Community Recovery in Graphs with Locality

Yuxin Chen[*] YXCHEN@STANFORD.EDU
Govinda M. Kamath[*] GKAMATH@STANFORD.EDU
Changho Suh[†] CHSUH@KAIST.AC.KR
David Tse[*+] DNTSE@STANFORD.EDU

[*] Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
[†] Department of Electrical Engineering, KAIST, Daejeon 305-701, Korea
[+] Department of EECS, University of California, Berkeley CA 94720, USA

Abstract

Motivated by applications in domains such as social networks and computational biology, we study the problem of community recovery in graphs with locality. In this problem, pairwise noisy measurements of whether two nodes are in the same community or different communities come mainly or exclusively from nearby nodes rather than uniformly sampled between all node pairs, as in most existing models. We present two algorithms that run nearly linearly in the number of measurements and which achieve the information limits for exact recovery.

1. Introduction

Clustering of data is a central problem that is prevalent across all of science and engineering. One formulation that has received significant attention in recent years is community recovery (Girvan & Newman, 2002; Fortunato, 2010; Porter et al., 2009), also referred to as correlation clustering (Bansal et al., 2004) or graph clustering (Jalali et al., 2011). In this formulation, the objective is to cluster individuals into different communities based on pairwise measurements of their relationships, each of which gives some noisy information about whether two individuals belong to the same community or different communities. While this formulation applies naturally in social networks, it has a broad range of applications in other domains including protein complex detection (Chen & Yuan, 2006), image segmentation (Shi & Malik, 2000; Globerson et al., 2015), shape matching (Chen et al., 2014a), etc. See (Abbe & Wainwright, 2015) for an introduction.

In recent years, there has been a flurry of works on designing community recovery algorithms based on idealised generative models of the measurement process. A particular popular model is the *Stochastic Block Model* (SBM) (Holland et al., 1983; Condon & Karp, 2001), where the n individuals to be clustered are modeled as nodes on a random graph with statistically more edges between nodes within the same community than between nodes across two different communities. A closely related model is the *Censored Block Model* (CBM) (Abbe et al., 2015), where one obtains noisy parity measurements on the edges of an Erdős-Rényi graph (Durrett, 2007). Both the SBM and the CBM can be unified into one model with noisy measurements which are randomly sampled on the edges of a complete graph, with the two models differing only in the measurement noise model. Thus, a central assumption underlying both models is that it is equally likely to obtain measurements between *any* pair of nodes. This is a very unrealistic assumption in many applications: nodes often have *locality* and it is more likely to obtain data on relationships between nearby nodes than far away nodes. For example, in friendship graphs, individuals that live close by are more likely to interact than nodes that are far away.

This paper focuses on the community recovery problem when the measurements are randomly sampled from graphs with locality structure rather than complete graphs. Our theory covers a broad range of graphs including rings, lines, 2-D grids, and small-world graphs (Fig. 1). Each of these graphs is parametrized by a locality radius r such that nodes within r hops are connected by an edge. We characterize the information limits for community recovery on these networks, i.e. the minimum number of measurements needed to exactly recover the communities as the number of nodes n scales. We propose two algorithms whose complexities are nearly linear in the number of measurements and which can achieve the information limits of all these networks for a very wide range of the radius r. In the special case when the radius r is so large that measurements at all locations are possible, we recover the exact recovery limit identified by (Hajek et al., 2015a) when measurements are randomly sampled from complete graphs.

It is worth emphasizing that various computationally feasible algorithms (Coja-Oghlan, 2010; Chaudhuri et al., 2012; Chen et al., 2014b; Abbe & Sandon, 2015) have been pro-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

posed for more general models beyond the SBM and the CBM, which accommodate multi-community models, the presence of outlier samples, the case where different edges are sampled at different rates, and so on. Most of these models, however, fall short of accounting for any sort of locality constraints. In fact, the results developed in prior literature often lead to unsatisfactory guarantees when applied to graphs with locality, as will be detailed in Section 3. Another recent work (Chen et al., 2015) has determined the order of the information limits in geometric graphs, with no tractable algorithms provided therein. In contrast, our findings uncover a curious phenomenon: the presence of locality does not lead to additional computational barriers: solutions that are information theoretically optimal can often be achieved computational efficiently and, perhaps more surprisingly, within nearly linear time.

2. Problem Formulation and An Application

2.1. Sampling Model

Measurement Graph. Consider a collection of n vertices $\mathcal{V} = \{1, \cdots, n\}$, each represented by a binary-valued vertex variable $X_i \in \{0, 1\}$, $1 \le i \le n$. Suppose it is only feasible to take pairwise samples over a restricted set of locations, as represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that comprises an edge set $\mathcal{E}$. Specifically, for each edge $(i, j) \in \mathcal{E}$ one acquires $N_{i,j}$ samples[1] $Y_{i,j}^{(l)}$ $(1 \le l \le N_{i,j})$, where each sample measures the parity of X_i and X_j. We will use $\mathcal{G}$ to encode the locality constraint of the sampling scheme, and shall pay particular attention to the following families of measurement graphs.

Complete graph: $\mathcal{G}$ is called a complete graph if every pair of vertices is connected by an edge; see Fig. 1(a).

Line: $\mathcal{G}$ is said to be a line $\mathcal{L}_r$ if, for some locality radius r, $(i, j) \in \mathcal{E}$ iff $|i - j| \le r$; see Fig. 1(b).

