{E}_1((n/\lambda)^S),$$

where S is the number of common edges between the fixed τ-path $(k_i, i_2^D(t), k_{i+D})_{t\in[\tau]}$ and the τ-path $(k_i, J_2^D(t), k_{i+D})_{t\in[\tau]}$ where $(J_2^D(t))_{t\in[\tau]}$ is chosen uniformly at random among $(\tau(D-1))$ sequences in $[n]\setminus k_1^K$.

To control this second moment, we will condition on the number of common edges between each path $J_2^D(t)$ in the randomly selected τ-path at its beginning and end with the beginning and end of some of the fixed paths $i_2^D(t')$, that we shall denote by X_t and Y_t. These satisfy the constraints $X_t, Y_t \geq 0$, $X_t + Y_t \leq D$. For $X_t + Y_t < D$, this forces the choice of $X_t + Y_t$ nodes among the $D-1$ to be chosen for path $J_2^D(t)$; for $X_t + Y_t = D$, this forces all the $D-1$ choices. Moreover, conditionally on $(X_t, Y_t)_{t\in[\tau]}$, the expectation of the variable $(n/\lambda)^S$ verifies

$$\mathbb{E}_1((n/\lambda)^S | (X_t, Y_t)_{t\in[\tau]}) \leq (n/\lambda)^{\sum_{t\in[\tau]} X_t+Y_t}(1+O(D/n))^{\tau D},$$

by the Markov chain bounds in Lemma 12. By assumption, $D \ll \sqrt{n}$ so that $(1+O(D/n))^D = 1+o(1)$. Thus, accounits for the $\tau!)^2$ choices of path correspondences between the beginnings and ends of the planted and random paths:

$$\begin{aligned}\mathbb{E}_1 Z_i^2 &\leq (\mathbb{E}_1(Z_i))^2(\tau!)^2\left[(n/\lambda)^D n^{D-1} + \textstyle\sum_{x,y\geq 0, x+y<D}(n/\lambda)^{x+y}n^{-(x+y)}(1+o(1))\right]^\tau\\ &\leq (\mathbb{E}_1(Z_i))^2(1+o(1))(\tau!)^2[n\lambda^{-D} + (\textstyle\sum_{x\geq 0}\lambda^{-x})^2]^2\\ &\leq (\mathbb{E}_1(Z_i))^2(1+o(1))(\tau!)^2\left(\tfrac{\lambda}{\lambda-1}\right)^{2\tau},\end{aligned}$$

where we used that $n\lambda^{-D} = o(1)$.

We now evaluate $\mathbb{E}_1(Z_iZ_j)$ for $i \neq j$. The Markov chain bounding technique of Lemma 12 directly applies to give:

$$\mathbb{E}_1(Z_iZ_j) \leq (\mathbb{E}(Z_i))^2(1+o(1)).$$

Finally we obtain:

$$\begin{aligned}\mathrm{Var}(|\pi(\ell)|) &= L\mathrm{Var}(Z_i) + L(L-1)\mathrm{Cov}(Z_i,Z_j)\\ &\leq \mathbb{E}_1(Z_i)^2\left[L(1+o(1))(\tau!)^2\left(\tfrac{\lambda}{\lambda-1}\right)^{2\tau} + L^2 o(1)\right]\\ &\leq \mathbb{E}_1(|\pi(\ell)|)^2\left[\tfrac{O(1)}{L} + o(1)\right].\end{aligned}$$

Since by assumption $L \gg 1$, Tchebitchev's inequality implies that the random variable $|\pi(\ell)|$ concentrates: for some suitable $\epsilon = o(1)$, one has

$$\mathbb{P}_1\left(\left|\frac{|\pi(\ell)|}{\mathbb{E}_1|\pi(\ell)|} - 1\right| \geq \epsilon\right) \leq \epsilon.$$

Denote by $\mathcal{A}$ the event $\mathcal{A} := \{|\frac{|\pi(\ell)|}{\mathbb{E}_1|\pi(\ell)|} - 1| \leq \epsilon\}$. It thus has probability at least $1-\epsilon$. Consider a bounded function f of the graph G. This concentration result allows us to establish the variation distance bound (8) as follows. For some arbitrary candidate τ-path $(i, i_2^D(t))_{t\in[\tau]}$, omitting for brevity the argument t below, write:

$$\mathbb{E}_1(f(G)|\mathcal{A}, \mathcal{K} = k_1^K, I(\ell) = i, I_2^D(\ell) = i_2^D) = \frac{\mathbb{E}_1[f(G)\mathbf{1}_{\mathcal{A}}\mathbf{1}_{(k_i,i_2^D,k_{i+D})\in G}\frac{1}{|\pi(\ell)|}]}{\mathbb{E}_1(\mathbf{1}_{\mathcal{A}}\mathbf{1}_{(k_i,i_2^D,k_{i+D})\in G}\frac{1}{|\pi(\ell)|})}.$$

