Shotgun Assembly of Erdos-Renyi Random Graphs

Martin Gjorgjevski

February 25, 2022

Abstract

The goal of this report is to introduce the topic "Shotgun assembly of random graphs". Given a collection of r-neighbourhoods of a graph, the goal is to reconstruct the original graph (either exactly or up to isomorphism). In earlier work it was shown by Mossel and Ross [MR15] that such a reconstruction is possible for Erdos-Renyi graphs with high probability when r=3 for p_n such that $\lim_{n\to\infty} \frac{np_n}{(log(n))^2} = \infty$. The question about reconstruction from 1-neighbourhoods and 2-neighborhoods remained open. In a recent work Mossel and Gaudio [GM22] answer these questions partially, leaving another set of open problems. We also consider other assembly problems such as the full binary tree and the random jigsaw puzzle.

Contents

1 Introduction

The deck of a graph G is the multiset of subgraphs $\{G - v | v \in V(G)\}$. An element of the deck is called a card. An old conjecture of Kelly abd Ulam [Kel57] states that if G and H are two graphs with isomorphic decks then G and H are isomorphic provided that they are graphs on at least 3 vertices. We will in particular consider the problem of reconstruction of random graphs, where typically results are of the form "with high probability G_n has certain property as $n \to \infty$". The reconstruction number [HP85] of a graph G is the minimal number k for which there exist k cards in the deck of G such that for any other graph H which in it's deck has k cards isomorphic to the previously specified cards of G then G and H are isomorphic. In this sense the conjecture of Kelly implies that the reconstruction number of any graph is finite. An important result of Bolobas [Bol90] states that the Erdos-Renyi Random Graph $G(n, p_n)$ with $c\log(n)/n < p_n < 1 - c\log(n)/n$ has reconstruction number 3 with high probability as $n \to \infty$.

We will focus on results about reconstruction of graphs from neighbourhoods. Instead of a deck, we assume we have acsess to the collection $\{N_r(v)|v \in G\}$ where $N_r(v)$ is the graph induced on vertices in G that are at distance at most r from v. We assume for the time being that the center of each such neighbourhood is labeled, all other vertices being unlabelled. The goal is to determine wheter the graph is uniquely reconstructible from the given neighbourhoods (either exactly or up to isomorphism) and in the case it is, to propose a method for efficient reconstruction. We say G is identifiable from its $r-$neighbourhoods if for all graphs H which have the same (potentially up to isomorphism) the same $r-$neighbourhoods as G, then G and H are isomorphic. The following lemma is known as the overlap method and it was used by Mossel and Ross [MR15] to show that for $\lim_{n\to\infty} \frac{np_n}{(log(n))^2} = \infty$, $G(n, p_n)$ is identifiable from 3-neighbourhoods.

Lemma If for any two vertices $u, v \in G$, $N_{r-1}(u)$ and $N_{r-1}(v)$ are not isomorphic, then G is identifiable from its $r-$neighbourhoods, and there is an efficient algorithm for the reconstruction.

Proof. We start with an arbitrary $r-$neighbourhood $N_r(v)$. Observe that $N_{r-1}(u)$ is contained in $N_r(v)$ for each neighbour u of v. Also, as we have acsess to $N_r(w)$, we also have acsess to $N_{r-1}(w)$ for all vertices w. Thus we start with a neighbour u of v and we can label it as there is only one neighbourhood $N_{r-1}(w)$ which will match with the observed neighbourhood in $N_r(v)$ (by assumption). In this way we can label all neighbours of v, and then we can proceed labeling vertices at distance 2 from v, and so on until the component of v is reconstructed. □

2 Assembly of Binary Trees

The shotgun assembly problem has been studied for various models of random graphs. As an introduction to the problems in graph assembly we consider the problem of assembling randomly colored trees.

Theorem [**MR15**] Let T_n be the full binary tree with 2^n leaves, each vertex is colored with one of q colors, independently and uniformly. We are given the 1-neighbourhoods of the $2^n - 2$ vertices that are not leaves and not the root. Let $\epsilon > 0$
If

$$\log(q)/n < \log(2) - \epsilon$$

then the probability of identifiability from 1-neighbourhods tends to zero. If

$$\log(q)/n > log(2) + \epsilon$$

then the probability of identifiability from 1-neighbourhoods tends to one.