Ring: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is said to be a ring $\mathcal{R}_r$ if, for some locality radius r, $(i, j) \in \mathcal{E}$ iff $i - j \in [-r, r] \pmod n$; see Fig. 1(c).

Grid: $\mathcal{G}$ is called a grid if (1) all vertices reside within a $\sqrt{n} \times \sqrt{n}$ square with integer coordinates, and (2) two vertices are connected by an edge if they are at distance not exceeding some radius r; see Fig. 1(d).

Small-world graphs: $\mathcal{G}$ is said to be a small-world graph if it is a superposition of a complete graph $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ and another graph $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ with locality. See Fig. 1(e).

Random Sampling. This paper focuses on a random sampling model, where the number of samples $N_{i,j}$ taken over $(i, j) \in \mathcal{E}$ is independently drawn and obeys[2] $N_{i,j} \sim \mathsf{Poisson}(\lambda)$ for some average sampling rate λ. This gives rise to an average total sample size

$$m := \sum\nolimits_{(i,j)\in\mathcal{E}} \mathbb{E}[N_{i,j}] = \lambda |\mathcal{E}| . \tag{1}$$

When m is large, the actual sample size sharply concentrates around m with high probability.

Measurement Noise Model. The acquired parity measurements are assumed to be independent given $N_{i,j}$; more precisely, conditional on $N_{i,j}$,

$$Y_{i,j}^{(l)} = Y_{j,i}^{(l)} \overset{\text{ind.}}{=} \begin{cases} X_i \oplus X_j, & \text{with probability } 1 - \theta \\ X_i \oplus X_j \oplus 1, & \text{else} \end{cases} \tag{2}$$

for some fixed error rate $0 < \theta < 1$, where $\oplus$ denotes modulo-2 addition. This is the same as the noise model in CBM (Abbe et al., 2015). The SBM corresponds to an asymmetric erasure model for the measurement noise, and we expect our results extend to that model as well.

2.2. Goal: Optimal Algorithm for Exact Recovery

This paper centers on exact recovery, that is, to reconstruct all input variables $\boldsymbol{X} = [X_i]_{1\le i\le n}$ precisely up to global offset. This is all one can hope for since there is absolutely no basis to distinguish $\boldsymbol{X}$ from $\boldsymbol{X} \oplus \mathbf{1} := [X_i \oplus 1]_{1\le i\le n}$ given only parity samples. More precisely, for any recovery procedure ψ the probability of error is defined as

$$P_{\mathrm{e}}(\psi) := \max_{\boldsymbol{X}\in\{0,1\}^n} \mathbb{P}\{\psi(\boldsymbol{Y}) \ne \boldsymbol{X} \text{ and } \psi(\boldsymbol{Y}) \ne \boldsymbol{X} \oplus \mathbf{1}\},$$

where $\boldsymbol{Y} := \{Y_{i,j}^{(l)}\}$. The goal is to develop an algorithm whose sample complexity approaches the information limit m^* (as a function of (n, θ)), that is, the minimum sample size m under which $\inf_\psi P_{\mathrm{e}}(\psi)$ vanishes as n scales.

2.3. Haplotype Phasing: A Motivating Application

Humans have 23 pairs of homologous chromosomes, one maternal and one paternal. Each pair are identical sequences of nucleotides A,G,C,T's except on certain documented positions called single nucleotide polymorphisms (SNPs), or genetic variants. The problem of haplotype phasing is that of determining which variants are on the same chromosome in each pair, and has important applications such as in personalized medicine and human genetics. The advent of next generation sequencing technologies allows haplotype phasing by providing linking reads between multiple SNP locations (Browning & Browning, 2011; Donmez & Brudno, 2011; Das & Vikalo, 2015).

One can formulate the problem of haplotype phasing as recovery of two communities of SNP locations, those with the variant on the maternal chromosome and those with the variant on the paternal chromosome (Si et al., 2014; Kamath et al., 2015). Each pair of linking reads gives a noisy measurement of whether two SNPs have the variant on the same chromosome or different chromosomes. While there

[1] We adopt the convention that $N_{i,j} \equiv 0$ for any $(i, j) \notin \mathcal{E}$.

[2] All results presented in this paper hold under a related model where $N_{i,j} \sim \mathsf{Bernoulli}(\lambda)$, as long as $|\mathcal{E}| \gg n \log n$ and $\lambda \le 1$ (which is the regime accommodated in all theorems).

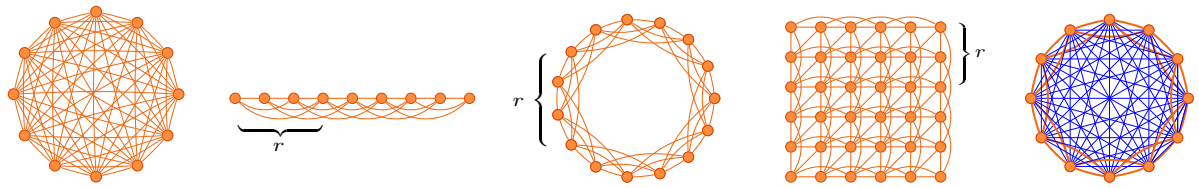


Figure 1. Examples of (a) complete graph, (b) line, (c) ring, (d) grid, and (e) small-world graph.

are of the order of $n = 10^5$ SNPs on each chromosome, the linking reads are typically only several SNPs or at most 100 SNPs apart, depending on the specific sequencing technology. Thus, the measurements are sampled from a line graph like in Fig. 1(b) with locality radius $r \ll n$.

