Analysis of a Canonical Labeling Algorithm for the Alignment of Correlated Erdős-Rényi Graphs

Osman Emre Dai

Department of Industrial and Systems Engineering,
Georgia Institute of Technology
oedai@gatech.edu

Negar Kiyavash

Department of Electrical & Computer Engineering and
Department of Industrial & Systems Engineering, Georgia
Institute of Technology
negar.kiyavash@ece.gatech.edu

Daniel Cullina

Department of Electrical Engineering, Princeton
University
dcullina@princeton.edu

Matthias Grossglauser

School of Computer and Communication Sciences, École
Polytechnique Fédérale de Lausanne
matthias.grossglauser@epfl.ch

ABSTRACT

Graph alignment in two correlated random graphs refers to the task of identifying the correspondence between vertex sets of the graphs. Recent results have characterized the exact information-theoretic threshold for graph alignment in correlated Erdős-Rényi graphs. However, very little is known about the existence of efficient algorithms to achieve graph alignment without seeds.

In this work we identify a region in which a straightforward $O(n^{11/5} \log n)$ -time canonical labeling algorithm, initially introduced in the context of graph isomorphism, succeeds in aligning correlated Erdős-Rényi graphs. The algorithm has two steps. In the first step, all vertices are labeled by their degrees and a trivial minimum distance alignment (i.e., sorting vertices according to their degrees) matches a fixed number of highest degree vertices in the two graphs. Having identified this subset of vertices, the remaining vertices are matched using an alignment algorithm for bipartite graphs.

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic inference problems**; • **Security and privacy** → **Data anonymization and sanitization**; **Social network security and privacy**;

KEYWORDS

Network alignment; de-anonymization

ACM Reference Format:

Osman Emre Dai, Daniel Cullina, Negar Kiyavash, and Matthias Grossglauser. 2019. Analysis of a Canonical Labeling Algorithm for the Alignment of Correlated Erdős-Rényi Graphs. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts)*, June 24–28, 2019, Phoenix, AZ, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3309697.3331505>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6678-6/19/06.

<https://doi.org/10.1145/3309697.3331505>

1 INTRODUCTION

Graph alignment refers to a class of computational techniques to identify node correspondences across related networks based on structural information. These techniques have applications in various domains where networks (e.g. social networks) exhibit a natural correlation due to underlying relations (e.g. real life relationships between individuals). The exact information-theoretic threshold for the alignment of correlated Erdős-Rényi graphs was recently characterized in [1]. In this paper we present a $O(n^{11/5} \log n)$ -time algorithm that requires no prior information about the alignment and characterize the regime where it reliably achieves perfect alignment for the Erdős-Rényi random graph model.

2 ALIGNMENT BY CANONICAL LABELING

Canonical labeling refers to the assignment of labels to a vertex set such that these labels are invariant to permutations of the vertex set and only depend on the graph's structure. Such labeling schemes allow the identification of the correspondence between the vertex sets of two isomorphic graphs. If the canonical labeling scheme is robust in the sense that small differences in the structure of the graph induce small perturbations on the labels of vertices, then such a scheme can also be used to align non-isomorphic yet sufficiently correlated graphs.

We consider a robust labeling scheme with two steps: In the first step vertices are labeled with their degrees and the small subset of the vertices with high-degrees are identified. These are referred to as 'anchors' and form a basis for the alignment of the rest of the graph. In the second step, the remaining vertices are labeled with signature vectors based on their adjacencies with the anchors identified in step one.

This second step ignores all edges between unidentified vertices, effectively treating the graph as a bipartite graph. Therefore, the second step may be considered separately as an algorithm to align two bipartite graphs with one unidentified partite set. We refer to the first step as the *anchor alignment algorithm* and the second step as the *bipartite alignment algorithm*.

Definition 2.1. The high-degree sorting function f_h takes as input a graph G on the vertex set V and lists the h highest-degree vertices, sorted by degree. We represent $f_h(G)$ is some vector $\mathbf{w} = (w_1, w_2, \dots, w_h) \in V^h$ as a vertex of distinct vectors.

Anchor alignment corresponds to the index-by-index alignment of the lists of vertices sorted by their degrees. After aligning the h highest-degree vertices from the two sets and forming the list of anchors, the algorithm labels the remaining vertices by binary vectors encoding their adjacency to the anchor vertices.

Definition 2.2. Given graph G and anchor vector $f_h(G) = \mathbf{w} = (w_1, w_2, \dots, w_h)$, the signature function sig_G takes as input vertex $u \in V(G)$ and returns the signature label of the vertex such that,

$$\text{sig}_G(u) \in \{0, 1\}^h \quad \text{and} \quad \text{sig}_G(u)_i = \mathbb{1}\{\{u, w_i\} \in E(G)\}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function of an event.