On $\mathcal{A}$ one has

$$\frac{1}{\mathbb{E}_1|\pi(\ell)|}\frac{1}{1+\epsilon} \leq \frac{1}{|\pi(\ell)|} \leq \frac{1}{\mathbb{E}_1|\pi(\ell)|}\frac{1}{1-\epsilon}.$$

This yields:

$$\frac{1-\epsilon}{1+\epsilon}\frac{\mathbb{E}_1[f(G)\mathbf{1}_{\mathcal{A}}\mathbf{1}_{(k_i,i_2^D,k_{i+D})\in G}]}{\mathbb{P}_1((k_i,i_2^D,k_{i+D})\in G)} \leq \mathbb{E}_1(f(G)|\mathcal{A},\mathcal{K}=k_1^K, I(\ell)=i, I_2^D(\ell)=i_2^D) \leq \frac{1+\epsilon}{1-\epsilon}\frac{\mathbb{E}_1[f(G)\mathbf{1}_{(k_i,i_2^D,k_{i+D})\in G}]}{\mathbb{P}_1(\mathcal{A}\cap(k_i,i_2^D,k_{i+D})\in G)}$$

By symmetry over all τ-paths in $\pi(\ell)$, denoting by Z the total number of possible such τ-paths in it ($Z \sim Ln^{\tau(D-1)}$), one has

$$\mathbb{P}_1(\mathcal{A}\cap(k_i,i_2^D,k_{i+D})\in G) = \frac{1}{Z}\mathbb{E}_1(|\pi(\ell)|\mathbf{1}_{\mathcal{A}}).$$

However by definition of $\mathcal{A}$ this is no smaller than

$$\frac{1}{Z}(1-\epsilon)\mathbb{E}_1|\pi(\ell)|\mathbb{P}_1(\mathcal{A}) \;\geq\; (1-\epsilon)^2\mathbb{P}_1((k_i,i_2^D,k_{i+D})\in G).$$

Finally we obtain:

$\frac{1-\epsilon}{1+\epsilon}\left[\mathbb{E}_1[f(G)|(k_i,i_2^D,k_{i+D})\in G] - ||f||_\infty\epsilon\right] \leq \mathbb{E}_1(f(G)|\mathcal{A},\mathcal{K}=k_1^K, I(\ell)=i, I_2^D(\ell)=i_2^D) \leq \cdots$
$\cdots \leq \frac{1+\epsilon}{(1-\epsilon)^3}\mathbb{E}_1[f(G)|(k_i,i_2^D,k_{i+D})\in G].$

The result of Equation (8) follows.

B.7. Proof of of Lemma 15, Equation (9)

We define the event $\mathcal{E}$ as, for some suitable constant $\alpha = \Omega(1)$:

$$\mathcal{E} := \cap_{\ell \in [K/M]} \mathcal{E}_\ell, \text{ where } \mathcal{E}_\ell := \{|\pi(\ell)| \geq \alpha \mathbf{E}_1 |\pi(\ell)|\}. \tag{25}$$

In the below display we let $I_2^D(\ell) = \cup_{t \in [\tau]} I_2^D(t, \ell)$, and $I_2^D(\ell) \cap I_2^D(\ell')$ the intersection of the two corresponding sets of nodes. We then have for arbitrary $\ell \neq \ell' \in [K/M]$:

$$\mathbb{E}_1(|I_2^D(\ell) \cap I_2^D(\ell')| \mathbf{1}_{\mathcal{E}}) = \sum_i \sum_j \mathbb{E}_1 \left(\frac{1}{|\pi(\ell)| \cdot |\pi(\ell')|} \sum_{i_2^D \in \pi_i} \sum_{j_2^D \in \pi_j} |i_2^D \cap j_2^D| \mathbf{1}_{\mathcal{E}} \right)$$

where the first summations are over $i \in [M(\ell-1)+1, M(\ell-1)+L]$ and $j \in [M(\ell'-1)+1, M(\ell'-1)+L]$. The expectation in the right-hand side does not depend on i and j, by symmetry. Moreover, on $\mathcal{E}$ we can upper bound the fraction in the expectation by $1/(\alpha \mathbb{E}_1 |\pi(\ell)|)^2$. Thus fixing some arbitrary $i \neq j$:

