

Predicting Signed Edges with $O(n^{1+o(1)} \log n)$ Queries

Michael Mitzenmacher*

Charalampos E. Tsourakakis†

Abstract

Social networks and interactions in social media involve both positive and negative relationships. Signed graphs capture both types of relationships: positive edges correspond to pairs of “friends”, and negative edges to pairs of “foes”. The *edge sign prediction problem*, which aims to predict whether an interaction between a pair of nodes will be positive or negative, is an important graph mining task for which many heuristics have recently been proposed [LHK10a; LHK10b].

Motivated by social balance theory, we model the edge sign prediction problem as a noisy correlation clustering problem with two clusters. We are allowed to query each pair of nodes whether they belong to the same cluster or not, but the answer to the query is corrupted with some probability $0 < q < \frac{1}{2}$. Let $c = \frac{1}{2} - q$ be the gap. We provide an algorithm that recovers the clustering with high probability in the presence of noise for any constant gap c with $O(n^{1+\frac{1}{\log \log n}} \log n)$ queries. Our algorithm uses simple breadth first search as its main algorithmic primitive. Finally, we provide a novel generalization to $k \geq 3$ clusters and prove that our techniques can recover the clustering if the gap is constant in this generalized setting.

1 Introduction

With the rise of social media, where both positive and negative interactions take place, signed graphs, whose study was initiated by Heider, Cartwright, and Harary [CH56; Hei46; Har53], have become prevalent in graph mining. A key graph mining problem is the *edge sign prediction problem*, which aims to predict whether an interaction between a pair of nodes will be positive or negative [LHK10a; LHK10b]. Recent works have developed numerous heuristics for this task that perform relatively well in practice [LHK10a; LHK10b].

In this work we propose a theoretical model for the edge sign prediction problem that highlights its intimate connections with the famous planted partition problem [ABH16; CK01; HWX16; McS01]. Specifically, we model the edge sign prediction problem as a noisy correlation clustering problem, where we are able to query a pair of nodes (u, v) to test whether they belong to the same cluster (edge sign $f(u, v) = +1$) or not (edge sign $f(u, v) = -1$). The query fails to return the correct answer with some probability $0 < q < \frac{1}{2}$. Correlation clustering is a basic data mining primitive with a large number of applications ranging from social network analysis [Har53; LHK10a] to computational biology [HEP+16]. Our theoretical model is inspired by the famous *balance theory*: “the friend of my enemy is my friend” [CH56; EK10; Hei46]. The details of our model follow.

Model I. Let $V = [n]$ be the set of n items that belong to two clusters, call them red and blue. Set $f : V \rightarrow \{\text{red}, \text{blue}\}$, $R = \{v \in V(G) : f(v) = \text{red}\}$ and $B = \{v \in V(G) : f(v) = \text{blue}\}$, where

*Harvard University, michaelm@eecs.harvard.edu

†Harvard University, babis@seas.harvard.edu

$0 \leq |R| \leq n$. The function f is unknown and we wish to recover the two clusters R, B by querying pairs of items. (We need not recover the labels, just the clusters.) For each query we receive the correct answer with probability $1 - q$, where $q > 0$ is the corruption probability. That is, for a pair of items u, v such that $f(u) = f(v)$, with probability q it is reported that $\tilde{f}(u) \neq \tilde{f}(v)$, and similarly if $f(u) \neq f(v)$ with probability q it is reported that $\tilde{f}(u) = \tilde{f}(v)$. Our goal is to perform as few queries as possible while recovering the underlying cluster structure.

Main result. Our main theoretical result is that we can recover the clusters (R, B) with high probability¹ in polynomial time. Our algorithm uses breadth first search (BFS) as its main algorithmic primitive. Our result is stated as Theorem 1.

Theorem 1. *There exists a polynomial time algorithm that performs $\Theta(n \log n (2c)^{-\frac{\log n}{\log \log n}})$ edge queries and recovers the clustering (R, B) whp for any gap $0 < c = \frac{1}{2} - q < \frac{1}{2}$.*

When c is constant, then the number of queries is $O(n^{1+\frac{1}{\log \log n}} \log n)$. A natural follow-up question that we address here is whether our results generalize to the case of more than two clusters. We provide a general model and show that our techniques recover the cluster structure whp as long as the gap c is constant.

Model II. Our model is now that there are k groups, that we number $\{0, 1, \dots, k - 1\}$ and that we think of as being arranged modulo k . Let $g(u)$ refer to the group number associated with a vertex u . We start by noting that if when querying an edge we returned only whether the the groups of the two edges were equal, it would be difficult to reconstruct the clusters; indeed, even with no errors, a chain of such responses along a path would not generally allow us to determine whether the endpoints of a path were in the same group or not. A model that provides more information and naturally generalizes the two cluster case is the following: when we query an edge $e = (x, y)$, we obtain

$$\tilde{f}(e) = \begin{cases} g(x) - g(y) \bmod k, & \text{with probability } 1 - q; \\ g(x) - g(y) + 1 \bmod k, & \text{with probability } q/2; \\ g(x) - g(y) - 1 \bmod k, & \text{with probability } q/2. \end{cases} \quad (1)$$

That is, we obtain the difference between the groups when no error occurs, and with probability q we obtain an error that adds or subtracts one to this gap with equal probability. When $q = 0$, so there are no errors from $\tilde{f}(e)$, the edge queries would allow us to determine the difference between the group numbers of vertices at the start and end of any path, and in particular would allow us to determine if the groups were the same. We also note that we choose this description for ease of exposition. More generally we could handle queries governed by more general error models, of the form:

$$\tilde{f}(e) = g(x) - g(y) + i \quad \text{with probability } q_i, 0 \leq i < k.$$

That is, the error does not depend on the group values x and y , but is simply independent and identically distributed over the values 0 to $k - 1$.

Theorem 2. *There exists a polynomial time algorithm that performs $O(n^{1+\frac{1}{\log \log n}} \log n)$ edge queries and recovers the k clusters under the model of equation (1) whp for any constant gap $0 < c < \frac{1}{2}$.*

¹An event A_n holds with high probability (whp) if $\lim_{n \rightarrow +\infty} \Pr[A_n] = 1$.