2.4. Other Useful Metrics and Notation

One key metric that captures the distinguishability between two probability measures P_0 and P_1 is the *Chernoff information* (Cover & Thomas, 2006), defined as

$$D^*(P_0, P_1) := -\inf_{0 \le \tau \le 1} \log \left\{ \sum\nolimits_y P_0^\tau(y) P_1^{1-\tau}(y) \right\}. \quad (3)$$

For instance, when $P_0 \sim \mathsf{Bernoulli}(\theta)$ and $P_1 \sim \mathsf{Bernoulli}(1-\theta)$, D^* simplifies to

$$D^* = \mathsf{KL}(0.5 \,\|\, \theta) = 0.5 \log \frac{0.5}{\theta} + 0.5 \log \frac{0.5}{1-\theta}, \quad (4)$$

where $\mathsf{KL}(0.5 \,\|\, \theta)$ is the Kullback-Leibler (KL) divergence between $\mathsf{Bernoulli}(0.5)$ and $\mathsf{Bernoulli}(\theta)$. Here and below, we shall use $\log(\cdot)$ to indicate the natural logarithm.

We denote by d_v and d_{avg} the vertex degree of v and the average vertex degree of $\mathcal{G}$, respectively.

3. Main Results

This section describes two nearly linear-time algorithms and presents our main results. The proofs of all theorems can be found in (Chen et al., 2016).

3.1. Algorithms

3.1.1. Spectral-Expanding

The first algorithm, called Spectral-Expanding, consists of three stages. For concreteness, we start by describing the procedure when the measurement graphs are lines / rings; see Algorithm 1 for a precise description of the algorithm and Fig. 2 for a graphical illustration.

Stage 1: spectral metrod on a core subgraph. Consider a subgraph $\mathcal{G}_{\mathrm{c}}$ induced by $\mathcal{V}_{\mathrm{c}} := \{1, \cdots, r\}$, and it is self-evident that $\mathcal{G}_{\mathrm{c}}$ is a complete subgraph. We run a spectral method (e.g. (Chin et al., 2015)) on $\mathcal{G}_{\mathrm{c}}$ using samples taken over $\mathcal{G}_{\mathrm{c}}$, in the hope of obtaining approximate recovery of $\{X_i \mid i \in \mathcal{V}_{\mathrm{c}}\}$. Note that the spectral method can be replaced by other efficient algorithms, including semidefinite programming (SDP) (Javanmard et al., 2015) and a variant of belief propagation (BP) (Mossel et al., 2013).

Stage 2: progressive estimation of remaining vertices. For each vertex $i > |\mathcal{V}_{\mathrm{c}}|$, compute an estimate of X_i by majority vote using *backward samples*—those samples linking i and some $j < i$. The objective is to ensure that a large fraction of estimates obtained in this stage are accurate. As will be discussed later, the sample complexity required for approximate recovery is much lower than that required for exact recovery, and hence the task is feasible even though we do not use any forward samples to estimate X_i.

Stage 3: successive local refinement. Finally, we clean up all estimates using both backward and forward samples in order to maximize recovery accuracy. This is achieved by running local majority voting from the neighbors of each vertex until convergence. In contrast to many prior work, we reuse all samples in all iterations. As we shall see, this stage is the bottleneck for exact information recovery.

Remark 1. The proposed algorithm falls under the category of a general paradigm, which starts with an approximate estimate (often via spectral methods) followed by iterative refinement. This paradigm has been successfully applied to a wide spectrum of applications ranging from matrix completion (Keshavan et al., 2010a; Jain et al., 2013) to phase retrieval (Chen & Candes, 2015) to community recovery (Chaudhuri et al., 2012; Abbe et al., 2016).

An important feature of this algorithm is its low computational complexity. First of all, the spectral method can be performed within $O(m_{\mathrm{c}} \log n)$ time by means of the power method, where m_{c} indicates the number of samples falling on $\mathcal{G}_{\mathrm{c}}$. Stage 2 entails one round of majority voting, whereas the final stage—as we will demonstrate—converges within at most $O(\log n)$ rounds of majority voting. Note that each round of majority voting can be completed in linear time, i.e. in time proportional to reading all samples. Taken collectively, we see that Spectral-Expanding can be accomplished within $O(m \log n)$ flops, which is nearly linear time.

Careful readers will recognize that Stages 2-3 bear similarities with BP, and might wonder whether Stage 1 can also be replaced with standard BP. Unfortunately, we are not aware of any approach to analyze the performance of vanilla BP without a decent initial guess. Note, however, that the spectral method is already nearly linear-time, and is hence at least as fast as any feasible procedure.