$$\begin{aligned} \mathbb{E}_1(|I_2^D(\ell) \cap I_2^D(\ell')| \mathbf{1}_{\mathcal{E}}) &\leq \tfrac{L^2}{(\alpha \mathbb{E}_1 |\pi(\ell)|)^2} \mathbb{E}_1 \left(\textstyle\sum_{i_2^D \in \pi_i} \sum_{j_2^D \in \pi_j} |i_2^D \cap j_2^D| \right) \\ &\leq \tfrac{L^2}{(\alpha \mathbb{E}_1 |\pi(\ell)|)^2} \textstyle\sum_{i_2^D, j_2^D} \mathbb{E}_1 \left(\mathbf{1}_{(k_i, i_2^D, k_{i+D}) \in G} \mathbf{1}_{(k_j, j_2^D, k_{j+D}) \in G} |i_2^D \cap j_2^D| \right), \end{aligned}$$

where summation is over all pairs of lists i_2^D and j_2^D of $\tau(D-2)$ distinct elements in $[n] \backslash k_1^K$. Denote by J_2^D one such list selected uniformly at random, and by i_2^D a fixed, arbitrary choice of one such list. One then has, recalling the expression of $\mathbb{E}_1 |\pi(\ell)| = L(\lambda/n)^{\tau D}(n-K) \cdots (n-K-\tau(D-1)+1)$:

$$\mathbb{E}_1(|I_2^D(\ell) \cap I_2^D(\ell')| \mathbf{1}_{\mathcal{E}}) \leq \frac{1}{\alpha^2} \mathbb{E}_1 \left(\left(\frac{n}{\lambda} \right)^S |i_2^D \cap J_2^D| \right), \tag{26}$$

where S denotes the number of edges in common between the two τ-paths i_2^D and J_2^D.

As in Lemma 12, we now define the Markov chain $\{Z'_t\}_{t \geq 0}$ on the three states $\{-1, 0, 1\}$, with transition probabilities given by the matrix

$$P := \begin{pmatrix} 1 - D/n' & D/n' & 0 \\ 1 - D/n' & (D-2)/n' & 2/n' \\ 1 - D/n' & (D-1)/n' & 1/n' \end{pmatrix},$$

where $n' = n - K - D$, and with initial condition $Z'_0 = -1$. These states are interpreted as follows: $Z'_t = -1$ if $J_{t+1} \notin i_2^D$, $Z'_t = 0$ if $J_t \notin i_2^D$ and $J_{t+1} \in i_2^D$, and $Z'_t = 1$ if $J_t, J_{t+1} \in i_2^D$. The same coupling argument as for Lemma 12 implies, letting $x = n/\lambda$, the following, where the subscript in the second expectation specifies the initial state of the Markov chain $\{Z'_t\}$:

$$\mathbb{E}_1 \left(\left(\frac{n}{\lambda} \right)^S |i_2^D \cap J_2^D| \right) \leq \mathbb{E}_{-1} \left(x^{\sum_{i=1}^{\tau(D-1)} Z_i'^+} \sum_{j=1}^{\tau(D-1)} \mathbf{1}_{Z'_j \geq 0} \right).$$

We introduce the notation $F_z(t) = (F_{z,-1}(t), F_{z,0}(t), F_{z,1}(t))$, where

$$F_{z,y}(t) := \mathbb{E}_z \left(x^{\sum_{s=1}^t Z_s'^+} \mathbf{1}_{Z'_t = y} \right).$$

It readily follows that

$$F_z(t) = (\mathbf{1}_{z=-1}, \mathbf{1}_{z=0}, \mathbf{1}_{z=1})M^t,$$

where

$$M := \begin{pmatrix} 1 - D/n' & D/n' & 0 \\ 1 - D/n' & (D-2)/n' & x * (2/n') \\ 1 - D/n' & (D-1)/n' & x/n' \end{pmatrix}.$$

This matrix M reads, as previously, $M_0 + O(D/n)$ where M_0 is given by (21).

Write then, using Markov's property:

$$\begin{aligned} \mathbb{E}_{-1}\left(x^{\sum_{i=1}^{\tau(D-1)} Z_i'^+} \sum_{j=1}^{\tau(D-1)} \mathbf{1}_{Z'_j \geq 0} \right) &= \sum_{j=1}^{\tau(D-1)} \sum_{z \in \{0,1\}} \mathbb{E}_{-1}\left(x^{\sum_{i=1}^{j} Z_i'^+} \mathbf{1}_{Z'_j = z} \right) \mathbb{E}_z\left(x^{\sum_{i=1}^{\tau(D-1)-j} Z_i'^+} \right) \\ &= \sum_{j=1}^{\tau(D-1)} \sum_{z \in \{0,1\}} F_{-1,z}(j) \sum_{y=-1,0,1} F_{z,y}(\tau(D-1) - j). \end{aligned}$$