Our proof techniques extend naturally to this model.

Roadmap. Section 2 presents some theoretical preliminaries. Section 3 presents our algorithmic contributions. Section 4 briefly reviews related work. Finally, Section 5 concludes the paper.

2 Theoretical Preliminaries

We use the following powerful probabilistic results for the proofs in Section 3.

Theorem 3 (Chernoff bound, Theorem 2.1 [JLR11]). *Let $X \sim \text{Bin}(n, p)$, $\mu = np$, $a \geq 0$ and $\varphi(x) = (1+x)\ln(1+x) - x$ (for $x \geq -1$, or ∞ otherwise). Then the following inequalities hold:*

$$\Pr[X \leq \mu - a] \leq e^{-\mu\varphi\left(\frac{-a}{\mu}\right)} \leq e^{-\frac{a^2}{2\mu}}, \quad (2)$$

$$\Pr[X \geq \mu + a] \leq e^{-\mu\varphi\left(\frac{a}{\mu}\right)} \leq e^{-\frac{a^2}{2(\mu+a/3)}}. \quad (3)$$

We define the notion of read- k families, a useful concept when proving concentration results for weakly dependent variables.

Definition 1 (Read- k families). *Let X_1, \dots, X_m be independent random variables. For $j \in [r]$, let $P_j \subseteq [m]$ and let f_j be a Boolean function of $\{X_i\}_{i \in P_j}$. Assume that $|\{j | i \in P_j\}| \leq k$ for every $i \in [m]$. Then, the random variables $Y_j = f_j(\{X_i\}_{i \in P_j})$ are called a read- k family.*

Theorem 4 (Concentration of Read- k families [GLS+15]). *Let Y_1, \dots, Y_r be a family of read- k indicator variables with $\Pr[Y_i = 1] = q$. Then for any $\epsilon > 0$,*

$$\Pr\left[\sum_{i=1}^r Y_i \geq (q + \epsilon)r\right] \leq e^{-D_{\text{KL}}(q+\epsilon||q) \cdot r/k} \quad (4)$$

and

$$\Pr\left[\sum_{i=1}^r Y_i \leq (q - \epsilon)r\right] \leq e^{-D_{\text{KL}}(q-\epsilon||q) \cdot r/k}. \quad (5)$$

Here, D_{KL} is Kullback-Leibler divergence defined as

$$D_{\text{KL}}(q||p) = q \log\left(\frac{q}{p}\right) + (1-q) \log\left(\frac{1-q}{1-p}\right).$$

We will use the following corollary of Theorem 5, which provides Chernoff-type bounds for read- k families. This is derived in a similar way that Chernoff multiplicative bounds are derived from Equations (3) and (2), see [McD98]. Notice that the main difference compared to the standard Chernoff bounds is the extra k factor in denominator of the exponent.

Theorem 5 (Concentration of Read- k families [GLS+15]). *Let Y_1, \dots, Y_r be a family of read- k indicator variables with $\Pr[Y_i = 1] = q$. Also, let $Y = \sum_{i=1}^r Y_i$. Then for any $\epsilon > 0$,*

$$\Pr[Y \geq (1 + \epsilon)\mathbb{E}[Y]] \leq e^{-\frac{\epsilon^2 \mathbb{E}[Y]}{2k(1+\epsilon/3)}} \quad (6)$$

$$\Pr[Y \leq (1 - \epsilon)\mathbb{E}[Y]] \leq e^{-\frac{\epsilon^2 \mathbb{E}[Y]}{2k}}. \quad (7)$$

3 Proposed Method

We prove our main result through a sequence of claims and lemmas. For completeness we include all proofs even if some claims are classic, e.g., Claim 1. At a high level, our proof strategy is as follows:

1. We compute the probability that a simple path between u and v provides us with the correct information on whether $f(u) = f(v)$ or not.
2. Let $L = \frac{\log n}{\log \log n}$. We show that there exist $N = (2c)^{-L} e^{\frac{4L}{5}}$ *almost edge-disjoint paths* of length $(1 + o(1))L$ between any pair of vertices with probability at least $1 - \frac{1}{n^3}$. The reader can think of the paths as being edge-disjoint, if that is helpful; we shall clarify both what we mean by *almost edge-disjoint paths* and how it affects the proof later in the paper.
3. For each path from the collection of N almost edge-disjoint paths, we compute the product of the sign of the edges along the path. Since the paths are not entirely edge disjoint, the corresponding random variables are weakly dependent. We use concentration of multivariate polynomials [GLS+15], see also [AS04; KV00], in combination with Claim 1 to show that using the majority of the N resulting signs to decide whether $f(u) = f(v)$ or not for a pair of nodes $u, v \in V(G)$ gives the correct answer with probability lower bounded by $1 - \frac{1}{n^3}$. Taking the union bound over $\binom{n}{2}$ pairs concludes the proof.

The pseudo-code is shown as Algorithm 1. The algorithm runs over each pair of nodes, and it invokes Algorithm 2 to construct almost edge-disjoint paths for each pair of nodes u, v using Breadth First Search. Note that since we perform $20n \log n (2c)^{-L}$ queries uniformly at random, the resulting graph is asymptotically equivalent to $G \sim G(n, \frac{40 \log n (2c)^{-L}}{n})$, see [FK15, Chapter 1]. Here, $G(n, p)$ is the classic Erdős-Rényi model (a.k.a random binomial graph model) where each possible edge between each pair $(u, v) \in \binom{[n]}{2}$ is included in the graph with probability p independent from every other edge.

