Rapport de stage

Francisco Calvillo

2021

Contents

1 Déroulement du stage

Malgré l'incertitude et les difficultés imposées par la crise sanitaire tout le long de l'année scolaire, j'ai eu la chance de pouvoir partir en stage à l'étranger du 9 février au 9 juin 2021. Ayant toute ma famille en Espagne et ayant la nationalité espagnole, j'ai pensé que ça pouvait être une bonne idée de chercher un stage dans une université espagnole: il me serait beaucoup plus facile de m'adapter aux mesures prises pour freiner le COVID, voire même en cas de reconfinement. Étant intéressé par le domaine des probabilités et des statistiques je me suis addressé à Rémy Mahfouf et à Giambattista Giacomin, qui m'ont vivement recommandé de réaliser mon stage à Barcelone sous la direction de Gabor Lugosi et m'ont mis en contact avec lui.

1.1 L'Université Pompeu Fabra

Je suis arrivé à Barcelone en début février pour m'installer dans une résidence universitaire qui se trouve devant le parc de La Ciutadella, à moins de cinc minutes à pied du quartier gothique et à quelques pas de l'Université Pompeu Fabra (UPF), où se trouve le bureau de Gabor. Même s'il n'existe pas un département de mathématiques à l'UPF, Gabor travaille avec toute une équipe de mathématiciens au département d'économie qui m'ont très bien accueilli. Aucun bureau n'était disponible pour moi au département d'économie, mais cela n'a posé aucun problème puisque j'avais la chance de pouvoir travailler à la bibliothèque de l'université, le Dipòsit de les Aigües: il s'agit d'un ancien château d'eau qui a été réhabilité en 1999 et qui fait partie du patrimoine culturel catalan.

1.2 Vie quotidienne et travail

L'objet de ce stage était de m'initier à la recherche mathématique. Avant mon arrivée, Gabor m'a montré un article [1] qu'il avait écrit il y a quelques années et qui parlait "d'archéologie" dans des arbres aléatoires, et m'a proposé d'étudier ce même sujet mais cette fois-ci dans un cas particulier de graphes aléatoires. C'est ce sur quoi nous avons travaillé pendant mes quatre mois de stage.

Au début nous travaillions tous les deux, en discutant de temps en temps avec les collègues de Gabor qui nous aidaient à nous débloquer. Au bout du premier mois, Simon Briend, un étudiant d'Orsay en M2, est venu pour faire lui aussi un stage sous la direction de Gabor. Intéressé par notre sujet de recherche, il a commencé à travailler avec nous.

En général, les matins je travaillais avec Simon à la bibliothèque de l'université, puis quasiment tous les après-midi nous nous réunissions quelques heures avec Gabor dans son bureau pour lui montrer nos progrès (lorsque nous en avions) et réfléchir au tableau avec lui. Je tiens à remercier Gabor d'avoir pris le temps pour que nous puissions nous réunir aussi régulièrement, ce fut un véritable plaisir d'être encadré par lui.

À plusieurs reprises, Gabor a dû voyager pour des courtes périodes, pendant lesquelles nous avons décidé de télétravailler (ce qui m'a permis de visiter ma famille à Valencia de temps en temps).

N'ayant pas de département de mathématiques à l'UPF, et s'agissant d'une année exceptionnelle à cause du COVID, il n'y avait pas vraiment de cours mathématiques intéressants que je pouvais suivre. Cependant, j'ai pu assister à un groupe de lecture, formé par plusieurs chercheurs de l'université et organisé par l'un des collègues de Gabor, Piotr Zwiernik, dans lequel chaque semaine nous décortiquions un chapitre du livre *High-Dimensional Statisctics* [2]. La plupart des présentations devaient se faire en visioconférence, mais nous avons pu nous réunir dans une salle pour faire quelques séances d'exercices. Malheureusement, le groupe de lecture a commencé plus tard que prévu, et nous n'avons pas pu traiter le livre dans sa totalité.

2 Travail mathématique effectué

Le problème proposé par Gabor pour ce stage de recherche est particulièrement inéressant: non seulement il s'agit d'un sujet qui pourrait être utile, par exemple, dans le monde des algorithmes de recommandation dans des plateformes numériques, il ne nécessite pas de prérequis mathématiques trop avancés, ce qui m'a permis de me lancer à réflechir et à me familiariser avec le problème dès le premier jour. Cependant, trouver une solution à celui-ci nous a pris plus de temps de ce qu'on avait prévu, et ce n'est que quelques jours après la fin oficielle de mon stage que nous en avons trouvé une. Le texte mathématique qui suit est en quelque sorte une prépublication faite en commun avec Simon Briend pendant ce mois d'août, qui résume les progrès que nous avons fait au bout de ces 4 mois de stage. Nous aimerions étendre nos résultats à d'autres modèles de graphes aléatoires similaires comme le modèle d'attachement préférentiel, pour ensuite le publier.

References

[1] Sébastien Bubeck, Luc Devroye, and Gábor Lugosi. *Finding Adam in random growing trees*. 2015. arXiv: 1411.3317 [math.PR].

[2] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. ISBN: 1108498027.

Graph archeology

Simon Briend, Francisco Calvillo,
under the supervision of Gabor Lugosi

Abstract

In this work we present a method to find the first vertex of a random graph randomly growing according to a complexified uniform attachment model, here we add from a non-uniform probability more edges on top of the edges from an uniform attachment model. Our goal is that given ϵ we output a set of $K = K(\epsilon)$ vertices such that with probability at least $1 - \epsilon$ the first vertex is in this set. We want to make sure that K does not depend on the size of the graph. Using a double-cycle counting we achieve exactly that in our base model and also in an extra-case, where we remove the uniform attachment tree and only observe the edges sampled from our non-uniform probability.