Previously given perturbation results give the existence of coefficients $[\beta_{z,r}]_{z \in \{-1,0,1\}, r \in \{0,1/\lambda,1\}}$ all in $O(1)$ such that

$$F_z(0) = \sum_{r \in \{0,1/\lambda,1\}} \beta_{z,r} v_r.$$

It follows that

$$F_z(\tau(D-1) - j) = \sum_{r \in \{0,1/\lambda,1\}} \beta_{z,r} \mu_r^{\tau(D-1)-j} v_r = O(1),$$

since $|\mu_r| \leq 1 + O(D/n)$ and $D^2 \ll n$. It follows that

$$\begin{aligned} \mathbb{E}_{-1}\left(x^{\sum_{i=1}^{\tau(D-1)} Z_i'^+ \sum_{j=1}^{\tau(D-1)} \mathbf{1}_{Z'_j \geq 0}} \right) &= \sum_{j=1}^{\tau(D-1)} F_{-1}(j) \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \times O(1) \\ &= \sum_{j=1}^{\tau(D-1)} \sum_{r \in \{0,1/\lambda,1\}} \beta_{-1,r} \mu_r^j v_r \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \times O(1). \end{aligned}$$

Since $F_{-1}(0) = u_1$, it holds that $\beta_{-1,1} = 1 + O(D/n)$, and $\beta_{-1,r} = O(D/n)$ for $r = 0, 1/\lambda$. The terms with $r = 0, 1/\lambda$ in the previous expression thus contribute at most $O(D^2/n)$. The terms with $r = 1$ give

$$\sum_{j=1}^{\tau(D-1)} \beta_{-1,r} \mu_1^j v_1 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \times O(1) = O(D^2/n),$$

by using the fact that $v_1 = (1, 0, 0) + O(D/n)$.

It remains to prove that the event $\mathcal{E}$ defined in (25) is such that $\mathbb{P}_1(\mathcal{E}) = 1 - o(1)$. It will suffice to prove that for all $\ell \in [K/M]$, $\mathbb{P}_1(\mathcal{E}_\ell) \geq 1 - o(M/K)$. To show this we shall leverage Janson's inequality, as described in Boucheron et al. (2013), p.205, Theorem 6.31. Applied to the random variable $|\pi(\ell)|$, it guarantees that for all $0 \leq t \leq \mathbb{E}|\pi(\ell)|$ one has

$$\mathbb{P}_1(|\pi(\ell)| \leq \mathbb{E}|\pi(\ell)| - t) \leq e^{-t^2/(2\Delta)}, \tag{27}$$

where Δ is the expected number of ordered pairs of τ-paths (P,Q) in $\pi(\ell)$ that share at least an edge. Paralleling our previous bound on the variance of $|\pi(\ell)|$, we distinguish the pairs of τ-paths (P,Q) according to whether they share the same starting point $i \in [(\ell-1)M+1,(\ell-1)M+L]$ or not to write $\Delta = \Delta_1 + \Delta_2$, and obtain:

$$\Delta_1 \leq L\tfrac{\lambda^{2D\tau}}{n^{2\tau}}(1+o(1))(\tau!)^2\left(\tfrac{\lambda}{\lambda-1}\right)^{2\tau 2},$$
$$\Delta_2 \leq L^2\tfrac{\lambda^{2D\tau}}{n^{2\tau}}O\left(\tfrac{D^2}{n}\right).$$

We moreover have that $\mathbb{E}|\pi(\ell)| \sim L\frac{\lambda^{D\tau}}{n^\tau}$, so that

$$\frac{(\mathbb{E}|\pi(\ell)|)^2}{\Delta} \geq \frac{\Omega(1)}{\frac{1}{L}+\frac{D^2}{n}}.$$

By our choices (7) for L and D, this lower bound is also $\Omega(1)L = C\Omega(1)\ln(n)$. Taking $t = (1-\alpha)\mathbb{E}|\pi(\ell)|$ for some $\alpha \in (0,1)$ in (27), we obtain

$$\mathbb{P}_1\left(|\pi(\ell)| \leq \alpha\mathbb{E}|\pi(\ell)|\right) \leq \exp(-\alpha^2 C\Omega(1)\ln(n)/2).$$

It readily follows that, for sufficiently large C, this probability can be made $o(n^{-3})$ (say), which suffices to conclude the proof of the Lemma.

Appendix C. Proofs for planted D-ary trees

C.1. Proof of Lemma 17

Proof The property $p_1 = 1$ is trivial. For $h \geq 1$, let Z $Poi(\lambda)$ be the number of children of the root o. Each of the Z children has independently a probability p_h of being the root of a D-ary tree of height h. Therefore, if we define Z_h to be the number of such children, we have

$$\mathcal{L}(Z_h \mid Z) \sim \mathrm{Bin}(Z, p_h).$$

By the splitting property of Poisson random variables, Z_h follows the distribtution $\mathrm{Poi}(\lambda p_h)$. But T contains a D-ary tree of height h rooted in o if and only if $Z_h \geq D$, and the lemma follows. ■