While the preceding paradigm is presented for lines / rings, it easily extends to a much broader family of graphs with locality (see (Chen et al., 2016)). The only places that need

Algorithm 1 : Spectral-Expanding

1. **Run spectral method (see (Chen et al., 2016)) on a core subgraph** induced by $\mathcal{V}_{\rm c}$, which yields estimates $X_j^{(0)}, 1 \le j \le |\mathcal{V}_{\rm c}|$.
2. **Progressive estimation**: for $i = |\mathcal{V}_{\rm c}|+1, \cdots, n$,
$$X_i^{(0)} \leftarrow \mathsf{majority}\left\{Y_{i,j}^{(l)} \oplus X_j^{(0)} \mid j:\ j<i,\ (i,j)\in\mathcal{E},\ 1\le l\le N_{i,j}\right\}.$$
3. **Successive local refinement:** for $t = 0, \cdots, T-1$,
$$X_i^{(t+1)} \leftarrow \mathsf{majority}\left\{Y_{i,j}^{(l)} \oplus X_j^{(t)} \mid j:\ j\neq i,\ (i,j)\in\mathcal{E},\ 1\le l\le N_{i,j}\right\},\quad 1\le i\le n.$$
4. **Output** $X_i^{(T)}, 1 \le i \le n$.

Here, $\mathsf{majority}\{\cdot\}$ represents the majority voting rule: for any sequence $s_1, \cdots, s_k \in \{0,1\}$, $\mathsf{majority}\{s_1, \cdots, s_k\}$ is equal to 1 if $\sum_{i=1}^k s_i > k/2$; and 0 otherwise.

Algorithm 2 : Spectral-Stitching

1. **Split** all vertices into several (non-disjoint) vertex subsets each of size W as follows
$$\mathcal{V}_l := \{i \mid (i-1)\,W/2+1 \le l \le (i-1)\,W/2+W\},\quad l=1,2,\cdots,$$
and **run spectral method on each subgraph induced by** $\mathcal{V}_l$, which yields estimates $\{X_j^{\mathcal{V}_l} \mid j\in\mathcal{V}_l\}$ for each $l \ge 1$.
2. **Stitching**: set $X_j^{(0)} \leftarrow X_j^{\mathcal{V}_1}$ for all $j \in \mathcal{V}_1$; for $l = 2, 3, \cdots,$
$$X_j^{(0)} \leftarrow X_j^{\mathcal{V}_l}\ (\forall j\in\mathcal{V}_l)\quad \text{if } \sum\nolimits_{j\in\mathcal{V}_l\cap\mathcal{V}_{l-1}} X_j^{\mathcal{V}_l}\oplus X_j^{\mathcal{V}_{l-1}} \le 0.5\,|\mathcal{V}_l\cap\mathcal{V}_{l-1}|\,;$$
$$\text{and}\quad X_j^{(0)} \leftarrow X_j^{\mathcal{V}_l}\oplus 1\ (\forall j\in\mathcal{V}_l)\quad \text{otherwise.}$$
3. **Successive local refinement** and output $X_i^{(T)}, 1 \le i \le n$ (see Steps 3-4 of Algorithm 1).

to be adjusted are:

(1) **The core subgraph** $\mathcal{V}_{\rm c}$. One would like to ensure that $|\mathcal{V}_{\rm c}| \gtrsim d_{\rm avg}$ and that the subgraph $\mathcal{G}_{\rm c}$ induced by $\mathcal{V}_{\rm c}$ forms a (nearly) complete subgraph, in order to guarantee decent recovery in Stage 1.

(2) **The ordering of the vertices**. Let $\mathcal{V}_{\rm c}$ form the first $|\mathcal{V}_{\rm c}|$ vertices of $\mathcal{V}$, and make sure that each $i > |\mathcal{V}_{\rm c}|$ is connected to at least an order of $d_{\rm avg}$ vertices in $\{1, \cdots, i-1\}$. This is important because each vertex needs to be incident to sufficiently many backward samples in order for Stage 2 to be successful.

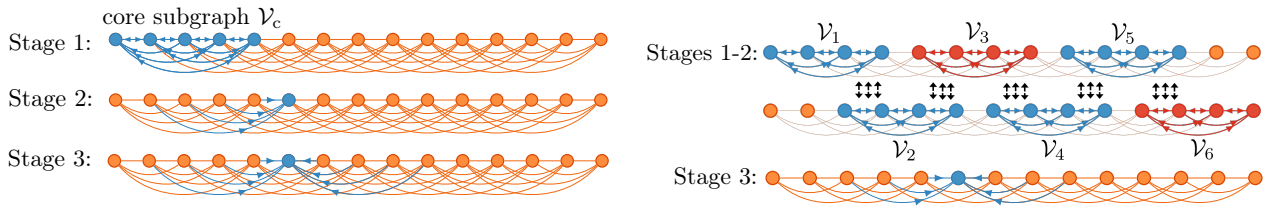
3.1.2. SPECTRAL-STITCHING

We now turn to the 2nd algorithm called Spectral-Stitching, which shares similar spirit as Spectral-Expanding and, in fact, differs from Spectral-Expanding only in Stages 1-2.

Stage 1: node splitting and spectral estimation. Split $\mathcal{V}$ into several overlapping subsets $\mathcal{V}_l$ ($l \ge 1$) of size W, such that any two adjacent subsets share $W/2$ common vertices. We choose the size W of each $\mathcal{V}_l$ to be r for rings / lines, and on the order of $d_{\rm avg}$ for other graphs. We then run spectral methods separately on each subgraph $\mathcal{G}_l$ induced by $\mathcal{V}_l$, in the hope of achieving approximate estimates $\{X_i^{\mathcal{V}_l} \mid i\in\mathcal{V}_l\}$—up to global phase—for each subgraph.