Proof. For the first claim, it suffices to consider the edges connecting vertices between levels $n-2$ and $n-1$. Furthermore, after relabeling the 2^{n-1} vertices on level $n-1$, we only consider the ones with odd labels as this colection has neighbourhoods which are independent (as they are disjoint). We note that if among these vertices there is a pair of distinct vertices which have the same color and also their parents have the same color, but their children do not (i.e. the neighbourhoods are not the same) then we can switch these two neighbourhoods in the reconstruction and with good probability we obtain a distinct three (even up to isomorphism). Thus we consider the number of such pairs $B_{n,q} = B$ which can be represented as a sum of indicators. The idea is to show that $EB \to \infty$, as then we would have by the second moment method:

$$P(B = 0) \leq Var(B)/(EB)^2 \leq \frac{1}{EB}$$

(by the positioning of the vertices in the selection and the fact that the coloring is uniform, it is easy to see that if $X_{a,b}$ is the indicator that same coloring appears on edges a and b are independent). For $X_{a,b} = 1$, the probability that the color of the central vertex of edge a and of edge b are the same is $1/q$, and the probability that the parents are of the same color is also $1/q$, and these events are independent by assumption. Also, the probability that the remaining two vertices in each neighbourhood are equal as sets is at most $2/q$ so that $EX_{a.b} \geq \frac{1}{q^2}(1 - \frac{2}{q})$. As there are $2^{2(n-2)}$ summands that make up B, we see that EB is asymptotically equivalent to $2^{2(n-2)}/q^2$, which tends to infinity under the first assumption.

For the second claim, we observe that if all edges are uniquely colored (in the sense that no two edges have vertices whose colors agree as sets) then identifiability is trivial. There are $2^{n+1} - 2 \leq 2^{n+1}$ edges in the binary tree, and the probability that two edges have the same color is $\leq 2/q^2$, so that the

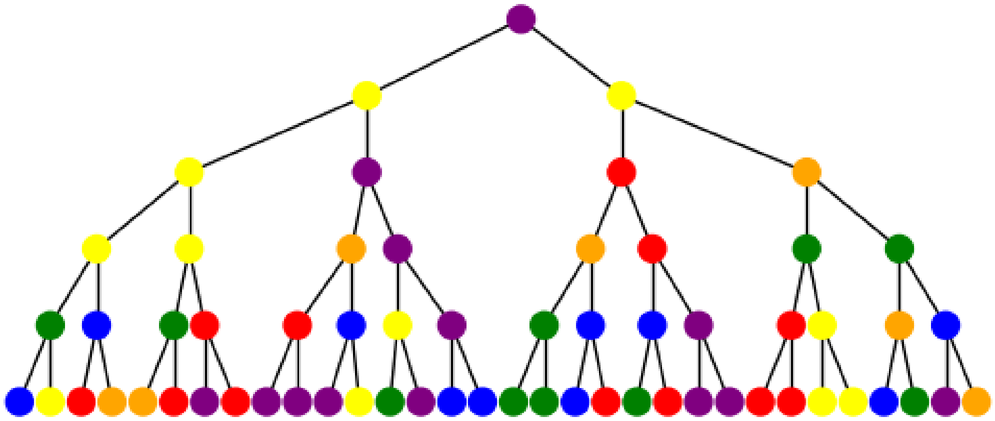

Figure 1: Randomly colored tree with 32 leaves and 6 colors

expected number of pairs of edges that are of the same color is $\leq 2^{2n+3}/q^2$ which under the second assumption tends to 0, and hence by the first moment method we conlude that with probabilty tending to 1, all edges are uniquely colored. □

3 Negative results

Theorem[MR15] Let G be an Erdos-Renyi graph on N vertices with $p_N = \lambda/n$. If $\sqrt{N}\lambda^r(1-\lambda/N)^{Nr} \to \infty$ then the probability of identifiability from r-neighbourhoods tends to 0.

Proof. The main idea is to consider certain configurations which prevent identifiability. Showing that they apear with large probability implies that identifiability is not possible. In particular we consider a graph which has two connected components, one of them being a path graph on $2r+1$ vertices and the other is a graph on $2r+5$ vertices (standard path graph on $2r+1$ vertices with additional 2 vertices at the endpoints of the path graph, also known as prongs- see Figure 2 below).