It turns out that our algorithm needs an average degree $O\left(\frac{\log n}{(2c)^L}\right)$ *only for the first level* of the trees T_u, T_v that we grow from u and v when we invoke Algorithm 2. For all other levels of the grown trees, we need the degree to be only $O(\log n)$. This difference in the branching factors exists in order to ensure that the number of leaves of trees T_u, T_v in Algorithm 2 is amplified by a factor of $\frac{1}{(2c)^L}$, which then allows us to apply Theorem 5. Using appropriate data structures, Algorithm 1 runs in $O(n^2(n+m)) = O(n^3 \log n (2c)^{-L})$. One can improve the run time in expectation by sampling $O(\log n)$ neighbors for each node in $O(\log n)$ time instead of $O(\log n (2c)^{-L})$ time using a standard sublinear sampling technique that generates geometric random variables between successive successes, see [Knu07; TKM11]. This results in total expected run time $O(n^3 \log n)$. Since we use a branching factor of $O(\log n)$ for all except the first two levels of T_u, T_v , we work with the $G(n, p)$ model with $p = \frac{40 \log n}{n}$ to construct the set of almost edge disjoint paths. (Alternatively, one can think that we start with the larger random graph with more edges, and then in the construction of the almost edge disjoint paths we subsample a smaller collection of edges to use in this stage.) The diameter of this graph *whp* grows asymptotically as L [Bol98] for this value of p . We use the $G(n, \frac{40 \log n (2c)^{-L}}{n})$ model only in Lemma 1 to prove that every node has degree at least $5 \log n (2c)^{-L}$.

The following result is well known but we present a proof for completeness.

Algorithm 1 2-Correlation Clustering with Noise

$L \leftarrow \frac{\log n}{\log \log n}$
Perform $20n \log n (2c)^{-L}$ queries uniformly at random.
Let $G(V, E, \tilde{f})$ be the resulting graph, $\tilde{f} : E \rightarrow \{+1, -1\}$
for each item pair u, v **do**
 $\mathcal{P}_{u,v} = \{P_1, \dots, P_N\} \leftarrow \text{Almost-Edge-Disjoint-Paths}(u, v)$
 $Y_i \leftarrow \prod_{e \in P_i} \tilde{f}(e)$ for $i = 1, \dots, N$
 $Y_{uv} \leftarrow \sum_{P \in \mathcal{P}_{u,v}} Y_P$
 if $Y_{uv} \geq 0$ **then**
 $\tilde{f}(u) = \tilde{f}(v)$
 else
 $\tilde{f}(u) \neq \tilde{f}(v)$
 end if
end for

Algorithm 2 Almost-Edge-Disjoint-Paths(u, v)

Require: $G(V, E, \tilde{f})$, $u, v \in V(G)$

$$\epsilon \leftarrow \frac{1}{\sqrt{\log \log n}}$$

Using Breadth First Search (BFS) grow a tree T_u starting from u as follows.

For the first level of the tree, we choose $4 \log n (2c)^{-\frac{\log n}{\log \log n}}$ neighbors of u .

For the rest of the tree we use a branching factor equal to $4 \log n$ until it reaches depth equal to ϵL . Similarly, grow a tree T_v rooted at v , node disjoint from T_u of equal depth.

From each leaf u_i (v_i) of T_u (T_v) for $i = 1, \dots, N$ grow node disjoint trees until they reach depth $(\frac{1}{2} + \epsilon)L$ with branching factor $4 \log n$. Finally, find an edge between T_{u_i}, T_{v_i}

Claim 1. Consider a path P_{uv} between nodes u, v of length L . Let $R_{uv} = \prod_{e \in P_{uv}} \tilde{f}(e)$. Then,

$$\Pr [R_{uv} = 1 | f(u) = f(v)] = \Pr [R_{uv} = -1 | f(u) \neq f(v)] = \frac{1 + (1 - 2q)^L}{2}$$

Proof. Here $R_{uv} = 1$ iff \tilde{f} agrees with the unknown clustering function f on u, v . This happens when the number of corrupted edges along that path P_{uv} is even. Let $Z_{uv} \sim \text{Bin}(L, q)$ be the number of corrupted edges/sign flips along the P_{uv} path. Clearly, $\Pr [Z_{uv} \text{ is even}] + \Pr [Z_{uv} \text{ is odd}] = 1$. Also,

$$(1 - 2q)^L = \sum_{k=0}^L \binom{L}{k} (-q)^k (1 - q)^{L-k} = \sum_{k=0}^{\lfloor \frac{L}{2} \rfloor} \binom{L}{2k} q^{2k} (1 - q)^{L-2k} - \sum_{k=0}^{\lfloor \frac{L}{2} \rfloor} \binom{L}{2k+1} q^{2k+1} (1 - q)^{L-(2k+1)} = \\ \Pr [Z_{uv} \text{ is even}] - \Pr [Z_{uv} \text{ is odd}].$$

Therefore $\Pr [Z_{uv} \text{ is even}] = \frac{1+(1-2q)^L}{2}$, and the result follows. ■

The next lemma is a direct corollary of the lower tail multiplicative Chernoff bound.

Lemma 1. Let $G \sim G(n, \frac{40 \log n}{(2c)^L n})$ be a random binomial graph. Then whp all vertices have degree greater than $5 \log n (2c)^{-L}$.

Proof. The degree $\deg(u)$ of a node $u \in V(G)$ follows the binomial distribution $\text{Bin}(n-1, \frac{40 \log n}{(2c)^L})$. Set $\delta = \frac{3}{4}$. Then

$$\Pr \left[\deg(u) < 5 \log n (2c)^{-L} \right] < e^{-\frac{\delta^2}{2} 40 \log n (2c)^{-L}} \ll n^{-1}.$$

Taking a union bound over n vertices gives the result. \blacksquare

Now we proceed to our construction of sufficiently enough almost edge-disjoint paths. Our construction is based on standard techniques in random graph theory [BFS+98; DFT15; FT12; Tso13], we include the full proofs for completeness.