1 Introduction

The first part of our work is based on an extension of the uniform attachment model. In this model, we start at time $t = 1$ with a single vertex, labeled 1, and at each step time $t \in \mathbb{N},\ t \geq 2$, we add the vertex t and the edge $\{i; t\}$ such that i is uniformly drawn from the $t - 1$ previous vertices: this is called a uniform random recursive tree (URRT). In our model we add some complexity. When our uniform attachment model reaches size n, what we observe is this graph on which are added some edges. Those edges are added as follows:

- We fix a constant $C_p \in (0, +\infty)$.
- For each pair of vertices i, j we draw Bernoullis independent of each other and of the branching process, of parameter $\frac{C_p}{\max(i,j)}$.
- If the Bernoulli realisation's is 1 we add the edge i, j if there were no edges i, j from the URRT. Those added edges will be called lianas.

What we observe is the unlabelled graph, and we cannot differentiate lianas from edges coming from the URRT. Then what differs from the URRT model is mainly that we lose the tree structure. Actually most of the statistics used to infer the root in an URRT used the tree structure, for example in [2] where one of the statistics is a measure of Jordan centrality in trees: looking at the vertices such that his biggest subtree is the smallest. However, even if in our model the tree structure is lost, we know from our model that our graph is the superposition of a tree sampled from URRT on which some edges had been added during the branching process. This view is close from the one presented in [3], and will be extensively used to understand the behaviour of the graph. But unlike in [3] we do not see the added edges as noise, on the opposite as they are added with non-uniform probability (they are more likely to connect old edges) that will add information for the root finding, rather than suppressing it.

In a second part we will extend our results to the case where there is not even an underlying tree in the graph, but only edges sampled according to the probability $C_p/\max(i,j)$. When $C_p > 1$ we are able to retrieve a set containing the root with high probability. We will use almost the same searching method as before but as one can expect by the removal of the tree, which added a lot of information, our method will be less performing and will result in $K(\epsilon)$ being bigger than in the previous model.

Let us remark that the choice of the parameter C_p has a mathematical origin and is not meaningless. One way to build a random graph having almost the same law as this one is the sequential branching model built as follow:

- Chose a parameter $p \in (0,1)$.
- Start with the vertices 1 and 2 connected by an edge.
- At each time we sample a Bernoulli of parameter p and we rather: -If the Bernoulli sample is 1: add a vertex and an edge connecting it to a uniformly chosen older vertex. - Otherwise add an edge uniformly chosen amongst all the possible edges of the graph (if we draw an already existing edge we do nothing).

Using this branching model we can show that as the size of the network grows, the probability of a liana (i,j) to exist is tending to $C_p/\max(i,j)$ for i and j indexes big enough, where $C_p = 2(1-p)/p$. This is where the expression of C_p comes from.

1.1 Related work

Random graphs have long been studied but root-finding algorithms look to be a more recent center of interest. It was investigated for uniform attachment model and preferential attachment model in 2015 by Bubeck, Devroye and Lugosi [2]. In this paper they give lower-bound on the size of $K(\epsilon)$ in a general case but also give results if they impose a polynomial time root finding algorithm. One of the key statistics there is Jordan centrality, that can not be used in our case. Going closer to our model was a paper from Crane and Xu in June 2021, [3], where they consider an preferential attachment model on top of which they add an Erdős–Rényi (when we add edges with non-uniform probability on top of an uniform attachment model). In this paper they see the added edges as noise and retrieve the root by creating a prior distribution on which they use Bayesian inference, before proving this Bayesian approach has frequentist properties.

However our methods are closer to the ones from Bubeck, Devroye and Lugosi as we rely on a measure (in our case being root of a double cycle) to infer the root. One of the first intuitions in those models would be to look at the degree centrality, as we expect the older vertices to have bigger degrees. In the case of the uniform attachment model, Devroye and Lu showed in 1995 [4] that it's not a way to retrieve the root as there exist a vertex with significantly bigger degree than the root. Doing the exact same computations as they use in their paper we could show that in our model, as the probability of the lianas are just C_p times the probability of the edges from an URRT, some vertices have significantly higher degree than the root.

1.2 Notation

In this whole work we will use the following notations. First of all we will call lianas the edges sampled on top of the URRT (as in the jungle liana are running from branches to branches, here it is the same, lianas are added on top of the tree). This vocabulary trick will make it easier to describe our ideas. For example we introduce $d_l(i)$, the degree of vertex i amongst the lianas, meaning the number of lianas connected to vertex i.

G will always denote the observed graph; while $DC_k(I)$ will denote the event: I is a root of double cycle of size smaller than k and $C_{K,L}(I)$ will be the number of valid double cycles rooted in I. Finally, $\rho(G)$ will denote the output of our root finding algorithm.

2 The URRT plus liana model

2.1 The double cycle statistics

We must define properly the statistics we want to consider. The idea is to consider double cycles and their root (the intersection point between two cycles). However we must be more precise regarding how we characterize double cycles, how we allow for common edges and vertices, what are the roots if several points are in both cycles. To do so let us focus on what kind of double cycles will have 1 as a root and define them as our valid double cycles.

Some double cycles with 1 as a root will appear as follow: if we consider a number of vertices M great enough we will be sure that 1 has at least two lianas edges linking it to the first M vertices. As there is a tree rooted in 1 those two lianas create two cycles that both contain 1, but these two cycles can intersect only in a very precise way.