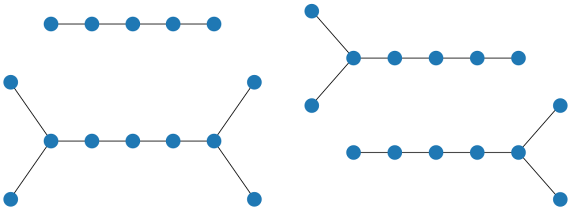

Figure 2: The occurence of this graph implies non-unique reconstruction

If such a configuration appears, it is clear that identifiability is impossible as by switching one of the prongs to the path graph, we obtain a non-isomorphic graph. By looking only at r-neighbourhoods it is not possible to tell which of the two configurations is from the original graph. An important thing to note is that when we assume that such a configuration does appear, we assume something about all potential edges connecting the configuration to the rest of the graph as well as the potential edges among the different vertices of the configuration itself. For α a collection of $2r+1+2r+5 = 4r+6$ points, we let X_α be the event that the configuration appears on the vertices labeled by α, and we let $B = \sum X_\alpha$, the sum being taken over all possible choices of α. Then $EB = \binom{N}{4r+6}EX_\alpha$. Next, we observe that if $\alpha \cap \beta \neq \emptyset$, then $X_\alpha X_\beta = 0$ (this is due to the remark that $X_\alpha = 1$ forces precisely which edges will appear among vertices in α, and thus it forces $X_\beta = 0$). Thus

$$Var(B) = (EB)(1 + \binom{N-4r-6}{4r+6}E(X_\beta|X_\alpha = 1))$$

due to the symmetry of the Erdos-Renyi graph. At last, we compute

$$EX_\alpha = \binom{4r+6}{2r+1}\binom{4r+5}{4}\binom{4}{2}2((2r+1)!/2)^2 p_N^{2r+2}(1-p_N)^{(4r+6)(N-3)+4}$$

where the first binomial coefficients stands for the choice of the path graph, the second stands for the choice of the prongs, and the next factor corresponds to permuting the $2r+1$ vertices from the $2r+1$ path graphs in the components. Similarly, one computes

$$E(X_\beta|X_{\alpha=1}) = \binom{4r+6}{2r+1}\binom{2r+5}{4}\binom{4}{2}2((2r+1)!/2)^2 p_N^{2(2r+2)}(1-p_N)^{(4r+6)(N-2r+3)+4}$$

Keeping in mind that $p_N = \lambda/N$ as well as the assumption of the theorem, it is easy to see that $EB^2 = o((EB)^2)$ so that by the second moment method $B > 0$ with high probability. □

4 1-Neighborhood reconstruction

For a graph G and an edge incident to the vertices u, v, we denote by $H_{u,v}$ the subgraph of G induced on the common neighbours of u and v. The method for proving that 1-neighbourhood reconstruction is possible relies on the following observation: if $H_{u,v}$ is unique up to isomorphims on graphs for each pair of vertices u, v, then 1-neighbourhood reconstruction is possible. Indeed, to determine if u and v are neighbours, we simply look at all possible neighbourhoods H_{u,u_0} and H_{v,v_0}. If two such neighbourhoods match (are isomorphic) then u and v are neighbours. There will be no ambiguity since if H_{v,v_0} matches with H_{u,u_0} then it follows that $v_0 = u$ and $u_0 = v$. Hence for each pair we u, v we can determine if there is an edge between u and v or not. However in practice it is difficult to determine if $H_{u,v}$ is unique up to isomorphism. The idea in proving that Erdos-Renyi Graphs are reconstructible from 1-neighbourhoods is to show that $H_{u,v}$ is unique with high probability, for each pair of vertices u, v. More specifically we have the following theorem.

Theorem [GM22] Let $p_n = n^{-\alpha}$ for $0 < \alpha < \frac{1}{3}$. Then $G(n, p_n)$ is exactly reconstructible from its $1-$neighbourhoods.