Lemma 2. *Let $G \sim G(n, p)$ where $p = \frac{40 \log n}{n}$. Fix $t \in \mathbb{Z}^+$ and $0 < \alpha < 1$. Then, whp there does not exist a subset $S \subseteq [n]$, such that $|S| \leq \alpha t L$ and $e[S] \geq |S| + t$.*

Proof. Set $s = |S|$. Then,

$$\begin{aligned} \Pr [\exists S : s \leq \alpha t L \text{ and } e[S] \geq s + t] &\leq \sum_{s \leq \alpha t L} \binom{n}{s} \binom{\binom{s}{2}}{s+t} p^{s+t} \leq \sum_{s \leq \alpha t L} \left(\frac{ne}{s} \right)^s \left(\frac{es^2 p}{2(s+t)} \right)^{s+t} \\ &\leq \sum_{s \leq \alpha t L} (e^{2+o(1)} \log n)^s \left(\frac{20es \log n}{n} \right)^t \leq \alpha t L \left((e^{2+o(1)} \log n)^{\alpha L} \left(\frac{20eat \log^2 n}{n \log \log n} \right)^t \right) < \frac{1}{n^{(1-\alpha-o(1))t}}. \end{aligned}$$

\blacksquare

Lemma 3. *Let T be a rooted tree of depth at most $\frac{4L}{7}$ and let v be a vertex not in T . Then with probability $1 - o(n^{-3})$, v has at most 10 neighbors in T , i.e., $|N(v) \cap T| \leq 10$.*

Proof. Let T be a rooted tree of depth at most $\frac{4L}{7}$ and let S consist of v , the neighbors of v in T plus the ancestors of these neighbors. Set $b = |N(v) \cap T|$. Then $|S| \leq 4bL/7 + 1 \leq 3bL/5$ and $e(S) = |S| + b - 2$. It follows from Lemma 2 with $\alpha = 3/5$ and $t = 8$, that we must have $b \leq 10$ with probability $1 - o(n^{-3})$. \blacksquare

We show that by growing trees iteratively we can construct sufficiently many edge-disjoint paths for n sufficiently large.

Lemma 4. *Let $\epsilon = \frac{1}{\sqrt{\log \log n}}$, and $k = \epsilon L$. For all pairs of vertices $x, y \in [n]$ there exists a subgraph $G_{x,y}(V_{x,y}, E_{x,y})$ of G as shown in figure 1, whp. The subgraph consists of two isomorphic vertex disjoint trees T_x, T_y rooted at x, y each of depth k . T_x and T_y both have a branching factor of $4 \log n (2c)^{-L}$ for the first level, and $4 \log n$ for the remaining levels. If the leaves of T_x are $x_1, x_2, \dots, x_\tau, \tau \geq (2c)^{-L} n^{4\epsilon/5}$ then $y_i = f(x_i)$ where f is a natural isomorphism. Between each pair of leaves $(x_i, y_i), i = 1, 2, \dots, m$ there is a path P_i of length $(1+2\epsilon)L$. The paths $P_i, i = 1, 2, \dots, \tau, \dots$ are edge disjoint.*

Proof. Because we have to do this for all pairs x, y , we note without further comment that likely (resp. unlikely) events will be shown to occur with probability $1 - o(n^{-2})$ (resp. $o(n^{-2})$).

To find the subgraph shown in Figure 1(b) we grow tree structures as shown in Figure 1(a). Specifically, we first grow a tree from x using BFS until it reaches depth k . Then, we grow a tree starting from y again using BFS until it reaches depth k . For the first level of both trees, we choose $\frac{5 \log n}{(2c)^L}$ neighbors of x, y respectively. For all other levels we use a branching factor equal to $4 \log n$.

Before we show how to continue our construction, we need to prune down the degree of $G([n] \setminus \{x, y\})$ so that the remaining subgraph behave as $G(n, p)$ with $p = \Theta(\frac{\log n}{n})$. This can be achieved for example either by considering a random subgraph of G according to $G([n] \setminus \{x, y\}, \frac{40 \log n}{n})$, applying Chernoff bounds as in Lemma 1 to show that each node has degree at least $5 \log n$, or by letting each node choose $5 \log n$ neighbors uniformly at random.

Finally, once trees T_x, T_y have been constructed, we grow trees from the leaves of T_x and T_y using BFS for depth $\gamma = (\frac{1}{2} + \epsilon)L$. Now we analyze these processes. Since the argument is the same we explain it in detail for T_x and we outline the differences for the other trees. We use the notation $D_i^{(\rho)}$ for the number of vertices at depth i of the BFS tree rooted at ρ .

First we grow T_x . As we grow the tree via BFS from a vertex v at depth i to vertices at depth $i + 1$ certain *bad* edges from v may point to vertices already in T_x . Lemma 3 shows with probability $1 - o(n^{-3})$ there can be at most 10 bad edges emanating from v .

Hence, we obtain the recursion

$$D_{i+1}^{(x)} \geq (5 \log n - 10) (D_i^{(x)} - 1) \geq 4 \log n D_i^{(x)}. \quad (8)$$

Therefore the number of leaves satisfies

$$D_k^{(x)} \geq \frac{1}{(2c)^L} (4 \log n)^{\epsilon L} \geq \frac{1}{(2c)^L} n^{4\epsilon/5}. \quad (9)$$

We can make the branching factors exactly $4 \log n (2c)^{-L}$ for the first level and $4 \log n$ for all remaining levels by pruning. We do this so that the trees T_x are isomorphic to each other. With a similar argument

$$D_k^{(y)} \geq \frac{1}{(2c)^L} n^{\frac{4}{5}\epsilon}. \quad (10)$$

The only difference is that now we also say an edge is bad if the other endpoint is in T_x . This immediately gives

$$D_{i+1}^{(y)} \geq (5 \log n - 20) (D_i^{(y)} - 1) \geq 4 \log n D_i^{(y)}$$

and the required conclusion (10).

Similarly, from each leaf $x_i \in T_x$ and $y_i \in T_y$ we grow trees $\widehat{T}_{x_i}, \widehat{T}_{y_i}$ of depth $\gamma = (\frac{1}{2} + \epsilon)L$ using the same procedure and arguments as above. Lemma 3 implies that there are at most 20 edges from the vertex v being explored to vertices in any of the trees already constructed (at most 10 to T_x plus any trees rooted at an x_i and another 10 for y). The number of leaves of each \widehat{T}_{x_i} now satisfies

$$\widehat{D}_\gamma^{(x_i)} \geq (4 \log n)^{\gamma+1} \geq n^{\frac{1}{2} + \frac{4}{5}\epsilon}.$$

The result is similar for $\widehat{D}_\gamma^{(y_i)}$.