If they exist, let us consider $(1, i)$ and $(1, j)$ two lianas. In the tree there is a single path from 1 to i and a single one from 1 to j. Those path can only have in common their beginning (first edges and vertices starting from 1, in the same order in both paths). Thus the union of those paths and of the lianas create two cycles, both containing the root 1 and being entangled in a very nice way:

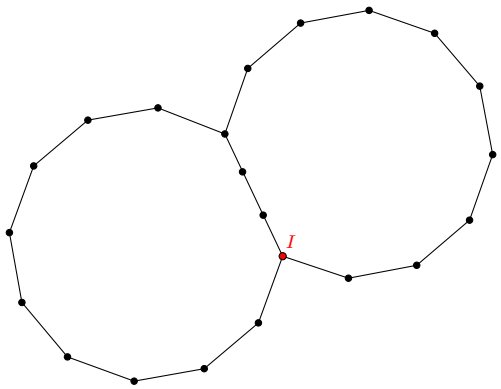
Definition 1 *In a graph G, we say that I is a root of a valid double cycle of size k, l if there exist $k+l-2-p$ different vertices $i_1, i_2, \dots i_{k+l-2-p}$, such that:*

- $I \to i_1 \to \cdots \to i_{k-1} \to I$ *is a cycle in the G.*
- $I \to i_{k-1-p} \to \cdots \to i_{k+l-2-p} \to I$ *is a cycle in G.*

The two cycles are disjoint, except for the common part $I \to i_{k-1-p} \to \cdots \to i_{k-1}$.

Remark: given this definition each valid double cycle can have two roots, one at both ends of the path in common in the two cycles. As an illustration here is the structure of the valid double cycles:

Figure 1: Illustration of a valid double cycle



It is now time to check that the statistic "being the root of a valid double cycle of small size" (small size will be defined in next section) is an actual way to retrieve the older vertex of the graph. We will denote $DC_k(I)$ the event of I being a root of a valid double cycle of size smaller than k.

2.2 Being sure the root is in a double cycle of small size

We want to make sure that our statistic is positive for the actual root of the graph. The idea will be to make sure that 1 is in at least one double cycle center of size M_ϵ with probability $1-\epsilon$. To do so we will use the tree structure of the URRT and the added vertex.

Since the tree part forms a fully connected component, as soon as two liana vertices starting from 1 are added we know that 1 will be a root of a valid double cycle. Thus we need to find k_ϵ such that:

$$\mathbb{P}\left[\sum_{i=2}^{k_\epsilon} \mathbb{I}\left(1 \underset{liana}{\rightarrow} i\right) \geq 2\right] \;\geq\; 1-\epsilon.$$

To do so let us use some concentration inequality from [1]. Let us denote $X_k = \sum_{i=2}^{k} \mathbb{I}\left(1 \underset{liana}{\rightarrow} i\right)$ We can upper bound:

$$\mathbb{P}\left[X_k \leq 2\right] \;\leq\; \exp\left(-\frac{\left(\mathbb{E}(X_k)-2\right)^2}{2\mathbb{E}(X_k)}\right).$$

As $\mathbb{E}(X_k) = \sum_{i=2}^{k} \frac{C_p}{i}$ we have:

$$C_p \left(\log(k) - 2\right) \leq \mathbb{E}(X_k) \leq C_p \log(k),$$

which leads to:

$$\mathbb{P}\left[X_k \leq 2\right] \;\leq\; \exp\left(-C_p/2 \;\log(k) + 4\right).$$

Thus, for $k_\epsilon = e^{4/C_p} \times \left(\frac{2}{\epsilon}\right)^{1/C_p}$ we have:

$$\mathbb{P}\left[\sum_{i=2}^{k_\epsilon} \mathbb{I}\left(1 \underset{liana}{\rightarrow} i\right) \geq 2\right] \;\geq\; 1 - \epsilon.$$

Now that we know that the vertex 1 is in a valid double cycle amongst the k_ϵ vertices we need to find out what are the sizes of those double cycles. A very naive approach would be to say they are of size at most k_ϵ but we know that the tree part of the graph has a diameter growing way slower than linear in the number of vertices. In fact we know from lemma 6 that the diameter of an URRT graph of k_ϵ vertices has a depth smaller than $\alpha C_p \log(k_\epsilon)$ with probability greater than $1 - \epsilon$.

We can consider $c > 0$ independent of C_p and ϵ such that, taking the intersection of those events, with probability greater than $1 - 2\epsilon$, 1 is a root of a valid double cycle of size at most $M_\epsilon := c \times \log(1/\epsilon)$.

2.3 Great indexes are not in double cycles

2.3.1 The number of large vertices being root of double cycles

First we want to upper-bound the expectation of the number of valid double cycles having I as a root. Let us give ourselves two positive integers $K < L$. We will denote $C_{K,L}(I)$ the number of valid double cycles having I as a root and with cycle of size K and L. We have:

$$C_{K,L}(I) = \underbrace{\sum_{p=0}^{K-1}}_{\substack{nbr\ common \\ edges}} \sum_{\substack{i_1 < \cdots < \\ i_{K+L-2-p}}} \sum_{\sigma \in S_{K+L-2-p}} \mathbb{I}\left(I \rightarrow i_{\sigma(1)} \ldots i_{\sigma(K-1)} \rightarrow I \rightarrow i_{\sigma(K-p)} \ldots i_{\sigma(K+L-2-p)} \rightarrow I\right).$$

So:

$$\mathbb{E}\left[C_{K,L}(I)\right] \;=$$

$$\sum_{p=0}^{K-1} \sum_{\substack{i_1<\cdots< \\ i_{K+L-2-p}}} \sum_{\substack{\sigma\in \\ S_{K+L-2-p}}} \mathbb{P}\left(I \to i_{\sigma(1)} \dots i_{\sigma(K-1)} \to I \to i_{\sigma(K-p)} \dots i_{\sigma(K+L-2-p)} \to I\right).$$

We want to upper-bound the probability of those double cycles. It would be very easy if every edge were independent but this is not the case as we know that part of the graph is a tree. However, using both:

- that for two independent Bernoulli B_1, B_2 of parameter p and q there exist a Bernoulli of parameter $p+q$ such that $\mathbb{I}\left(B_1 + B_2 \geq 1\right) \leq B_3$.
- a negative correlation between the probability of edges (see 2) and independence of the lianas from all the other random variables.

we are sure that the probability of the double cycles are upper-bounded by the product of the probability of the same double cycle if each edge was drawn independently from a binomial of parameter $(1+C_p)/\max(i,j)$.