Stage 2: stiching the estimates. The aim of this stage is to stitch together the outputs of Stage 1 computed in isolation for the collection of overlapping subgraphs. If approximate recovery (up to some global phase) has been achieved in Stage 1 for each $\mathcal{V}_l$, then the outputs for any two adjacent subsets are positively correlated only when they have matching global phases. This simple observation allows us to calibrate the global phases for all preceding estimates, thus yielding a vector $\{X_i^{(0)}\}_{1\le i\le n}$ that is approximately faithful to the truth modulo some global phase.

The remaining steps of Spectral-Stitching follow the same local refinement procedure as in Spectral-Expanding; see Algorithm 2. As can be seen, the first 2 stages of Spectral-Stitching—which can also be completed in nearly lin-



(a) Spectral-Expanding (b) Spectral-Stitching

Figure 2. Illustration of the information flow in Spectral-Expanding and Spectral-Stitching.

ear time—are more "symmetric" than those of Spectral-Expanding. More precisrely, Spectral-Expanding emphasizes a single core subgraph $\mathcal{G}_c$ and computes all other estimates based on $\mathcal{G}_c$, while Spectral-Stitching treats each subgraph $\mathcal{G}_l$ almost equivalently. This symmetry nature might be practically beneficial when the acquired data deviate from our assumed random sampling model.

3.2. Theoretical Guarantees: Rings

We start with the performance of our algorithms for rings. This class of graphs—which is spatially invariant—is arguably the simplest model exhibiting locality structure.

3.2.1. Minimum Sample Complexity

Encouragingly, the proposed algorithms succeed in achieving the minimum sample complexity, as stated below.

Theorem 1. *Fix $\theta > 0$ and any small $\epsilon > 0$. Let $\mathcal{G}$ be a ring $\mathcal{R}_r$ with locality radius r, and suppose*

$$m \geq (1+\epsilon)\, m^*, \tag{5}$$

where

$$m^* = \frac{n \log n}{2\left(1 - e^{-\mathsf{KL}(0.5\|\theta)}\right)}. \tag{6}$$

Then with probability approaching one[3]*, Spectral-Expanding (resp. Spectral-Stitching) converges to the ground truth within $T = O(\log n)$ iterations, provided that $r \gtrsim \log^3 n$ (resp. $r \geq n^\delta$ for an arbitrary constant $\delta > 0$).*

Conversely, if $m < (1-\epsilon)\, m^*$, then the probability of error $P_e(\psi)$ is approaching one for any algorithm ψ.

Remark 2. When $r = n - 1$, a ring reduces to a complete graph (or an equivalent Erdős-Rényi model). For this case, computationally feasible algorithms have been extensively studied (Swamy, 2004; Jalali et al., 2011; Chen et al., 2014c;b;a), most of which focus only on the scaling results. Recent work (Hajek et al., 2015a; Jog & Loh, 2015) succeeded in characterizing the sharp threshold for this case, and it is immediate to check that the sample complexity we derive in (6) matches the one presented in (Hajek et al., 2015a; Jog & Loh, 2015).

[3]More precisely, the proposed algorithms succeed with probability exceeding $1 - c_1 r^{-9} - C_2 \exp\{-c_2 \frac{m}{n}(1 - e^{-D^*})\}$ for some constants $c_1, c_2, C_2 > 0$.

Remark 3. Theorem 1 requires $r \gtrsim \mathsf{poly}\log(n)$ because each node needs to be connected to sufficiently many neighbors in order to preclude "bursty" errors. The condition $r \gtrsim \log^3 n$ might be improved to a lower-order $\mathsf{poly}\log(n)$ term using more refined analyses. When $r \lesssim \log n$, one can compute the maximum likelihood (ML) estimate via dynamic programming (Kamath et al., 2015) within polynomial time.

Theorem 1 uncovers a surprising insensitivity phenomenon for rings: as long as the measurement graph is sufficiently connected, the locality constraint does not alter the sample complexity limit and the computational limit at all. This subsumes as special cases two regimes that exhibit dramatically different graph structures: (1) complete graphs, where the samples are taken in a global manner, and (2) rings with $r = O(\mathsf{poly}\log(n))$, where the samples are constrained within highly local neighborhood.

Notably, both (Abbe et al., 2015) and (Hajek et al., 2015b) have derived general sufficient recovery conditions of SDP which, however, depend on the second-order graphical metrics of $\mathcal{G}$ (Durrett, 2007) (e.g. the spectral gap or Cheeger constant). When applied to rings (or other graphs with locality), the sufficient sample complexity given therein is significantly larger than the information limit[4]. This is in contrast to our finding, which reveals that for many graphs with locality, both the information and computation limits often depend only upon the vertex degrees independent of these second-order graphical metrics.

3.2.2. Bottlenecks for Exact Recovery

Before explaining the rationale of the proposed algorithms, we provide here some heuristic argument as to why $n \log n$ samples are necessary for exact recovery and where the recovery bottleneck lies.