Observe next that BFS does not condition on the edges between the leaves X_i, Y_i of the trees \widehat{T}_{x_i} and \widehat{T}_{y_i} . That is, we do not need to look at these edges in order to carry out our construction. On the other hand we have conditioned on the occurrence of certain events to imply a certain growth rate. We handle this technicality as follows. We go through the above construction and halt if ever we find that we cannot expand by the required amount. Let \mathbf{A} be the event that we do not halt the construction i.e. we fail the conditions of Lemmas 2 or 3. We have $\Pr[\mathbf{A}] = 1 - o(1)$ and so,

$$\Pr[\exists i : e(X_i, Y_i) = 0 \mid \mathbf{A}] \leq \frac{\Pr[\exists i : e(X_i, Y_i) = 0]}{\Pr(\mathbf{A})} \leq 2n^{\frac{4\epsilon}{5}} (1-p)^{n^{1+\frac{8\epsilon}{5}}} \leq n^{-n^\epsilon}.$$

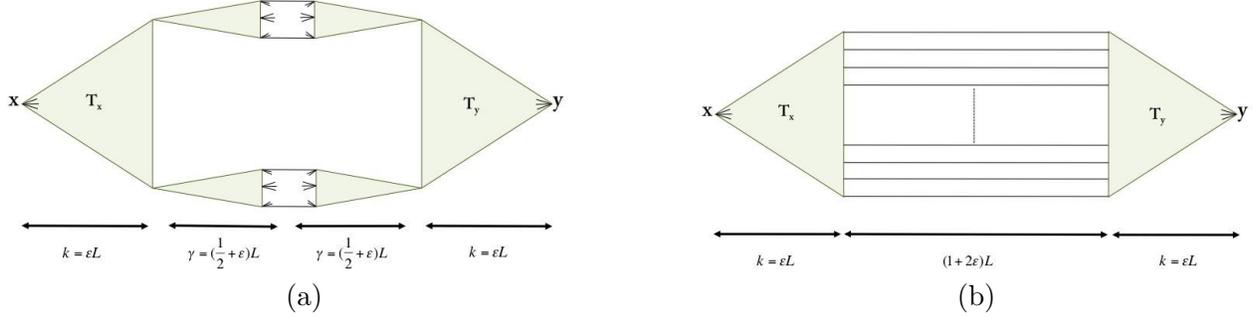


Figure 1: Illustration of construction in Lemma 4. (a) By repeatedly growing trees appropriately, (b) we create for each pair of nodes x, y two node disjoint trees T_x, T_y of depth $k = \epsilon L$ whose leaves can be matched via a natural isomorphism and linked with edge disjoint paths of length $(1 + o(1))L$

We conclude that *whp* there is always an edge between each X_i, Y_i and thus a path of length at most $(1 + 2\epsilon)L$ between each x_i, y_i . ■

The proof of Theorem 1 follows. Set $q = \frac{1}{2} - c$.

Proof of Theorem 1. Fix a pair of nodes $x, y \in V(G)$. Let Y_1, \dots, Y_N be the signs of the N edge disjoint paths connecting them, i.e., $Y_i \in \{-1, +1\}$ for all i . Also let $Y = \sum_{i=1}^N Y_i$. Notice that $\{Y_1, \dots, Y_N\}$ is a read- k family where $k = \frac{N}{4 \log n (2c)^{-L}}$.

By the linearity of expectation

$$\mathbb{E}[Y] = N(2c)^L \geq n^{\frac{4}{5}\epsilon} (2c)^L.$$

By applying Theorem 5 we obtain

$$\Pr[Y < 0] = \Pr[Y - \mathbb{E}[Y] < -\mathbb{E}[Y]] \leq \exp\left(-\frac{n^{4/5\epsilon} (2c)^L}{\frac{2n^{4/5\epsilon}}{4(2c)^{-L} \log n}}\right) = o(n^{-2}).$$

■

Planted bisection model. Before we prove Theorem 2 we discuss the connection between our formulation and the well studied planted bisection model. Suppose that n is even, and the graph has two clusters of equal size. The probabilities of connecting are p within each cluster, and $q < p$ across the clusters. Now recall our setting as described in Model I, and consider just the edges that correspond to queries that return $+1$. These form a graph drawn from the planted bisection model where $p = \frac{1+c}{2} \times \frac{40 \log n}{n}$, $q = \frac{1-c}{2} \times \frac{40 \log n}{n}$. Therefore, one can apply existing methods for exact recovery, e.g., [ABH16; McS01] instead of our method when the sizes of the two clusters are (roughly) equal. It is worth noting that despite the wide variety of techniques that appear in the context of the planted partition model, including the EM algorithm [SN97], spectral methods [McS01; Vu14], semidefinite programming [ABH16; HWX16; MS15], hill-climbing [CI01], Metropolis algorithm [JS98], modularity based methods [BC09], our edge-disjoint path technique is novel in this context.

Hajek, Wu, and Xu proved that when each cluster has $\Theta(n)$ nodes, the average degree has to scale as $\frac{\log n}{(\sqrt{1-q}-\sqrt{q})^2}$ for exact recovery [HWX16]. Also, they showed that using semidefinite

programming (SDP) exact recovery is achievable at this threshold [HWX16]. Note that as $q \rightarrow \frac{1}{2}$, this lower bound scales as $\Theta(\frac{\log n}{(1-2q)^2}) = \Theta(\frac{\log n}{(2c)^2})$. It is an interesting theoretical problem to explore if the techniques we develop in this work, or similar techniques can get closer to this lower bound.