Thus we can use 4.3 with parameter C_p+1 to be sure that the most likely double cycle, in a setting where each edge is sampled from a Bernoulliof parameter $(C_p+1)/\max(i,j)$, is the monotonic one (again, see 4.3 for details).

So, denoting $\mathbb{P}_{ind}$ the probability measure in the setting where each edge is sampled from independent Bernoulli of parameter $(C_p+1)/\max(i,j)$, we have:

$$\mathbb{E}\left[C_{K,L}(I)\right] \;\leq$$

$$\sum_{p=0}^{K-1} (K+L-2-p)! \sum_{\substack{i_1<\cdots< \\ i_{K+L-2-p}}} \mathbb{P}_{ind}\left(I \to i_1 \dots i_{K-1} \to I \to i_{K-p} \dots i_{K+L-2-p} \to I\right).$$

We now have to get into this computation which is easy but quite arduous to write, for fixed p we have:

$$\sum_{\substack{i_1,\dots, \\ i_{K+L-2-p}}} \mathbb{P}_{ind}\left(I \to i_1 \dots i_{K-1-p} \to i_{K+L-2-2p} \dots i_{K+L-2-p} \to I \to i_{K-p} \dots i_{K+L-2-p} \to I\right) \leq$$

$$(C_p+1)^{K+L-2-p} \sum_{\substack{I<i_1,\dots, \\ i_{K+L-2-p}}} \left(\frac{1}{i_{K+L-2-p}} \frac{1}{i_{K+L-2-2p}} \prod_{i=1}^{K+L-2-p} \frac{1}{i_j}\right) + \qquad (A)$$

$$(C_p+1)^{K+L-2-p}\sum_{m=1}^{K-p-1}\sum_{\substack{i_1,\dots i_m<I<i_{m+1},\\ \cdots i_{K+L-2-p}}}\frac{1}{I}\frac{1}{i_{K+L-2-2p}}\frac{1}{i_{K+L-2-p}}\prod_{j=2}^{K+L-2-p}\frac{1}{i_j}+ \qquad (B)$$

$$(C_p+1)^{K+L-2-p}\sum_{m=K-p}^{K+L-2-p}\sum_{\substack{i_1,\dots i_m<I<i_{m+1},\\ \cdots i_{K+L-2-p}}}\frac{1}{I^2}\frac{1}{i_{K+L-2-2p}}\frac{1}{i_{K+L-2-p}}\prod_{j=2}^{K-p-1}\frac{1}{i_j}\prod_{j=K-p+1}^{K+L-2-p}\frac{1}{i_j}+ \quad (C)$$

$$(C_p+1)^{K+L-2-p}\sum_{\substack{i_1,\dots,\\ i_{K+L-2-p}<I}}\frac{1}{I^3}\frac{1}{i_{K+L-2-2p}}\prod_{j=2}^{K-p-1}\frac{1}{i_j}\prod_{j=K-p+1}^{K+L-2-p}\frac{1}{i_j}. \qquad (D)$$

Then we can upper-bound each term. For A we end up summing things to the third power so the rest of the sum is of order $1/I^2$.

$$A \leq (C_p+1)^{K+L-2-p}\frac{1}{I^2}.$$

For B the terms until i_m sum to I, and the terms after i_{m+1} are like in A summing to $1/I^2$. Thus we end up having:

$$B \leq (C_p+1)^{K+L-2-p}(K-1-p)\frac{1}{I^2}.$$

For C the terms before i_m once again sum to I but this time the terms after i_{m+1} are squared so the rest of the sum is of order $1/I$. Anyway we end up having:

$$C \leq (C_p+1)^{K+L-2-p}(L-1)\frac{1}{I^2}.$$

Finally for D the first indexes sum to at most I so we end up again with:

$$D \leq (C_p+1)^{K+L-2-p}\frac{1}{I^2}.$$

Putting it all together we have:

$$\mathbb{E}\left[C_{K,L}(I)\right] \leq \sum_{p=0}^{K-1}(C_p+1)^{K+L-2-p}(K+L-p)!\frac{1}{I^2} \leq 2(C_p+1)^{K+L}\frac{(K+L)!}{I^2}.$$

2.3.2 Large indexes are not roots of valid double cycles

Now that we control the expectation of $C_{K,L}(I)$ we can show that great indexes will not be positive to our statistic. We recall that our statistic for graph archeology is to output roots of double cycles of size smaller than $M_\epsilon = c \times \log(1/\epsilon)$. Then, using union bound:

$$\mathbb{P}\left(\exists I \geq K,\ s.t.\ DC_{M_\epsilon}(I)\right) \ \leq \ \sum_{I \geq K} \sum_{k,l \leq M_\epsilon} \mathbb{P}(C_{k,l}(j) \geq 1),$$

And Markov's inequality:

$$\mathbb{P}\left(\exists I \geq K,\ s.t.\ DC_{M_\epsilon}(I)\right) \ \leq \ \sum_{I \geq K} \sum_{k,l \leq M_\epsilon} 2\left(C_p+1\right)^{k+l} \frac{(k+l)!}{I^2}.$$

And finally we yield:

$$\mathbb{P}\left(\exists I \geq K,\ s.t.\ DC_{M_\epsilon}(I)\right) \ \leq \ 4\left(C_p+1\right)^{2M_\epsilon} (2M_\epsilon)! \frac{1}{K}.$$

Thus we can assure that with probability greater than $1-\epsilon$ our method outputs at most:

$$K_\epsilon := 4\frac{1}{\epsilon}\left(C_p+1\right)^{2M_\epsilon} (2M_\epsilon)!\ ;$$

which is of order $c' \log(1/\epsilon)^{c' \log(1/\epsilon)}$, for some $c' > 0$.