Proof. For vertices a,b, let $W_{a,b}$ denote the number of shared neighbours of a and b. $W_{a,b}$ has distribution $bin(n-2, p_n^2)$. Given four vertices u, v, x, y let Y be the number of common neighbours of u, v, x, y and let G_1 be the graph induced on the neighbours of u, v that are not the neighbours of both x and y. Also, we let Z to be the number of vertices from $\{x, y\}$ that are connected to both u and v. Let G_2 be the graph on the common vertices of x and y. If $H_{x,y}$ and $H_{u,v}$ are isomorphic, then G_1 must be isomorphic to a subgraph of G_2. Hence, we focus on bounding the probability of the event $G_1 \subseteq G_2$.

$$\begin{aligned}&P(H_{x,y} = H_{u,v}|W_{u,v} = W_{x,y} = \lambda + Z, Y = \mu, EG_1 = k) \leq \\ &P(G_1 \subseteq G_2|W_{u,v} = W_{x,y} = \lambda + Z, Y = \mu, EG_1 = k) \leq \\ &\binom{\lambda+2}{\lambda-\mu}(\lambda-\mu)! p_n^{k-2\mu-1} \leq \exp\left(\log(n)\right)(2\alpha+1)\lambda + \alpha - k\alpha)\end{aligned}$$

To see this note that we can further condition on the vertices of G_2, and then observe that at most $2\mu + 1$ edges have already been revealed in G_2 so that at most $k - 2\mu - 1$ must appear in G_2. The combinatorial term counts the number of ways this can happen.
Next we note that EG_1 is binomially distributed given $W_{u,v} - Y - Z$, so that concentration inequalities imply that the probability of taking values far away from the mean is exponentially small. Let $c > 0$, to be determined. We have

$$P(W_{u,v} \geq n^c(n-2)p_n^2) \leq \exp(-\Theta(n^{1+c-2\alpha}))$$

which is a high probability bound for $c > 2\alpha - 1$, in particular it is $o(1/n^4)$. Similarly, a concentration inequality gives that

$$P(W_{u,v} - Y - Z \leq \frac{1}{2}np_n^2) = o(\frac{1}{n^4})$$

From here we also have concentration for $E(G_1)$:

$$\begin{aligned}
&P(E(G_1) \leq (1-\epsilon)\binom{\frac{1}{2}np_n^2}{2}) \\
&\leq P(E(G_1) \leq (1-\epsilon)\binom{\frac{1}{2}np_n^2}{2}|W_{u,v} - Y - Z \geq \frac{1}{2}np_n^2)P(W_{u,v} - Y - Z \geq \frac{1}{2}np_n^2) \\
&+ P(W_{u,v} - Y - Z \leq \frac{1}{2}np_n^2)) \\
&\leq \exp(-\frac{\epsilon^2}{2}p_n\binom{\frac{1}{2}np_n^2}{2}) + o(\frac{1}{n^4}) \\
&\leq \exp(-\Theta(n^{2-5\alpha}) + o(\frac{1}{n^4})
\end{aligned}$$

Taking all of these observations into account, we finally get

$$P(H_{x,y} = H_{u,v}) \leq \exp(\Theta(n^{1+c-2\alpha} - \Theta(n^{2-5\alpha})) + o(\frac{1}{n^4})$$

This is a high probability bound when $1 + c - 2\alpha < 2 - 5\alpha$, that is when $c < 1 - 3\alpha$. For $0 < \alpha < \frac{1}{3}$, choosing $\max(0, 2\alpha - 1) < c < 1 - 3\alpha$ gives that $P(H_{x,y} = H_{u,v}) = o(\frac{1}{n^4})$ and we conclude by an union bound. □

In the paper of Gaudio and Mossel [GM22] the following result is also established

Theorem Let $p_n = n^{-\alpha}$. Then for $\frac{1}{2} < \alpha < 1$ reconstruction of $G(n, p_n)$ from 1-neighbourhoods is impossible with high probability.

5 2-Neighborhood reconstruction

Theorem [GM22] For $\alpha \in (0, \frac{1}{2}) \cup (\frac{1}{2}, \frac{3}{5})$ and $p_n = n^{-\alpha}$, $G(n, p_n)$ is exactly reconstructible from 2-neighbourhoods.

Proof. We only show this result for $\alpha \in (0, \frac{1}{2})$. The result is a consequence of the following results:

1 If $p_n = c\frac{\sqrt{\log(n)}}{n}$ with $c > \sqrt{2}$ then the diameter of $G(n, p_n)$ is 2 with high probability

2 If $p \leq \frac{1}{2}$ and $p = \omega(\frac{\log^4(n)}{n \log\log(n)}$ then with high probabilty any two distinct vertices u and v have different degree neighbourhoods. Moreover, with high probability one can produce a canonical labelling of G by sorting the vertices in lexicographic order by their degree neighbourhoods.