Proof of Theorem 2. Since the proof of Theorem 2 overlaps with the proof of Theorem 1, we outline the main differences. Let us return to the basic version of Model II, and let $X(e) \in \{-1, 0, 1\}$ for $e = (x, y)$ be

$$\tilde{f}(e) - (g(x) - g(y)) \bmod k.$$

Then given a path between two vertices u and v ,

$$g(v) = g(u) + \sum_{e \in P_{uv}} \tilde{f}(e) - \sum_{e \in P_{uv}} X(e) \bmod k.$$

Our question is now what is $Z_{uv} = \sum_{e \in P_{uv}} X(e) \bmod k$. We would like that Z_{uv} be (even slightly) more highly concentrated on 0 than on other values, so that when $g(u) = g(v)$, we find that the sum of the return values from our algorithm, $\sum_{e \in P_{uv}} \tilde{f}(e) \bmod k$, is most likely to be 0. We could then conclude by looking over many almost edge-disjoint paths that if this sum is 0 over a plurality of the paths, then u and v are in the same group *whp*.

For our simple error model, the sum $\sum_{e \in P_{uv}} X(e) \bmod k$ behaves like a simple lazy random walk on the cycle of values modulo k , where the probability of remaining in the same state at each step is q . Let us consider this Markov chain on the values modulo k ; we refer to the values as states. Let p_{ij}^t be the probability of going from state i to state j after t steps in such a walk. It is well known that one can derive explicit formulae for p_{ij}^t ; see e.g. [Fel68, Chapter XVI.2]. It also follows by simply finding the eigenvalues and eigenvectors of the matrix corresponding to the Markov chain and using that representation. One can check the resulting forms to determine that p_{0j}^t is maximized when $j = 0$, and to determine the corresponding gap $\max_{j \in [1, k-1]} |p_{00}^t - p_{0j}^t|$. Based on this gap, we can apply Chernoff-type bounds as in Theorem 5 to show that the plurality of almost edge-disjoint paths will have error 0, allowing us to determine whether the endpoints of the path x and y are in the same group with high probability.

The simplest example is with $k = 3$ groups, where we find

$$p_{00}^t = \frac{1}{3} + \frac{2}{3} (1 - 3q/2)^t,$$

and

$$p_{01}^t = p_{02}^t = \frac{1}{3} - \frac{1}{3} (1 - 3q/2)^t.$$

In our case $t = L$, and we see that for any $q < 2/3$, p_{00}^t is large enough that we can detect paths using the same argument as in Model I.

For general k , we use that the eigenvalues of the matrix

$$\begin{bmatrix} 1-q & q & 0 & \dots & q \\ q & 1-q & q & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q & 0 & 0 & \dots & 1-q \end{bmatrix}$$

are $1 - q + q \cos(2\pi j/k)$, $j = 0, \dots, k-1$, with the j -th corresponding eigenvector being $[1, \omega^j, \omega^{2j}, \dots, \omega^{j(k-1)}]$ where $\omega = e^{2\pi i/k}$ is a primitive k th root of unity. Here, i is not an index but the square root of -1,

i.e., $i = \sqrt{-1}$. In this case we have

$$p_{00}^t = \frac{1}{k} + \frac{1}{k} \sum_{j=1}^{k-1} (1 - q + q \cos(2\pi j/k))^t.$$

Note that $p_{00}^t > 1/k$. Some algebra reveals that the next largest value of p_{0j}^t belongs to p_{01}^t , and equals

$$p_{01}^t = \frac{1}{k} + \frac{1}{k} \sum_{j=1}^{k-1} \omega^{-j} (1 - q + q \cos(2\pi j/k))^t.$$

We therefore see that the error between ends of a path again have the plurality value 0, with a gap of at least

$$p_{00}^t - p_{01}^t \geq 2(1 - \cos(2\pi/k))(1 - q + q \cos(2\pi/k))^t.$$

This gap is constant for any constant $k \geq 3$ and $q \leq 1/2$. ■

As we have already mentioned, the same approach could be used for the more general setting where

$$\tilde{f}(e) = g(x) - g(y) + j \quad \text{with probability } q_j, 0 \leq j < k,$$

but now one works with the Markov chain matrix

$$\begin{bmatrix} q_0 & q_1 & q_2 & \cdots & q_{k-1} \\ q_{k-1} & q_0 & q_1 & \cdots & q_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & q_3 & \cdots & q_0 \end{bmatrix}.$$

4 Related Work

Fritz Heider introduced the notion of a signed graph, with $+1$ or -1 labels on the edges, in the context of balance theory [Hei46]. The key subgraph in balance theory is the *triangle*: any set of three fully interconnected nodes whose product of edge signs is negative is not balanced. The complete graph is balanced if every one of its triangles is balanced. Early work on signed graphs focused on graph theoretic properties of balanced graphs [CH56]. Harary proved the famous balance theorem which characterizes balanced graphs [Har53].

Bansal et al. [BBC04] studied Correlation Clustering: given an undirected signed graph partition the nodes into clusters so that the total number of disagreements is minimized. This problem is NP-hard [BBC04; SST04]. Here, a disagreement can be either a positive edge between vertices in two clusters or a negative edge between two vertices in the same cluster. Note that in Correlation Clustering the number of clusters is not specified as part of the input. The case when the number of clusters is constrained to be at most two is known as 2-Correlation-Clustering.

Minimizing disagreements is equivalent to maximizing the number of agreements. However, from an approximation perspective these two versions are different: minimizing is harder. For minimizing disagreements, Bansal et al. [BBC04] provide a 3-approximation algorithm for 2-Correlation-Clustering, and Giotis and Guruswami provide a polynomial time approximation scheme (PTAS) [GG06]. Ailon et al. designed a 2.5-approximation algorithm [ACN08] that was further improved by Coleman et al. to a 2-approximation [CSW08]. We remark that the notion of *imbalance* studied by Harary is the 2-Correlation-Clustering cost of the signed graph. Mathieu and

Schudy initiated the study of noisy correlation clustering [MS10]. They develop various algorithms when the graph is complete, both for the cases of a random and a semi-random model. Later, Makarychev, Makarychev, and Vijayaraghavan proposed an algorithm for graphs with $O(npoly \log n)$ edges under a semi-random model [MMV15].