C.2. Proof of Theorem 16

Proof Let $h_0 > 0$ to be fixed later on ; there exists $\kappa > 0$ such that

$$\frac{(\lambda x)^D}{D!} \leq \psi_D(\lambda x) \leq e^{\kappa(D-1)p_{h_0}}\frac{(\lambda x)^D}{D!} \tag{28}$$

for all $x \leq \epsilon$. Therefore, for $h \geq h_0$, one has

$$D\ln(p_h) + (D-1)c_{\lambda,D} \leq \ln(p_{h+1}) \leq D\ln(p_h) + (D-1)(c_{\lambda,D} + \kappa\, p_{h_0}). \tag{29}$$

Iterating inequality (29), we get that for all $h \geq 0$:

$$D^h\left(\ln(p_{h_0}) + c_{\lambda,D}\right) - c_{\lambda,D} \leq \ln(p_{h+h_0}) \leq D^h\left(\ln(p_{h_0}) + c_{\lambda,D} + \kappa\, p_{h_0}\right) - c_{\lambda,D} - \kappa\, p_{h_0} \tag{30}$$

Choose h_0 such that $\alpha := -(\ln(p_{h_0}) + c_{\lambda,D} + \kappa\, p_{h_0}) > 0$, and let

$$h_* = \left\lfloor \frac{\ln\left(\frac{\ln(n)}{\alpha}\right)}{\ln(D)} \right\rfloor + h_0$$

Then $h_* + 1 = \frac{\ln\left(\frac{\ln(n)}{\alpha}\right)}{\ln(D)} + h_0 + \delta$ for some $\delta > 0$. Thus, using (30), we find

$$\ln(p_{h_*+1}) \leq -D^{\delta} \ln(n) - c_{\lambda,D} - \kappa\, p_{h_0}$$

which yields that $p_{h_*+1} = o\left(\frac{1}{n}\right)$ as required.

On the other hand, for almost all λ there is a choice of h_0 such that

$$h_* < \frac{\ln\left(\frac{\ln(n)}{\alpha}\right)}{\ln(D)} + h_0 - \ln\left(\frac{\alpha}{\ln(p_{h_0}) + c_{\lambda,D}}\right)$$

by continuity of the right-hand side. Then, for some $\delta' > 0$, we have

$$\ln(p_{h_*}) \geq -D^{-\delta'} \ln(n) - c_{\lambda,D}$$

which implies the second result of Theorem 16.

■

C.3. Proof of Lemma 20

This lemma is a classical result in sparse random graph theory (see e.g. Bordenave et al. (2015)) ; we reproduce it here for the sake of self-containedness. First, a result on the size of neighbourhoods in G :

Lemma 27 (Lemma 29 in Bordenave et al. (2015)) *For a vertex v in G, let $S_t(v)$ denote the size of the t-neighbourhood of v. Then there exists a constant C such that with high probability, for every vertex $v \in G$ and $t \geq 0$:*

$$S_t(v) \leq C \ln(n) \alpha^t$$

We'll also use a bound on the number of vertices whose neighbourhood contains a cycle ; its proof, as well as the preceding lemma, can be found in Bordenave et al. (2015).

Lemma 28 (Lemma 30 in Bordenave et al. (2015)) *Assume that $\ell = o(\ln(n))$. Then w.h.p there are at most $\ln(n)\lambda^{2\ell}$ vertices whose ℓ-neighbourhood contains a cycle. Moreover, with high probability the graph G is ℓ tangle-free, i.e. no vertex has more than one cycle in its ℓ-neighbourhood.*

We can now prove the first part of our lemma : consider the classical breadth-first exploration process which starts with $A_0 = \{v\}$ and at step $t \leq 0$, considers (if possible) a vertex $v_t \in A_t$ at minimal distance from v and reveals its neighbors N_{t+1} in $[n] \setminus \bigcup_t A_t$. It then updates A_{t+1} as $A_t \cup N_{t+1}$ and repeats the process. We denote by $\mathcal{F}_t$ the filtration generated by $A_0, \ldots, A_t$.

Proof (First part of Lemma 20). Let τ be the stopping time at which $(G, v)_\ell$ has been revealed.By the two previous lemmas, with probability at least $1 - c\lambda^{2\ell}/n$, the neighbourhood $(G, v)_\ell$ is a tree. Therefore, we can mirror the discovery process in (T, o), where at each step we discover the children of v_t. To establish the desired coupling result, we then only need to focus on the number of children of each node.