The bipartite alignment algorithm aligns vertices such as to minimize the Hamming distance between pairs of signatures of aligned vertices. In our analysis we consider a naive approach, aligning each vertex in one graph to the vertex with the closest signature in the other graph. The steps of the alignment algorithm are summarized in Algorithm 1.

ALGORITHM 1 AnchorSignAlign

Input: $G_a = (V_a; E_a)$, $G_b = (V_b; E_b)$, h

Output: Estimated alignment $\hat{M} : V_b \rightarrow V_a$

1: **Step 1: Anchor alignment**

2: $\mathbf{w}_a = f_h(G_a)$

3: $\mathbf{w}_b = f_h(G_b)$

4: **for** $i \in [h]$ **do**

5: $\hat{M}(\mathbf{w}_{b,i}) = \mathbf{w}_{a,i}$

6: **end for**

7: **Step 2: Bipartite alignment**

8: $H_a = \{\mathbf{w}_{a,i} : i \in [h]\}$

9: $H_b = \{\mathbf{w}_{b,i} : i \in [h]\}$

10: **for** vertex $v \in V_b \setminus H_b$ **do**

11: $\hat{M}(v) = \arg\min_{u \in V_a \setminus H_a} |\text{sig}_a(u) - \text{sig}_b(v)|$

12: **end for**

3 CORRELATED ERDŐS-RÉNYI GRAPHS

We perform our analysis on correlated Erdős-Rényi (ER) graphs [1]. Under the basic ER model of random graphs, $G \sim ER(n; p)$ is a random graph on n vertices where any two vertices share an edge with probability p independent from the rest of the graph. Under the correlated graph model, $(G_a, G_b) \sim ER(n; (p_{11}, p_{10}, p_{01}, p_{00}))$ are a pair of graphs on the same set of n vertices where the occurrences of an edge $e = \{u, v\}$ between any pair of vertices u, v is independent and identically distributed with the following probabilities: $\Pr[e \in E(G_a), e \in E(G_b)] = p_{11}$, $\Pr[e \in E(G_a), e \notin E(G_b)] = p_{10}$, etc. We denote the vector of probabilities as $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$. Note that all probabilities are functions of n . We limit our interest to sparse graphs and only consider $\mathbf{p}$ such that $\lim_{n \rightarrow \infty} p_{00} = 1$.

4 OUTLINE AND INTUITION FOR THE ANALYSIS

The two steps of Algorithm 1 dictate opposing bounds on the value of the parameter h . The bipartite alignment phase requires distinct signatures, which is guaranteed only if the length of the signature vectors (h) is large enough. However, the performance of the anchor alignment phase degrades as h grows larger. Our analysis consists

of determining upper and lower bounds on h and identifying the region for h which satisfies both bounds.

First we present a sufficient condition to perfectly align the h highest-degree vertices in correlated ER graphs. This gives an upper bound on h . The result is derived by determining the conditions that guarantee, with high probability, that the h highest-degree vertices have large enough degree separation. It is then unlikely that any two high-degree vertices have their degree order reversed.

In order to understand the constraints on the bipartite matching phase, we then analyze the closely related problem of correlated random bipartite graphs. We try to match one of the partite sets based on the complete knowledge of the matching of the other partite set. This allows us to derive a lower bound on the size h of the identified set. This problem closely relates to the bipartite alignment phase of Algorithm 1; in both cases we assume to have complete knowledge of the alignment of one partite set (i.e. the set of anchor vertices) and try to match the other side by only considering edges that connect these two sets. But in the general case the edge random variables between the high-degree set and the remaining vertices are not independent. Fortunately the dependence is weak and it is possible to handle this issue by requiring the anchor set to be robust to the addition or removal of a pair of vertices.

5 RESULT

Our main result is a condition under which Algorithm 1 successfully recovers the true graph alignment.

THEOREM 5.1. Let $G_a = (V; E_a)$ and $G_b = (V; E_b)$ such that $(G_a, G_b) \sim ER(n, \mathbf{p})$ where $\mathbf{p}$ is a function of n with $p_{11} \leq o(1)$,

$$p_{11} \geq \omega\left(\frac{\log^{7/5} n}{n^{1/5}}\right) \quad \text{and} \quad p_{01} + p_{10} \leq o\left(\frac{p_{11}^5}{\log^6 n}\right),$$

Then Algorithm 1 with parameter h such that

$$\frac{\log n + \omega(1)}{p_{11}} \leq h \leq O\left(\frac{\log n}{p_{11}}\right)$$

exactly recovers the alignment between the vertex sets of G_a and G_b with probability $1 - o(1)$.

Figure 1: Comparison of regions of achievability for symmetric noise ($p_{10} = \Theta(p_{01})$): (A) region achievable by Algorithm 1, (A $\cup$ B) theoretical achievability region [1].

REFERENCES

- [1] D. Cullina and N. Kiyavash, “Improved achievability and converse bounds for erdos-erenyi graph matching,” in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science - SIGMETRICS 2016*, ACM Press, 2016.