Without loss of generality, assume $\boldsymbol{X} = [0, \cdots, 0]^\top$. Suppose the genie tells us the correct labels of all nodes except v. Then all samples useful for recovering X_v reside on the edges connecting v and its neighbors, and there are $\mathsf{Poisson}(\lambda d_v)$ such samples. Thus, this comes down to

[4]For instance, the sufficient sample complexity given in (Abbe et al., 2015) scales as $\frac{n \log n}{h_{\mathcal{G}} D^*}$ with $h_{\mathcal{G}}$ denoting the Cheeger constant. Since $h_{\mathcal{G}} = O(1/n)$ for rings / lines, this results in a sample size that is about n times larger than the information limit.

testing between two conditionally i.i.d. distributions with a Poisson sample size of mean λd_v. From the large deviation theory, the ML rule fails in recovering X_v with probability

$$P_{\mathrm{e},v} \approx \exp\left\{-\lambda d_v (1 - e^{-D^*})\right\}, \tag{7}$$

where D^* is the large deviation exponent. The above argument concerns a typical error event for recovering a single node v, and it remains to accommodate all vertices. Since the local neighborhoods of two vertices v and u are nearly non-overlapping, the resulting typical error events for recovering X_v and X_u become almost independent and disjoint. As a result, the probability of error of the ML rule ψ_{ml} is approximately lower bounded by

$$P_{\mathrm{e}}(\psi_{\mathrm{ml}}) \gtrsim \sum\nolimits_{v=1}^{n} P_{\mathrm{e},v} \approx n \exp\left\{-\lambda d_{\mathrm{avg}} (1 - e^{-D^*})\right\}, \tag{8}$$

where one uses the fact that $d_v \equiv d_{\mathrm{avg}}$. Apparently, the right-hand side of (8) would vanish only if

$$\lambda d_{\mathrm{avg}} (1 - e^{-D^*}) > \log n. \tag{9}$$

Since the total sample size is $m = \lambda \cdot \frac{1}{2} n d_{\mathrm{avg}}$, this together with (9) confirms the sample complexity lower bound

$$m = \frac{1}{2} \lambda n d_{\mathrm{avg}} > \frac{n \log n}{2\,(1 - e^{-D^*})} = m^*.$$

As we shall see, the above error events—in which only a single variable is uncertain—dictate the hardness of exact recovery.

3.2.3. Interpretation of Our Algorithms

The preceding argument suggests that the recovery bottleneck of an optimal algorithm should also be determined by the aforementioned typical error events. This is the case for both Spectral-Expanding and Spectral-Stitching, as revealed by the intuitive arguments below. While the intuition is provided for rings, it contains all important ingredients that apply to many other graphs.

To begin with, we provide an heuristic argument for Spectral-Expanding.

(i) Stage 1 focuses on a core complete subgraph $\mathcal{G}_{\mathrm{c}}$. In the regime where $m \gtrsim n \log n$, the total number of samples falling within $\mathcal{G}_{\mathrm{c}}$ is on the order of $\frac{|\mathcal{V}_{\mathrm{c}}|}{n} \cdot m \geq |\mathcal{V}_{\mathrm{c}}| \log n$, which suffices in guaranteeing partial recovery using spectral methods (Chin et al., 2015). In fact, the sample size we have available over $\mathcal{G}_{\mathrm{c}}$ is way above the degrees of freedom of the variables in $\mathcal{G}_{\mathrm{c}}$ (which is r).

(ii) With decent initial estimates for $\mathcal{G}_{\mathrm{c}}$ in place, one can infer the remaining pool of vertices one by one using existing estimates together with backward samples. One important observation is that each vertex is incident to many—i.e. about the order of $\log n$—backward samples. That said, we are effectively operating in a high signal-to-noise ratio (SNR) regime. While existing estimates are imperfect, the errors occur only to a small fraction of vertices. Moreover, these errors are in some sense randomly distributed and hence fairly spread out, thus precluding the possibility of bursty errors. Consequently, one can obtain correct estimate for each of these vertices with high probability, leading to a vanishing fraction of errors in total.

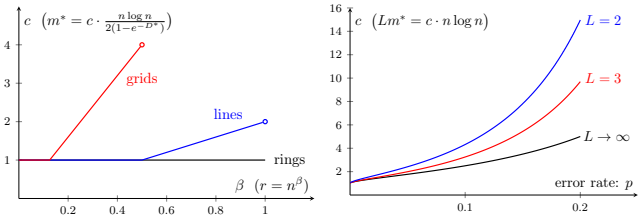


Figure 3. (Left) Minimum sample complexity m^* vs. locality radius r; (Right) Minimum number Lm^* of vertices being measured (including repetition) vs. single-vertex error rate p.

(iii) Now that we have achieved approximate recovery, all remaining errors can be cleaned up via local refinement using all backward and forward samples. For each vertex, since only a vanishingly small fraction of its neighbors contain errors, the performance of local refinement is almost the same as in the case where all neighbors have been perfectly recovered.

The above intuition extends to Spectral-Stitching. Following the argument in (i), we see that the spectral method returns nearly accurate estimates for each of the subgraph $\mathcal{G}_{\mathrm{l}}$ induced by $\mathcal{V}_l$, except for the global phases. Since any two adjacent $\mathcal{G}_{\mathrm{l}}$ and $\mathcal{G}_{\mathrm{l}+1}$ have sufficient overlaps, this allows us to calibrate the global phases for $\{X_i^{\mathcal{V}_l} : i \in \mathcal{V}_l\}$ and $\{X_i^{\mathcal{V}_{l+1}} : i \in \mathcal{V}_{l+1}\}$. Once we obtain approximate recovery for all variables simultaneously, the remaining errors can then be cleaned up by Stage 3 as in Spectral-Expanding.

We emphasize that the first two stages of both algorithms—which aim at approximate recovery—require only $O(n)$ samples (as long as the pre-constant is sufficiently large). In contrast, the final stage is the bottleneck: it succeeds as long as local refinement for each vertex is successful. The error events for this stage are almost equivalent to the typical events singled out in Section 3.2.2, justifying the information-theoretic optimality of both algorithms.