Given $\mathcal{F}_t$, the number of discovered neighbors y_{t+1} of the node v_t has distribution $\mathrm{Bin}(n_t, \lambda/n)$, where

$$n_t = n - \sum_{s=0}^{t} y_s$$

Therefore, given $\mathcal{F}_t$, the total variation distance between the number of children of v_t in $(G, v)_\ell$ and in $(T, o)_\ell$ is

$$\left|\mathrm{Bin}(n_t, \frac{\lambda}{n}) - \mathrm{Poi}(\lambda)\right|_{\mathrm{var}}$$

The Stein-Chen method (see for example Barbour and Chen (2005)) yields that

$$\left|\mathrm{Bin}(n_t, \frac{\lambda}{n}) - \mathrm{Poi}\left(\lambda\frac{n_t}{n}\right)\right|_{\mathrm{var}} \leq \frac{\lambda}{n},$$

and a classical bound for Poisson law (see again Barbour and Chen (2005)) that

$$\left|\mathrm{Poi}\left(\lambda\frac{n_t}{n}\right) - \mathrm{Poi}(\lambda)\right|_{\mathrm{var}} \leq \lambda\left(1 - \frac{n_t}{n}\right)$$

From Lemma 27, we find that $n_t \geq n - C\ln(n)\lambda^\ell$ with probability greater than $1 - 1/n$, and thus

$$|P_{t+1} - Q_{t+1}|_{\mathrm{var}} \leq \frac{\lambda}{n} + \lambda\frac{C\ln(n)\lambda^\ell}{n},$$

where P_{t+1} is the distribution of y_{t+1} given $\mathcal{F}_t$ and Q_{t+1} is a $\mathrm{Poi}(\lambda)$ random variable independent of $\mathcal{F}_t$. This finishes the proof of the first part of the lemma. ■

For the second part, note that there exists a coupling (X, X') such that $X \sim \mathrm{Poi}(\lambda)$, $X' \sim \mathrm{Poi}(\lambda')$ and $X' > X$ a.s. (take for example $X' = X + Z$ where $Z \sim \mathrm{Poi}(\lambda' - \lambda)$).

The proof is then straightforward : for every vertex v, we produce a coupling between the exploration process of $(G, v)_\ell$ and $(T', o')_\ell$ such that at each step t, the number of neighbors y_t of v_t in G is less than in T'.

C.4. Proof of Theorem 18

Proof We first apply the first part of Lemma 20 to $\ell = \underline{h} = O(\ln\ln(n))$. Then, for at least $n - O(\ln(n)^\alpha)$ vertices v (for some $\alpha > 0$), there is a coupling between $(T, o)_{\underline{h}}$ and $(G, v)_{\underline{h}}$. Since in $(T, o)_{\underline{h}}$, there is a copy of Γ in $(T, o)_{\underline{h}}$ with probability $\Omega(n^{-c})$. It follows that w.h.p there is $\omega(1)$ copies of Γ in G.

Now, assume that $h = \overline{h} + C$, where C is large enough such that for some $\lambda' > \lambda$, there are no trees of height h in (T', o') with probability $1 - o(1/n)$.

For every $v \in G$ such that the h-neighbourhood of v is a tree, we can produce a coupling of $(G, v)_h$ and $(T', o')_h$ such that $(G, v)_h \subseteq (T', o')_h$ with probability 1. Thus, with high probability, no vertex whose h-neighbourhood is a tree contains a copy of Γ in said neighbourhood.

Assume now that there is one cycle in the h-neighbourhood of v. With high probability, there is only one cycle going through v in the neighbourhood. Thus, there are only two vertices in the neighbors of v whose offspring contains a cycle. With probability $1 - O(n^{-c})$, no other neighbour of v is the root of a D-ary tree of height $h-1$. If $D > 2$, then there is no copy of Γ rooted in v ; if $D = 2$, then both neighbors of v in the cycle must be roots of disjoints binary trees of size $h-1$, in which case the cycle edge does not help.

To summarize, the probability of presence of a copy of Γ rooted at v is upper bounded by $o(1/n)$ if the h-neighbourhood of v is cycle-free, and by $O(n^{-c})$ if it is not. Since there are $O(\ln(n)^\alpha)$ such vertices, w.h.p there is no copy of Γ in G. ■

C.5. Proof of Lemma 22

Proof In view of Lemma 5, we aim to bound the ratio

$$\mathbb{E}_0(L^2) = \frac{\mathbb{E}_0(X_\Gamma^2)}{\mathbb{E}_0(X_\Gamma)^2}$$

. As before, let $\Gamma_1, \dots, \Gamma_m$ be the copies of Γ in the complete graph K_n, and let $X_i = \mathbf{1}_{\Gamma_i \in G}$.

We follow the proof sketch from Bollobás (2001) : write

$$\mathbb{E}_0\left(X_\Gamma^2\right) = \sum_{i,j} \mathbb{E}_0\left(X_i X_j\right) = \mathbb{E}' + \mathbb{E}'', \tag{31}$$

where $\mathbb{E}'$ is the sum over Γ_i, Γ_j having disjoint vertex sets.