2.4 The final result

Putting it all together we know that if we chose $M_\epsilon = c \times \log(1/\epsilon)$ for some constant $c > 0$, and look at the set $Out(\epsilon)$ of all the I such that $DC_{M_\epsilon}(I)$, then:

- With probability greater than $1-2\epsilon$, we output the root, meaning that $1 \in Out(\epsilon)$.
- With probability greater than $1-\epsilon$ we output less than $K(\epsilon) = 4\frac{1}{\epsilon}\left(C_p+1\right)^{2M_\epsilon} (2M_\epsilon)!$ points, meaning $|Out(\epsilon)| \leq K(\epsilon)$.

Thus, we chose as a final output :

- If $|Out(\epsilon)| \leq K(\epsilon)$, then $\rho_\epsilon(G) = Out(\epsilon)$
- Otherwise to construct $\rho_\epsilon(G)$ we randomly pick $K(\epsilon)$ points from $Out(\epsilon)$, with any random law.

With such an output $\rho_\epsilon(G)$ we output a set of at most $K(\epsilon)$ vertices, containing the root with a probability higher than $1-3\epsilon$ (as this event holds in the intersection of two events of probability higher than $1-\epsilon$ and $1-2\epsilon$).

3 Extension to the "no tree" model

In this section, we will consider the same model but without the tree part. This means that in our graph G, for each pair of vertices i, j the edge $\{i, j\}$ exists with a probability $\frac{C_p}{\max(i,j)}$, independently from any other edge. Without the tree structure, the argument that has been given in 2.2 to show that the first vertex is the root of a double cycle does not hold anymore.

3.1 There is a double cycle whose root is the first vertex

The idea is to show that the root is connected with at least 3 edges to some connected set S in $\{2, ..., M\}$: if this happens, then we know that 1 is in the center of some double cycle of size at most $|S|$.

3.1.1 Existence of a big connected set

Given an integer $M \geq 1$, we want to show that there is a big connected component in the subgraph $\{2, ..., M\}$. To do so, we will use some well known results concerning the size of the connected components in the Erdös-Rényi model, and use the fact that, from the proper point of view, the subgraph $\{2, ..., M\}$ contains an Erdös-Rényi random graph.

We may assume that for any pair of vertices $i, j \in \{2, ..., M\}$ the edge $\{i, j\}$ is represented by a Bernoulli $X_{i,j}$ of parameter $\frac{C_p}{\max(i,j)}$ that can be written as follows:

$$X_{i,j} = \mathbb{I}_{X^{(1)}_{\hat{\sigma}(i),\hat{\sigma}(j)} + X^{(2)}_{i,j} \geq 1}$$

where $X^{(1)}_{i,j}$ is a Bernoulli of parameter $\frac{C_p}{M}$, $X^{(2)}_{i,j}$ is a Bernoulli of parameter $\frac{C_p(\frac{1}{\max(i,j)} - \frac{1}{M})}{1-\frac{C_p}{M}}$ and $\hat{\sigma}$ is a uniform random permutation of the set $\{2, ..., M\}$

such that the set of random variables $\{\hat{\sigma}\} \cup \{X_{i,j}^{(1)}, X_{i,j}^{(2)}\}_{1 \leq i,j \leq M}$ is mutually independent, and it is independent from any other edge of the graph.

Now, consider the random graph whose vertices are $\{2, ..., M\}$ and whose edges $\{i, j\}$ are given by the random variables $X_{i,j}^{(1)}$, for any $2 \leq i, j \leq M$. By construction, this is a random graph $G(M, \frac{C_p}{M})$ and its image under the mapping $\hat{\sigma}^{-1}$ is a subgraph of our initial graph G. Let us use the following result from [6]

Lemma 1 *Assume $C_p > 4 \log 2$. Let Y_M be the size of the biggest component of $G(M, \frac{C_p}{M})$ and let α be a positive number that satisfies $\alpha < 1 - \frac{4 \log 2}{C_p}$.*

Then

$$\mathbb{P}(Y_M < \alpha M) = O(1) \left(\frac{2}{\exp(\frac{C_p(1-\alpha)}{4})} \right)^M.$$

This means that, if $C_p > 4 \log 2$, then with probability at least $1 - \epsilon$ we can find a subset of αM vertices in $\{2, ..., M\}$ that is connected in $G(M, \frac{C_p}{M})$, and therefore its image by $\hat{\sigma}^{-1}$ is also connected in our initial graph, as long as $M \geq c \log(1/\epsilon)$ for some constant $c > 0$.

From now on, in this section, we will assume that $C_p > 4 \log 2$ and we will fix some constant $0 < \alpha < 1 - \frac{4 \log 2}{C_p}$ independent of M and ϵ.

3.1.2 The root is connected with at least 3 edges to some connected set

Let A denote the event "$Y_M \geq \alpha M$". Assuming that A occurs, we can define a random variable $\hat{S}$ by choosing randomly and uniformly one set amongst all sets of αM vertices that are connected in $G(M, \frac{C_p}{M})$. Since in $G(M, \frac{C_p}{M})$ all edges exist with the same probability, we know that all vertices in $\{2, ..., M\}$ have the same probability to be in $\hat{S}$.