Using these 2 results, in the setting $\alpha \in (0, \frac{1}{2})$, we note that we observe the graph from any 2-neighbourhood (by 1). Then we produce a canonical labeling of each such neighbourhood (possible by 2). Finally, we sort the vertices lexicographically by their degree neighbourhoods to determine the canonical label of each vertex.

□

Negative results about reconstrucion with 2-neighbourhoods are also partially known. We have the following theorem

Theorem [GM22] Let $\alpha \in (\frac{3}{4}, 1)$, $p_n = n^{-\alpha}$. With high probability $G(n, p_n)$ is not reconstructible from 2−neighbourhoods.

We conclude this section with another negative result.

Theorem [GM22] Let $\alpha \in (\frac{2}{3}, 1)$, $p_n = n^{-\alpha}$. With high probability, $G(n, p_n)$ cannot be reconstructed from its 2-neighbourhoods using the overlap method.

Proof. The idea is to show that with large probability there are two isomorphic star graphs among the 1-neighbourhoods of G, so the neighbourhoods $\{N_1(v)|v \in V(G)\}$ are not unique. To this end, let $\beta = \frac{\frac{2}{3}+\alpha}{2}$. For arbitrary vertex v, by Markov's inequality we have

$$P(deg(v) > n^{1-\beta}) \leq n^{\beta-\alpha} = o(1)$$

Let S_v be the indicator that the neighbourhood $N_1(v)$ is a star graph of degree at most $n^{1-\beta}$. Then we have

$$P(S_v = 1) \geq P(S_v = 1|deg(v)) \leq n^{1-\beta})(1 - o(1))$$

Next, we observe that $P(S_v = 1|deg(v) = j) \geq P(S_v = 1|deg(v) = n^{1-\beta})$ simply by observing that knowing the degree of v also means (by symmetry) that we may assume we know the neighbours of v. So we conclude that

$$P(S_v = 1) \geq 1 - P(Z \geq 1) - o(1) \geq 1 - EZ - o(1)$$

where Z has is binomially distributed with parameters $bin(\binom{n^{1-\beta}}{2}, p_n)$ as the event $S_v = 1$ given $degv = n^{1-\beta}$ implies that there is no edge between the remainig $\binom{n^{1-\beta}}{2}$ vertices that are neighbours of v. $EZ = o(1)$ in this case, so finally we get that

$$P(S_v = 1) = 1 - o(1)$$

Finally, $P(\sum_{v\in V} S_v < \frac{3n}{4}) = P(\sum_{v\in V}(1 - S_v) > \frac{n}{4}) = o(1)$ by Markov's inequality. Hence, with probability $1 - o(1)$ at least $\frac{3n}{4}$ neighbours are stars with no more than $n^{1-\beta}$ neighbours, and by the pigeonhole principle this means that there is a k_n such that there are at least

$$\frac{3n}{4(n^{1-\beta} + 1)} \geq \frac{n}{2n^{1-\beta}}$$

stars of degree k_n. □

6 Finding the centers of the neighbourhoods

Lemma [GM22] Suppose $p_n = n^{-\alpha}$ for $\alpha \in (0, 1)$. Then with high probability $v \in V$ is the only vertex connected to every other vertex in $N_1(v)$.

Proof. As the degree of v is distributed as a binomial variable $bin(n-1, p_n)$, by standard concentration inequalities, we can restrict our attention to the case where $deg(v) = m \geq (1 - \epsilon)p_n(n - 1)$. In that case the probability that a given w of v is conneceted to all other neighbours of v is p_n^{m-1}. The claim follows by an union bound. □

Identification of the center of a given vertex from $2-$neighbourhoods is also possible for $\alpha \in (\frac{1}{2}$. To find the center, we first remove the vertices in the neighbourhood with degree $\leq \frac{1}{2}n^{1-\alpha}$. Then with highprobability the center is the vertex with the highest degree in the remaining graph.

7 Random Jigsaw puzzles

Consider an n by n grid. We color the edges that connect two orthogonally adjacent cells by one of q colors uniformly and independenlty. To represent this construction as a graph, we consider the centers of the cells as vertices and place edges between any two vertices that belong in adjacent cells with the same color as the edge separating these two cells. This is done even for the cells at the boundary where extra vertices are created so that each vertex has degree 4 and the edges are colored with one of q colors. We observe the colored 1-neighbourhood of each vertex. What is the relationship between n and q so that with high probability there is identifiability?