We can easily compute $\mathbb{E}'$:

$$\mathbb{E}' = \binom{n}{K}\binom{n-K}{K}\left(\frac{K!}{|\mathrm{Aut}(\Gamma)|}\right)^2 p^{2K-2} \sim \frac{n^{2K}p^{2K-2}}{|\mathrm{Aut}(\Gamma)|^2} \sim \mathbb{E}_0\left(X_\Gamma\right)^2$$

We therefore need to show that $\mathbb{E}'' = o\left(\mathbb{E}_0\left(X_\Gamma\right)^2\right)$; to this end, note that if Γ_i and Γ_j are such that $v(\Gamma_i \cup \Gamma_j) = s$, then $e(\Gamma_i \cap \Gamma_j) \leq 2K - s - 1$ (since $\Gamma_i \cap \Gamma_j$ is a forest of size $2K - s$) and therefore $e(\Gamma_i \cup \Gamma_j) \geq s - 1$.

Grouping the terms of $\mathbb{E}''$ by the size of $\Gamma_i \cup \Gamma_j$, we get

$$\begin{aligned}
\mathbb{E}'' &\leq \sum_{s=K}^{2K-1} \binom{n}{s}\binom{s}{s-K, s-K, 2K-s}\left(\frac{K!}{|\mathrm{Aut}(\Gamma)|}\right)^2\left(\frac{\lambda}{n}\right)^{s-1} \\
&= \frac{n}{\lambda|\mathrm{Aut}(\Gamma)|^2}\sum_{s=K}^{2K-1}\frac{n^{\underline{s}}\,\lambda^s}{n^s}\frac{K!^2}{(s-K)!^2(2K-s)!} \\
&= \frac{n}{\lambda|\mathrm{Aut}(\Gamma)|^2}\sum_{s=K}^{2K-1}\lambda^s\frac{K!^2}{(s-K)!^2(2K-s)!}\left(1 + O\left(\frac{K^2}{n}\right)\right) \\
&\leq \frac{n\lambda^{K-1}(1+o(1))}{|\mathrm{Aut}(\Gamma)|^2}\sum_{u=0}^{K-1}\lambda^u\frac{K!^2}{u!^2(K-u)!},
\end{aligned}$$

where we made the change of variables $u = s - K$. Now, write

$$\frac{K!^2}{u!^2(K-u)!} = \binom{K}{u}\frac{K!}{u!} \leq \binom{K}{u}K^{K-u},$$

and we get

$$\begin{aligned}
\mathbb{E}'' &\leq \frac{n\lambda^{K-1}K^K}{|\mathrm{Aut}(\Gamma)|^2}(1+o(1))\sum_{u=0}^{K-1}\binom{K}{u}\left(\frac{\lambda}{K}\right)^u \\
&\leq \frac{n\lambda^{K-1}K^K}{|\mathrm{Aut}(\Gamma)|^2}(1+o(1))\left(1+\frac{\lambda}{K}\right)^K \\
&\leq \frac{n\lambda^{K-1}K^K e^{\lambda}}{|\mathrm{Aut}(\Gamma)|^2}(1+o(1)) \\
&= O\left(\mathbb{E}_0(X_\Gamma)^2 \times \frac{K^K}{n\lambda^K}\right)
\end{aligned}$$

When $K \leq \frac{\ln(n)}{\ln\ln(n)}$, we find that $\mathbb{E}'' = o\left(\mathbb{E}_0[X_\Gamma]^2\right)$, as requested. But $K = \frac{D^{h+1}-1}{D-1} \leq \frac{\ln(n)}{\ln\ln(n)}$ whenever

$$h \leq \underline{h} - \frac{\ln(\underline{h})}{\ln(D)} + \frac{\ln\left(1-\frac{1}{D}\right)}{\ln(D)},$$

which is the condition mentioned in Theorem 21.

■

C.6. Proof of Theorem 24

Proof For $0 \leq p \leq h$, let L_p be the set of vertices at depth p of Γ, and T_p the set of vertices at depth $\leq p$.

The strategy of proof is as follows : we aim to prove that there exists a universal constant δ such that given G and

$$\mathcal{T} := \sigma(T_{h-1}) \subset G,$$

the location of the first $h-1$ rows of Γ, we have with high probability on G

$$\mathbb{P}_1\left((\mathrm{ov}(\hat{\mathcal{K}}) \leq (1-\delta)K \;\middle|\; G, \mathcal{T}\right) = 1 - o(1) \tag{32}$$

In what follows, we will consider $\mathcal{T}$ to be fixed, and G drawn under $\mathbb{P}_1$.

Let $\varepsilon > 0$ to be adapted later, and consider two cases :

- $|\hat{\mathcal{K}} \cap \mathcal{T}| \leq (1-\varepsilon)|\mathcal{T}|$: in this case, we easily get

$$\begin{aligned}
\mathrm{ov}(\hat{\mathcal{K}}) &\leq D^h + (1-\varepsilon)\frac{D^h-1}{D-1} \\
&= K - \varepsilon\frac{D^h-1}{D-1} \\
&= K - \varepsilon\frac{K-1}{D} \\
&= (1-\frac{\varepsilon}{D})K + o(K),
\end{aligned}$$

from which equation (32) follows since ε is independent from G and $\mathcal{T}$.