In order to show that the root is connected with at least 3 edges to the set $\hat{\sigma}^{-1}\hat{S}$, first we want the root to be connected to a lot of vertices in $\{2, ..., M\}$. Now, we use the concentration inequality from [1] used before in 2.2:

$$\mathbb{P}\left[\sum_{j=2}^{M} X_{1,j} \leq \frac{C_p}{2} \log M \right] \leq \exp\left(- \frac{(\sum_{j=2}^{M} \frac{C_p}{j} - \frac{C_p}{2} \log M)^2}{2 \sum_{j=2}^{M} \frac{C_p}{j}} \right)$$
$$\leq \lambda^{C_p/8} M^{-C_p/8}$$

for some constant $\lambda > 0$ independent from C_p. Thus, for $M \geq \lambda (\frac{1}{\epsilon})^{8/C_p}$ we have:

$$\mathbb{P}\left[\sum_{j=2}^{M} X_{1,j} \geq \frac{C_p}{2} \log M \right] \geq 1 - \epsilon.$$

Let B denote the event "$\sum_{j=2}^{M} X_{1,j} \geq \frac{C_p}{2} \log M$". Assuming that B occurs, we can define a random set $\hat{R}$ by choosing uniformly $\frac{C_p}{2} \log M$ vertices amongst all vertices in $\{2, ..., M\}$ that are connected to the root. The sum $\sum_{j=2}^{M} X_{1,j} \mathbb{I}_{j \in \hat{\sigma}^{-1}\hat{S}}$ counts the number of edges connecting the root to the set $\hat{\sigma}^{-1}\hat{S}$, which is connected in the initial graph G. First, we have:

$$\sum_{j=2}^{M} X_{1,j} \mathbb{I}_{j \in \hat{\sigma}^{-1}\hat{S}} \geq \sum_{j=2}^{M} \mathbb{I}_{j \in \hat{R}} \mathbb{I}_{j \in \hat{\sigma}^{-1}\hat{S}} = \sum_{j=2}^{M} \mathbb{I}_{\hat{\sigma}^{-1}(j) \in \hat{R}} \mathbb{I}_{j \in \hat{S}}$$

Note that the probability

$$\mathbb{P}[X_{1,\hat{\sigma}^{-1}(j)} = 1] = \sum_{i=2}^{M} \frac{1}{M-1} \mathbb{P}\big[X_{1,i} = 1 \big| \hat{\sigma}^{-1}(j) = i\big] = \frac{1}{M-1} \sum_{i=2}^{M} \frac{C_p}{i}$$

does not depend on j. Thus, the random vertices $\hat{\sigma}^{-1}(2), ..., \hat{\sigma}^{-1}(M)$ are in $\hat{R}$ with the same probability. Moreover, $\hat{S}$ is independent from $\hat{R}$ and $\hat{\sigma}$. It follows that the sum

$$\sum_{j=2}^{M} \mathbb{I}_{\hat{\sigma}^{-1}(j) \in \hat{R}} \mathbb{I}_{j \in \hat{S}}$$

counts the number of vertices that belong to the set $\hat{S}$ after making $\frac{C_p \log M}{2}$ draws, without replacement, from the set $\{2, ..., M\}$. This means that the sum follows the hypergeometric distribution $H(M, \alpha M, \frac{C_p \log M}{2})$.

Let us use some tail inequality from [7]. If Z is an hypergeometric random variable $H(N, M, n)$ and $t \geq 0$, we have:

$$\mathbb{P}[Z \leq \mathbb{E}[Z] - tn] \leq e^{-2t^2 n}$$

Here, by taking $t = \alpha - \frac{6}{C_p \log M}$ it follows:

$$\mathbb{P}\left[\sum_{j=2}^{M} \mathbb{I}_{\hat{\sigma}^{-1}(j) \in \hat{R}} \mathbb{I}_{j \in \hat{S}} \leq 3 \middle| A \cap B\right] \leq e^{-C_p \alpha^2 \log M + \frac{12\alpha}{C_p} - \frac{36}{C_p \log M}} \leq e^{\frac{12\alpha}{C_p}} M^{-C_p \alpha^2}$$

Thus, for $M \geq e^{\frac{12}{C_p^2 \alpha}} \frac{1}{\epsilon}^{\frac{1}{C_p \alpha^2}}$, we have:

$$\mathbb{P}\left[\sum_{j=2}^{M} \mathbb{I}_{\hat{\sigma}^{-1}(j) \in \hat{R}} \mathbb{I}_{j \in \hat{S}} \geq 3 \middle| A \cap B\right] \geq 1 - \epsilon$$

Now, using the fact that the events A and B are independent, we know that if C denotes the event " $\sum_{j=2}^{M} \mathbb{I}_{\hat{\sigma}^{-1}(j)\in\hat{R}}\mathbb{I}_{j\in\hat{S}} \geq 3$", then there is a constant $c > 0$ (this time depending on C_p and α) such that if $M_\epsilon = c \times \max\left((\frac{1}{\epsilon})^{\frac{8}{C_p}}, (\frac{1}{\epsilon})^{\frac{1}{C_p\alpha^2}}\right)$ then:

$$\mathbb{P}[A \cap B \cap C] \geq 1 - 3\epsilon$$

so 1 is the root of some double cycle of size at most αM_ϵ with probability greater than $1 - 3\epsilon$.

3.2 Great indexes are still not in double cycles

Since we do not have any tree part in our graph, the computation that has been made in 2.3.1 to upper-bound the expectation of the number of valid double cycles having I as a root becomes easier: the edges are now independent, and each edge is drawn from a binomial of parameter $C_p/\max(i,j)$. By doing the exact same computations, we get:

$$\mathbb{E}\left[C_{K,L}(I)\right] \leq \ 2C_p^{K+L}\frac{(K+L)!}{I^2}.$$

and by making the same reasoning as in 2.3.2 we have:

$$\mathbb{P}\left(\exists I \geq K, \ s.t. \ DC_{\alpha M_\epsilon}(I)\right) \ \leq \ 4C_p^{2\alpha M_\epsilon}\left(2\alpha M_\epsilon\right)!\frac{1}{K}.$$

Thus we can assure that with probability greater than $1 - \epsilon$ our method outputs at most:

$$K_\epsilon := 4\frac{1}{\epsilon}C_p^{2\alpha M_\epsilon}(2\alpha M_\epsilon)!$$

vertices.