When the graph is not complete Correlation Clustering and MINIMUM MULTICUT reduce to one another leading to a $O(\log n)$ approximations [CGW03; DI03]. The case of constrained size clusters has recently been studied by Puleo and Mileknovic [PM15]. Finally, by using the Goemans-Williamson SDP relaxation for MAX CUT [GW95], one can obtain a 0.878 approximation guarantee for 2-Correlation-Clustering problem as noted by [CSW08].

Chen et al. [CJS+14; CSX12] consider also model I as described in Section 1 and provide a method that can reconstruct the clustering for random binomial graphs with $O(npoly \log n)$ edges. Their method exploits low rank properties of the cluster matrix, and requires certain conditions, including conditions on the imbalance between clusters, see [CSX12, Theorem 1, Table 1] to be true. Their methods is based on a convex relaxation of a low rank problem. Also, closely related to our work lies the work of Cesa-Bianchi et al. [CBGV+12] that take a learning-theoretic perspective on the problem of predicting signs. They consider three types of models: batch, online, and active learning, and provide theoretical bounds for prediction mistakes for each setting. They use the correlation clustering objective as their learning bias, and they show that the risk of the empirical risk minimizer is controlled by the correlation clustering objective. Chian et al. point out that the work of Candès and Tao [CRT06] can be used to predict signs of edges, and also provide various other methods, including singular value decomposition based methods, for the sign prediction problem [CHN+14]. The incoherence is the key parameter that determines the number of queries, and is equal to the group imbalance $\tau = \max_{\text{cluster } C} \frac{n}{|C|}$. The number of queries needed for exact recovery under Model I is $O(\tau^4 n \log^2 n)$.

5 Conclusion

In this work we have studied the edge sign prediction problem, showing that under our proposed correlation clustering formulation and a fully random noise model querying $O(n^{1+o(1)} \log n)$ pairs of nodes uniformly at random suffices to recover the clusters efficiently, *whp*. We also provided a generalization of our model and proof approach to more than two clusters. While our work here is theoretical, in future work we plan to apply our method and additional heuristics to real data, and compare against related alternatives. From a theoretical perspective, it is an interesting problem to close the gap between our upper bound and the known $\frac{\log n}{(2c)^2}$ lower bound for exact recovery [HWX16] using techniques based on BFS.

Acknowledgment

We would like to thank Bruce Hajek and Zeyu Zhou for detecting an error in an earlier version of our work. We also want to thank Yury Makarychev, Konstantin Makarychev, and Aravindan Vijayaraghavan for their feedback.

This work was supported in part by NSF grants CNS-1228598, CCF-1320231, CCF-1563710, and CCF-1535795.

References

- [ABH16] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. “Exact recovery in the stochastic block model”. In: *IEEE Transactions on Information Theory* 62.1 (2016), pp. 471–487 (cit. on pp. 1, 8).
- [ACN08] Nir Ailon, Moses Charikar, and Alantha Newman. “Aggregating inconsistent information: ranking and clustering”. In: *Journal of the ACM (JACM)* 55.5 (2008), p. 23 (cit. on p. 10).
- [AS04] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2004 (cit. on p. 4).
- [BBC04] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. “Correlation clustering”. In: *Machine Learning* 56.1-3 (2004), pp. 89–113 (cit. on p. 10).
- [BC09] Peter J Bickel and Aiyou Chen. “A nonparametric view of network models and Newman–Girvan and other modularities”. In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21068–21073 (cit. on p. 8).
- [BFS+98] Andrei Z Broder, Alan M Frieze, Stephen Suen, and Eli Upfal. “Optimal construction of edge-disjoint paths in random graphs”. In: *SIAM Journal on Computing* 28.2 (1998), pp. 541–573 (cit. on p. 6).
- [Bol98] Béla Bollobás. “Random graphs”. In: *Modern Graph Theory*. Springer, 1998 (cit. on p. 4).
- [CBGV+12] Nicolo Cesa-Bianchi, Claudio Gentile, Fabio Vitale, Giovanni Zappella, et al. “A Correlation Clustering Approach to Link Classification in Signed Networks.” In: *COLT*. 2012, pp. 34–1 (cit. on p. 11).
- [CGW03] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. “Clustering with qualitative information”. In: *Proceedings. 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2003, pp. 524–533 (cit. on p. 11).
- [CH56] Dorwin Cartwright and Frank Harary. “Structural balance: a generalization of Heider’s theory.” In: *Psychological review* 63.5 (1956), p. 277 (cit. on pp. 1, 10).
- [CHN+14] Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit S Dhillon, and Ambuj Tewari. “Prediction and clustering in signed networks: a local to global perspective.” In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1177–1213 (cit. on p. 11).
- [CI01] Ted Carson and Russell Impagliazzo. “Hill-climbing finds random planted bisections”. In: *Proceedings of the twelfth annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics. 2001, pp. 903–909 (cit. on p. 8).
- [CJS+14] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. “Clustering partially observed graphs via convex optimization.” In: *Journal of Machine Learning Research* 15.1 (2014), pp. 2213–2238 (cit. on p. 11).
- [CK01] Anne Condon and Richard M Karp. “Algorithms for graph partitioning on the planted partition model”. In: *Random Structures and Algorithms* 18.2 (2001), pp. 116–140 (cit. on p. 1).