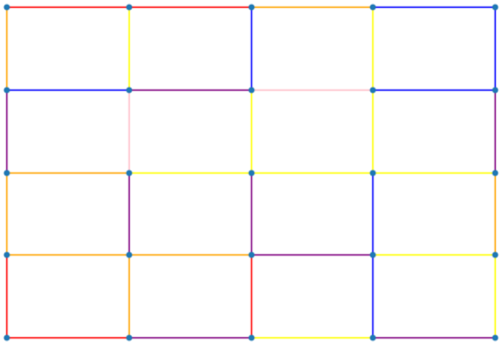

Figure 3: Random Jigsaw puzzle with 40 edges and 6 colors

Theorem [MR15] If $q = o(n^{2/3})$, then the probability of identifiability goes to 0 as $n \to \infty$.

Proof. A pair of pieces is called aligned if the position of their centers are of the form $(j, 2i)$ and $(j, 2i+1)$. We consider the event $X_{i,j,k,l}$ to be the event that the pieces $(j, 2i)$ and $(k, 2l)$ have the same color except at the edge above which is different and $(j, 2i+1)$ and $(k, 2l+1)$ also have the same color except for the edges below these verticies (which are neccesairily different). In the event $X_{i,j,k,l} = 1$, we can easily switch these edges that do not match and get a different solution to the jigsaw puzzle. Again, the idea is to use the second moment method to show that the number of such events is ¿0 with high probability. To this end, let $Y = \sum X_{i,j,k,l}$ where the sum is taken over all distinct pairs $(i, j) \neq (k, l)$. It is easy to see that these events are pairwise independent (this is obvious when the four considered pieces are far apart and follows by the fact that the coloring is uniform in other cases). Hence,

$$Var(Y) = \sum Var(X_{i,j,k,l}) \leq \sum EX_{i,j,k,l} = EY$$

Also it is easy to see that $EX_{i,j,k,l} = q^{-6}(1 - \frac{1}{q})$ so that EY behaves like $nq^{-6} \to \infty$ when $q = o(n^{2/3})$, concluding the proof. □

It is easy to get an upper bound on the number of colors needed to ensure identifiability with high probability. Intuitively, the more colors are allowed, the easier it is to get an identifiable graph. A naive idea which provides one such upper bound is to pick q in such a way that with high probability all edges are colored differently. Then obviously reconstrucion is possible by overlaping edges with the same color. A simple calculation shows that if $q >> n^4$ then this is satisfied. A stronger result (which uses percolation theory) can be found in Mossel[2] which states that:

Theorem [MR15] If $q = \omega(n^2)$ then assembly of the random jigsaw puzzle is possible with probability tending to 1. Moreover, there is an efficient algorithm for this assembly that works with high probability.

8 Open Problems

The question about reconstruction from 1−neighbourhoods and 2−neighbourhoods has been partially answered in the paper of Gaudio and Mossel [GM22] . The following questions remain open:

1 Is reconstruction from 2−neighbourhoods possible for $\alpha = \frac{1}{2}$ or $\alpha \in [\frac{3}{4}, 1]$?

2 Is there an efficient algorithm for reconstruction from 2−neighbourhoods besides the case $\alpha \in (0, \frac{1}{2})$?

References

[AS16] Noga Alon and Joel H. Spencer. *The Probabilistic Method*. 4th. Wiley Publishing, 2016. ISBN: 1119061954.

[Bol90] Béla Bollobás. “Almost every graph has reconstruction number three”. In: *J. Graph Theory* 14 (1990), pp. 1–4.

[GM22] Julia Gaudio and Elchanan Mossel. *Shotgun Assembly of Erdos-Renyi Random Graphs*. 2022. arXiv: 2010.14661 [math.PR].

[HP85] Frank Harary and Michael Plantholt. “The graph reconstruction number”. In: *Journal of Graph Theory* 9.4 (1985), pp. 451–454. DOI: https://doi.org/10.1002/jgt.3190090403. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jgt.3190090403. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jgt.3190090403.

[Kel57] Paul J. Kelly. “A congruence theorem for trees.” In: *Pacific Journal of Mathematics* 7.1 (1957), pp. 961–968. DOI: pjm/1103043674. URL: https://doi.org/.

[MR15] Elchanan Mossel and Nathan Ross. *Shotgun assembly of labeled graphs*. 2015. arXiv: 1504.07682 [math.PR].