3.3 Final output

The final result in this section is the same given in 2.4, but this time by taking $M_\epsilon = c\times\max\left((\frac{1}{\epsilon})^{\frac{8}{C_p}}, (\frac{1}{\epsilon})^{\frac{1}{C_p\alpha^2}}\right)$ for some constant $c > 0$ depending on C_p and α. This means that we output a set of at most $K(\epsilon) = 4\frac{1}{\epsilon}C_p^{2\alpha M_\epsilon}(2\alpha M_\epsilon)!$ vertices, which is of order $\left(1/\epsilon^{c'}\right)!$ for some $c' > 0$. Thus, by losing the tree information, our method can still find the root, but the size of the output becomes quite bigger.

4 Annex

4.1 Negative correlation between edges of the tree

Lemma 2 *Let us consider the uniform attachment model. Then, for any $j \in \mathbb{N}^*$ and a sequence of edges $e_1, \dots e_j$ such that $\forall i \in \{1, \dots, j\};\ e_i = (a_i \to b_i)$ with $a < b$. Then:*

If the b_i are disjoint, meaning $\#\{b_i\}_{i\in\{1,j\}} = j$, the following holds:

$$\mathbb{P}\left(e_1, \dots, e_j\right) \ = \ \prod_{i=1}^{j} \mathbb{P}(e_i).$$

If there happens to be two disjoint indexes $i \neq j$ such that $b_i = b_j$ then:

$$\mathbb{P}\left(e_1, \dots, e_j\right) \ = \ 0$$

Proof:

In the case of non-disjoint b_i the result is trivial because in a tree each vertex has a single parent.

When the b_i are disjoint, we will prove the result by induction on j. For $j = 1$ the result is trivially true. Then let us fix $j \leq 1$ and suppose the result true for this fixed j. Let us have $j + 1$ edges $e_i = (a_i, b_i)$ with disjoint b_i, we order them by the value of b_i, meaning $b_1 < \dots < b_{j+1}$. Then:

$$\mathbb{P}\left(e_1, \dots, e_{j+1}\right) \ = \ \mathbb{P}\left(e_1, \dots, e_j\right) \times \mathbb{P}\left(e_{j+1} \mid e_1, \dots e_j\right).$$

But according to the uniform attachment model, as $b_{j+1} > b_i$ for all $i \in \{1, \dots, j\}$ we have that e_{j+1} is independent from $e_1, \dots, e_j$. Then $\mathbb{P}\left(e_{j+1} \mid e_1, \dots e_j\right) = \mathbb{P}((e_{j+1})$, using the induction hypothesis we end up having:

$$\mathbb{P}\left(e_1, \dots, e_{j+1}\right) \ = \ \prod_{i=1}^{j+1} \mathbb{P}\left(e_i\right),$$

which proves the lemma.

4.2 Most likely cycle

Lemma 3 *Amongst the cycles of size K on the fixed set $i_1 < \cdots < i_K$, the most likely is the monotonic one:*

$$\{i_1 \to \cdots \to i_k \to i_1\}.$$

Proof:

Let us have a permutation $\sigma \in S_K$. we can write:

$$\mathbb{E}\left[1_{Cycle(i_{\sigma_1} \to \cdots \to i_{\sigma_K} \to i_{\sigma_1})}\right] = \prod_{j=1}^{K} \frac{1}{i_j^{E_{\sigma(j)}}}.$$

Where the E depend on σ, the comparative value of I to the i and are such that:

- $E_{\sigma(i)} \in \{0, 1, 2\}$, because each i is involved in 2 edges.
- In fact we have exactly: $E_{\sigma(j)} = \mathbb{I}(\sigma(j-1) \leq \sigma(j)) + \mathbb{I}(\sigma(j) \geq \sigma(j+1))$; where we obviously look at j modulo K in this writing.

As what matters is only the order in the loop we can fix $\sigma(1) = 1$ without losing any generality. Now let us consider the second smallest vertex, i_2. If $\sigma(2) \neq 2$ (nor K, id est 1 and 2 are not consecutive in the loop) let us change σ by σ' such that $\sigma'(2) = 2$ and the order of the other indexes is the same in σ and σ'. Let us denote l the index such that $\sigma(l) = 2$. For example we would go from cycle:

$$1 \to 3 \to 5 \to 2 \to 4$$

to the cycle:

$$1 \to 2 \to 3 \to 5 \to 4$$

and here l would be 4.

As the exponents E depend only on their neighbours the only exponents $E_{\sigma'(j)}$ that can differ from $E_{\sigma'(j)}$ are:

- $E_{\sigma'(2)} = E_{\sigma(l)} + 1$
- $E_{\sigma'(1)} = E_{\sigma(1)} = 0$ is actually not changing.
- $E_{\sigma'(3)} = E_{\sigma(2)}$ because the previous neighbour of $\sigma(2)$ is still smaller than $i_{\sigma(2)}$ (as i_2 is the second smallest) and the following neighbour is the same.

- If $\sigma(l-1) \leq \sigma(l+1)$ then $E_{\sigma'(l)} = E_{\sigma(j)} - 1$; $E_{\sigma'(l+1)} = E_{\sigma(l+1)}$
- If $\sigma(l-1) \geq \sigma(l+1)$ then $E_{\sigma'(l)} = E_{\sigma(j)}$; $E_{\sigma'(l+1)} = E_{\sigma(l+1)} - 1$

In any case, since we removed a power of 1 to a big index and added one to a small one, we have:

$$\mathbb{E}\left[1_{Cycle(i_{\sigma'_1} \to \cdots \to i_{\sigma'_K} \to i_{\sigma'_1})}\right] \;\geq\; \mathbb{E}\left[1_{Cycle(i_{\sigma_1} \to \cdots \to i_{\sigma_K} \to i_{\sigma_1})}\right].$$

Iterating this method from 3^{rd} smallest index to the K^{st} one proves the lemma.