- [CRT06] Emmanuel J Candès, Justin Romberg, and Terence Tao. “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information”. In: *IEEE Transactions on information theory* 52.2 (2006), pp. 489–509 (cit. on p. 11).
- [CSW08] Tom Coleman, James Saunderson, and Anthony Wirth. “A local-search 2-approximation for 2-correlation-clustering”. In: *European Symposium on Algorithms (ESA)*. Springer, 2008, pp. 308–319 (cit. on pp. 10, 11).
- [CSX12] Yudong Chen, Sujay Sanghavi, and Huan Xu. “Clustering sparse graphs”. In: *Advances in neural information processing systems*. 2012, pp. 2204–2212 (cit. on p. 11).
- [DFT15] Andrzej Dudek, Alan M Frieze, and Charalampos E Tsourakakis. “Rainbow connection of random regular graphs”. In: *SIAM Journal on Discrete Mathematics* 29.4 (2015), pp. 2255–2266 (cit. on p. 6).
- [DI03] Erik D Demaine and Nicole Immorlica. “Correlation clustering with partial information”. In: *Approximation, Randomization, and Combinatorial Optimization (APPROX-RANDOM)*. Springer, 2003, pp. 1–13 (cit. on p. 11).
- [EK10] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010 (cit. on p. 1).
- [FK15] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2015 (cit. on p. 4).
- [FT12] Alan Frieze and Charalampos E Tsourakakis. “Rainbow connectivity of sparse random graphs”. In: *Approximation, Randomization, and Combinatorial Optimization (APPROX-RANDOM)*. Springer, 2012, pp. 541–552 (cit. on p. 6).
- [Fel68] William Feller. *An introduction to probability theory and its applications: volume I*. Vol. 3. John Wiley & Sons London-New York-Sydney-Toronto, 1968 (cit. on p. 9).
- [GG06] Ioannis Giotis and Venkatesan Guruswami. “Correlation clustering with a fixed number of clusters”. In: *Proceedings of the seventeenth annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics. 2006, pp. 1167–1176 (cit. on p. 10).
- [GLS+15] Dmitry Gavinsky, Shachar Lovett, Michael Saks, and Srikanth Srinivasan. “A tail bound for read-k families of functions”. In: *Random Structures & Algorithms* 47.1 (2015), pp. 99–108 (cit. on pp. 3, 4).
- [GW95] Michel X Goemans and David P Williamson. “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. In: *Journal of the ACM (JACM)* 42.6 (1995), pp. 1115–1145 (cit. on p. 11).
- [HEP+16] Jack P Hou, Amin Emad, Gregory J Puleo, Jian Ma, and Olgica Milenkovic. “A new correlation clustering method for cancer mutation analysis”. In: *arXiv preprint arXiv:1601.06476* (2016) (cit. on p. 1).
- [HWX16] Bruce Hajek, Yihong Wu, and Jiaming Xu. “Achieving exact cluster recovery threshold via semidefinite programming”. In: *IEEE Transactions on Information Theory* 62.5 (2016), pp. 2788–2797 (cit. on pp. 1, 8, 9, 11).
- [Har53] Frank Harary. “On the notion of balance of a signed graph.” In: *The Michigan Mathematical Journal* 2.2 (1953), pp. 143–146 (cit. on pp. 1, 10).

- [Hei46] Fritz Heider. “Attitudes and cognitive organization”. In: *The Journal of psychology* 21.1 (1946), pp. 107–112 (cit. on pp. 1, 10).
- [JLR11] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random Graphs*. Vol. 45. John Wiley & Sons, 2011 (cit. on p. 3).
- [JS98] Mark Jerrum and Gregory B Sorkin. “The Metropolis algorithm for graph bisection”. In: *Discrete Applied Mathematics* 82.1 (1998), pp. 155–175 (cit. on p. 8).
- [KV00] Jeong Han Kim and Van H Vu. “Concentration of multivariate polynomials and its applications”. In: *Combinatorica* 20.3 (2000), pp. 417–434 (cit. on p. 4).
- [Knu07] Donald E Knuth. “Seminumerical algorithms”. In: (2007) (cit. on p. 4).
- [LHK10a] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. “Predicting positive and negative links in online social networks”. In: *Proceedings of the 19th international conference on World Wide Web (WWW)*. ACM. 2010, pp. 641–650 (cit. on p. 1).
- [LHK10b] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. “Signed networks in social media”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1361–1370 (cit. on p. 1).
- [MMV15] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. “Correlation clustering with noisy partial information”. In: *Proceedings of the Conference on Learning Theory (COLT)*. Vol. 6. 2015, p. 12 (cit. on p. 11).
- [MS10] Claire Mathieu and Warren Schudy. “Correlation clustering with noisy input”. In: *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2010, pp. 712–728 (cit. on p. 11).
- [MS15] Andrea Montanari and Subhabrata Sen. “Semidefinite programs on sparse random graphs and their application to community detection”. In: *arXiv preprint arXiv:1504.05910* (2015) (cit. on p. 8).
- [McD98] Colin McDiarmid. “Concentration”. In: *Probabilistic methods for algorithmic discrete mathematics*. Springer, 1998, pp. 195–248 (cit. on p. 3).
- [McS01] Frank McSherry. “Spectral partitioning of random graphs”. In: *Proceedings. 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2001, pp. 529–537 (cit. on pp. 1, 8).
- [PM15] Gregory J Puleo and Olgica Milenkovic. “Correlation clustering with constrained cluster sizes and extended weights bounds”. In: *SIAM Journal on Optimization* 25.3 (2015), pp. 1857–1872 (cit. on p. 11).
- [SN97] Tom AB Snijders and Krzysztof Nowicki. “Estimation and prediction for stochastic blockmodels for graphs with latent block structure”. In: *Journal of classification* 14.1 (1997), pp. 75–100 (cit. on p. 8).
- [SST04] Ron Shamir, Roded Sharan, and Dekel Tsur. “Cluster graph modification problems”. In: *Discrete Applied Mathematics* 144.1 (2004), pp. 173–182 (cit. on p. 10).
- [TKM11] Charalampos E Tsourakakis, Mihail N Kolountzakis, and Gary L Miller. “Triangle Sparsifiers.” In: *J. Graph Algorithms Appl.* 15.6 (2011), pp. 703–726 (cit. on p. 4).
- [Tso13] Charalampos E. Tsourakakis. “Mathematical and Algorithmic Analysis of Network and Biological Data”. PhD thesis. Carnegie Mellon University, 2013 (cit. on p. 6).

- [Vu14] Van Vu. “A simple SVD algorithm for finding hidden partitions”. In: *arXiv preprint arXiv:1404.3918* (2014) (cit. on p. 8).