3.3. Theoretical Guarantees: Inhomogeneous Graphs

The proposed algorithms are guaranteed to succeed for a much broader class of graphs with locality beyond rings, including those that exhibit inhomogeneous vertex degrees. The following theorem formalizes this claim for two of the most important instances: lines and grids.

Theorem 2. *Theorem 1 continues to hold for the following families of measurement graphs:*

(1) Lines with $r = n^\beta$ for some constant $0 < \beta < 1$, where

$$m^* = \frac{\max\{1/2,\ \beta\}\, n \log n}{1 - e^{-\mathsf{KL}(0.5\,\|\,\theta)}}; \tag{10}$$

(2) Grids with $r = n^\beta$ for some constant $0 < \beta < 0.5$, where

$$m^* = \frac{\max\{1/2,\ 4\beta\}\, n \log n}{1 - e^{-\mathsf{KL}(0.5\,\|\,\theta)}}. \tag{11}$$

Remark 4. Careful readers will note that for lines (resp. grids), m^* does not converge to $\frac{n \log n}{2(1-e^{-\mathsf{KL}(0.5\|\theta)})}$ as $\beta \to 1$ (resp. $\beta \to 0.5$), which is the case of complete graphs. This arises because m^* experiences a more rapid drop in the regime where $\beta = 1$ (resp. $\beta = 0.5$). For instance, for a line with $r = \gamma n$ for some constant $\gamma > 0$, one has $m^* = \frac{(1-\gamma/2) n \log n}{1-e^{-\mathsf{KL}(0.5\|\theta)}}$. In addition, the result extends to small-world graphs. See (Chen et al., 2016) for details.

Theorem 2 characterizes the effect of locality radius upon the sample complexity limit; see Fig. 3 for a comparison of three classes of graphs. In contrast to rings, lines and grids are spatially varying models due to the presence of boundary vertices, and the degree of graph inhomogeneity increases in the locality radius r. To be more concrete, consider, for example, the first $d_{\text{avg}}/\log n$ vertices of a line, which have degrees around $d_{\text{avg}}/2$. In comparison, the set of vertices lying away from the boundary have degrees as large as d_{avg}. This tells us that the first few vertices form a weakly connected component, thus presenting an additional bottleneck for exact recovery. This issue is negligible unless the size of the weakly connected component is exceedingly large. As asserted by Theorem 2, the minimum sample complexity for lines (resp. grids) is identical to that for rings unless $r \gtrsim \sqrt{n}$ (resp. $r \gtrsim n^{1/8}$). Note that the curves for lines and grids (Fig. 3) have distinct hinge points primarily because the vertex degrees of the corresponding weakly connected components differ.

More precisely, the insights developed in Section 3.2.2 readily carry over here. Since the error probability of the ML rule is lower bounded by (8), everything boils down to determining the smallest λ (called λ^*) satisfying

$$\sum\nolimits_{v=1}^{n} \exp\left\{-\lambda^* d_v \left(1 - e^{-D^*}\right)\right\} \to 0,$$

which in turn yields $m^* = \frac{1}{2}\lambda^* d_{\text{avg}} n$. The two cases accommodated by Theorem 2 can all be derived in this way.

3.4. Connection to Low-Rank Matrix Completion

One can aggregate all correct parities into a matrix $\boldsymbol{Z} = [Z_{i,j}]_{1\le i,j\le n}$ such that $Z_{i,j} = 1$ if $X_i = X_j$ and $Z_{i,j} = -1$ otherwise. It is straightforward to verify that $\operatorname{rank}(\boldsymbol{Z}) = 1$, with each $Y_{i,j}^{(l)}$ being a noisy measurement of $Z_{i,j}$. Thus, our problem falls under the category of low-rank matrix completion, a topic that has inspired a flurry of research (e.g. (Candes & Recht, 2009; Keshavan et al., 2010b; Candès et al., 2011; Chandrasekaran et al., 2011; Chen et al., 2013)). Most prior works, however, concentrated on samples taken over an Erdős–Rnyi model, without investigating sampling schemes with locality constraints. One exception is (Bhojanapalli & Jain, 2014), which explored the effectiveness of SDP under general sampling schemes. However, the sample complexity required therein increases significantly as the spectral gap of the measurement graph drops, which does not deliver optimal guarantees. We believe that the approach developed herein will shed light on solving general matrix completion problems from samples with locality.

4. Extension: Beyond Pairwise Measurements

The proposed algorithms are applicable to numerous scenarios beyond the basic setup in Section 2.1. This section presents one important extension.

In some applications, each measurement may cover more than two nodes in the graph. In the haplotype phasing application, for example, a new sequencing technology called 10X (10x, 2016) generates barcodes to mark reads from the same chromosome (maternal or paternal), and more than two reads can have the same barcode. For concreteness, we suppose the locality constraint is captured by rings, and consider the type of multiple linked samples as follows.

Measurement (hyper)-graphs. Let $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ be a ring $\mathcal{R}_r$, and let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hypergraph such that (i) every hyperedge is incident to L vertices in $\mathcal{V}$, and (ii) all these L vertices are mutually connected in $\mathcal{G}_0$.