4.3 Most likely double cycle

In this subsection we still consider a model where the presence of each edge is sampled from independent Bernoulli, of parameter $c/\max(i,j)$ for edge i,j.

Lemma 4 *Amongst the disjoint double cycles centered in I, of size K, L on the fixed set $i_1 < \cdots < i_{K+L-2}$, the most likely is the monotonic one, composed of the cycles:*

$$\{I \to i_1 \to \cdots \to i_{k-1} \to I\} \;\&\; \{I \to i_K \to \cdots \to i_{k+L-2} \to I\}.$$

Proof:

We showed that the most likely cycle was the monotonic one. So let us fix a partition of $\{1, \ldots K+L-2\}$ in a set $A = \{a_1 < \cdots < a_{K-1}$ and $B = \{b_1 < \cdots < b_{L-1}$. A will be the points in the first cycle, B the ones in the second. Then the most likely disjoint double cycle rooted in I in this configuration is:

$$\{I \to i_{a_1} \to \cdots \to i_{a_{K-1}} \to I\} \;\&\; \{I \to i_{b_1} \to \cdots \to i_{b_{L-1}} \to I\}.$$

Its probability is:

$$\mathbb{P}\left(\{I \to i_{a_1} i_{a_{K-1}} \to I\} \& \{I \to i_{b_1} i_{b_{L-1}} \to I\}\right) \;=\; \frac{1}{I^2} \times \prod_{j=1}^{K+L-2} \frac{1}{i_j} \times \frac{1}{\max(I, \{i_a\}_{a \in A})} \times \frac{1}{\max(I, \{i_b\}_{b \in B})}.$$

Thus finding the most likely double cycle is to maximize this expression in A and B. I.e. choosing A and B such that $\max\left(\{i_a\}_{a \in A}\right) \times \max\left(\{i_b\}_{b \in B}\right)$ is minimal. To do so one can take $A = \{1, \ldots, K-1\}$ and $B = \{K, \ldots, K+L-2\}$.

This proves the lemma.

Lemma 5 *Amongst the valid double cycles centered in I, of size K, L with p common edges on the fixed set $i_1 < \cdots < i_{K+L-2-p}$, the most likely is the monotonic one, with common edges being between the bigger vertices. More precisely:*

$$\{I \to i_1 \to \cdots \to i_{k-1-p} \to i_{K+L-2-2p} \cdots \to i_{k+L-2-p} \to I\} \,\&\, \{I \to i_{K-p} \to \cdots \to i_{k+L-2-p} \to I\}.$$

Proof:

We will use lemma 4. We just have to remark that the probability of this valid double cycle with p common vertices is the probability of the double cycle with no intersection where we just double the common vertices (to put one in each cycle, not commonly in both cycles), times 1 over the probability of the common path. From this observation the proof of lemma 5 is evident.

4.4 Diameter of uniform attachment model

Lemma 6 *For H_{k_ϵ} the height of a uniform attachment tree of size k_ϵ (see 2.2 for definition of k_ϵ), there exists a constant $\alpha > 0$ such that:*

$$\mathbb{P}\left(H_{k_\epsilon} \geq \alpha C_p \log(k_\epsilon)\right) \;\leq\; \epsilon.$$

Proof:

From theorem 6.32 in [5] we know that H_n has the following properties:

$$\mathbb{E}\left[H_n\right] \;\sim\; e\log(n);$$

and:

$$\mathbb{P}\left(|H_n - \mathbb{E}\left[H_n\right] \geq \mu\right) \;=\; O\left(e^{-c\mu}\right)$$

for some $c > 0$. If we write:

$$\mathbb{P}\left(H_n \geq \alpha C_p \log(n)\right) \;=\; \mathbb{P}\left(H_n - \mathbb{E}[H_n] \geq \alpha C_p \log(n) - \mathbb{E}[H_n]\right);$$

from the result above we can pick c' and c'' such that:

$$\mathbb{P}\left(H_n \geq \alpha C_p \log(n)\right) \;\leq\; c' \times \exp\left(-c''(\alpha C_p - e + o(1))\log(n)\right) \;\leq\; \frac{C'}{n^{c''(\alpha C_p - 2e)}}$$

plugging in the value of k_ϵ, there is a constant $c > 0$ which does not depend neither on C_p nor on ϵsuch that

$$\mathbb{P}\left(H_n \geq \alpha C_p \log(n)\right) \;\leq\; c \times \epsilon^{\frac{1}{C_p} c''(\alpha C_p - 2e)};$$

which is smaller than ϵ for a constant α big enough, independent from C_p and ϵ.

References

[1] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, 2013, x–481 p. URL: https://hal.archives-ouvertes.fr/hal-00794821.

[2] Sébastien Bubeck, Luc Devroye, and Gábor Lugosi. *Finding Adam in random growing trees*. 2015. arXiv: 1411.3317 [math.PR].

[3] Harry Crane and Min Xu. *Root and community inference on the latent growth process of a network using noisy attachment models*. 2021. arXiv: 2107.00153 [stat.ME].

[4] Luc Devroye and Jiang Lu. "The Strong Convergence of Maximal Degrees in Uniform Random Recursive Trees and Dags". In: *Random Struct. Algorithms* 7.1 (1995), pp. 1–14. DOI: 10.1002/rsa.3240070102. URL: https://doi.org/10.1002/rsa.3240070102.

[5] Michael Drmota. *Random Trees*. Springer-Verlag Wien, 2009, pp. XVII, 458. URL: https://www.springer.com/gp/book/9783211753552.

[6] Edgar M. Palmer. *Graphical Evolution: An Introduction to the Theory of Random Graphs*. John Wiley Sons, Inc., 1985, p. 46. ISBN: 0471815772.

[7] Matthew Skala. *Hypergeometric tail inequalities: ending the insanity*. 2013. arXiv: 1311.5939 [math.PR].