- if $|\hat{\mathcal{K}} \cap \mathcal{T}| > (1-\varepsilon)|\mathcal{T}|$, we need the following lemma :

Lemma 29 *Let $\sigma(L_{h-1}) = \{i_1, \dots i_{D^{h-1}}\}$, and define $n_k = |\mathcal{N}(i_k)|$ and $m_k = |\hat{\mathcal{K}} \cap \mathcal{N}(i_k)|$. Then*

$$\mathbb{E}_1\left(|\hat{\mathcal{K}} \cap \sigma(L_h)| \,\middle|\, G, \mathcal{T}\right) = D\sum_k \frac{m_k}{n_k}$$

Proof (of lemma 29). Given $\mathcal{T}$, all vertices that are neighbours of a vertex in $\sigma(L_{h-1})$ are equally likely to belong to Γ, since all D-ary trees in G have the same probability of generating G.

Therefore, given G, $\sigma(L_{h-1}) = \{i_1, \dots i_{D^{h-1}}\}$, the random variable $N_k = |\hat{\mathcal{K}} \cap \mathcal{K} \cap \mathcal{N}(i_k)|$ follows a hypergeometric law of parameters (n_k, D, m_k). If follows that

$$\mathbb{E}_1(N_k) = D\frac{m_k}{n_k}$$

Now, with high probability the neighbourhoods $\mathcal{N}(i_k)$ are disjoint and the variables N_k are thus independent. Since $|\hat{\mathcal{K}} \cap \sigma(L_h)| = \sum_k N_k$ whenever the $\mathcal{N}(i_k)$ are disjoint, the lemma follows. ■

We can now prove our main theorem : notice that $|\mathcal{N}(i_k)| \sim D + \mathrm{Poi}(\lambda)$ since $K = o(n)$, so w.h.p a proportion α (for a universal constant α) of the i_k are such that $|\mathcal{N}(i_k)| \geq D+1$. Moreover,

$$S := \sum_k m_k = K - |\hat{\mathcal{K}} \cap \mathcal{T}| < D^h + \varepsilon|\mathcal{T}| = (1 + \frac{\varepsilon}{D-1})D^h + o(D^h)$$

Thus, $S \leq (1+\varepsilon')D^h$ for some $\varepsilon' > 0$.

Let I_1 be the set of indices such that $n_k = D$; we have

$$\begin{aligned}\sum_k \frac{m_k}{n_k} &= \sum_{k\in I_1} \frac{m_k}{n_k} + \sum_{k\notin I_1} \frac{m_k}{n_k} \\ &\leq \sum_{k\in I_1} \frac{m_k}{D} + \sum_{k\notin I_1} \frac{m_k}{D+1}\end{aligned}$$

Let $S_1 = \sum_{k\in I_1} m_k$; we know that

$$S_1 \leq D|I_1| \leq D(1-\alpha)D^{h-1},$$

since $m_k \leq n_k = D$ on I_1, which yields

$$\begin{aligned}\sum_k \frac{m_k}{n_k} &\leq \frac{S_1}{D} + \frac{S-S_1}{D+1} \\ &= \frac{S}{D+1} + \frac{S_1}{D(D+1)} \\ &\leq D^{h-1}\left((1+\varepsilon)\frac{D}{D+1} + (1-\alpha)\frac{1}{D+1}\right) \\ &\leq D^{h-1}\left(1 - \frac{\alpha - D\varepsilon}{D+1}\right)\end{aligned}$$

Choosing ε such that $\alpha - D\varepsilon > 0$, we eventually find

$$\mathbb{E}_1\left(|\hat{\mathcal{K}} \cap L_h| \,\middle|\, G, \mathcal{T}\right) \leq (1-\gamma)D^h \tag{33}$$

for some $\gamma > 0$.

Finally, we can bound $\hat{\mathcal{K}} \cap \mathcal{K}$:

$$\begin{aligned}
\mathbb{E}_1\left(|\hat{\mathcal{K}} \cap \mathcal{K}| \,\middle|\, G, \mathcal{T}\right) &\leq |\mathcal{T}| + \mathbb{E}_1\left(|\hat{\mathcal{K}} \cap L_h| \,\middle|\, G, \mathcal{T}\right) \\
&\leq (1-\gamma)D^h + |\mathcal{T}| \\
&\leq K - \gamma D^h + o(D^h) \\
&\leq (1 - \gamma\frac{D-1}{D})K + o(K),
\end{aligned}$$

which completes the proof of Theorem 24. ■