Noise model. On each hyperedge $e = (i_1, \cdots, i_L) \in \mathcal{G}$, we obtain $N_e \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda)$ multi-linked samples $\{Y_e^{(l)} \mid 1 \le l \le N_e\}$. Conditional on N_e, each sample $Y_e^{(l)}$ is an independent copy of

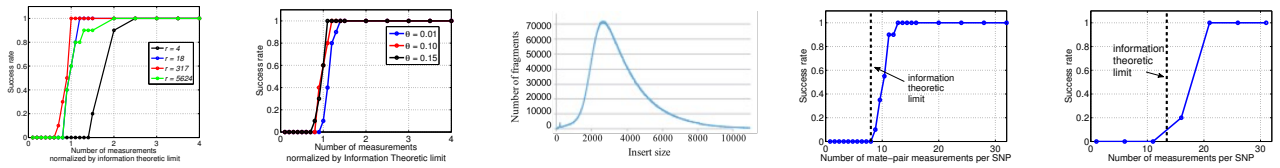
$$Y_e = \begin{cases} (Z_{i_1}, \cdots, Z_{i_L}), & \text{with prob. } 0.5, \\ (Z_{i_1} \oplus 1, \cdots, Z_{i_L} \oplus 1), & \text{else}, \end{cases} \tag{12}$$

where Z_i is a noisy measurement of X_i such that $Z_i = X_i$ with probability $1 - p$ and $Z_i = X_i \oplus 1$ otherwise. Here, p represents the error rate for measuring a single vertex. For the pairwise samples considered before, one can think of the parity error rate θ as $\mathbb{P}\{Z_i \oplus Z_j \neq X_i \oplus X_j\}$ or, equivalently, $\theta = 2p(1-p)$.

We emphasize that a random global phase is incorporated into each sample (12). That being said, each sample reveals only the *relative* similarity information among these L vertices, without providing further information about the absolute cluster membership. Interestingly, the proposed algorithms with slight modification (see (Chen et al., 2016)) are still information-theoretically optimal.

Theorem 3. *Fix $L \ge 2$. Theorem 1 continues to hold under the above L-wise sampling model, with m^* replaced by*

$$m^* := \frac{n \log n}{L\left(1 - e^{-D(P_0, P_1)}\right)}.$$



(a) (b) (c) (d) (e)

Figure 4. Empirical success rate of Spectral-Expanding for: (a) Rings $\mathcal{R}_r$; (b) Rings $\mathcal{R}_{18}$ with varied measurement error rate θ; (c) Insert size distribution (Illumina, 2012); (d) Performance on a simulation of haplotype; (e) Performance on a simulation of haplotype. For (a) and (b), the x-axis is the sample size m normalized by the information limit m^*.

	$r = n^{0.25}$	$r = n^{0.5}$	$r = n^{0.75}$
Time (seconds/run)	3.55	6.45	58.4

Table 1. The time taken to run Spectral-Expanding on a MacBook Pro equipped with a 2.9 GHz Intel Core i5 and 8GB of memory over rings $\mathcal{R}_r$, where $n = 100,000$, $\theta = 10\%$ and $m = 1.5m^*$.

Here,

$$P_0 = (1-p)\mathsf{Binomial}\,(L-1,p) + p\mathsf{Binomial}\,(L-1,1-p)\,;$$
$$P_1 = p\mathsf{Binomial}\,(L-1,p) + (1-p)\,\mathsf{Binomial}\,(L-1,1-p)\,.$$

Remark 5. A closed-form expression of $D(P_0, P_1)$ can be found in (Chen et al., 2016).

With Theorem 3 in place, we can determine the benefits of multi-linked sampling. To enable a fair comparison, we evaluate the sampling efficiency in terms of Lm^* rather than m^*, since Lm^* captures the total number of vertices (including repetition) one needs to measure. As illustrated in Fig. 3, the sampling efficiency improves as L increases, but there exists a fundamental lower barrier given by $\frac{n\log n}{1-e^{-\mathsf{KL}(0.5\|p)}}$. This lower barrier, as plotted in the black curve of Fig. 3, corresponds to the case where L is approaching infinity.

5. Numerical Experiments

To verify the practical applicability of the proposed algorithms, we have conducted simulations in various settings. All these experiments focused on graphs with $n = 100,000$ vertices, and used an error rate of $\theta = 10\%$ unless otherwise noted. For each point, the empirical success rates averaged over 10 Monte Carlo runs are reported.

(a) *Regular rings.* We ran Algorithm 1 on rings $\mathcal{R}_r$ for various values of locality radius r (Fig. 4(a)), with the runtime reported in Table 1;

(b) *Rings with different error rates.* We varied the error rate θ for rings with $r = 18 = n^{0.25}$, and plotted the empirical success rate (Fig. 4(d)).

We have also simulated a model of the haplotype phasing problem by assuming that the genome has a SNP periodically every 1000 base pairs. The insert length distribution, i.e. the distribution of the genomic distance between the linking reads, is given in Fig. 4(c) for Illumina reads, and a draw from $\mathsf{Poisson}(3.5)$ truncated within the interval $1,\cdots,9$ is a reasonable approximation for the number of SNPs between two measured SNPs. We then ran the simulation on the rings $\mathcal{R}_9$, with non-uniform sampling weight. Using the nominal error rate of $p = 1\%$ for the short reads, the error rates of the measurements is $2p(1-p) \approx 2\%$. The empirical performance is shown in Fig. 4(d).

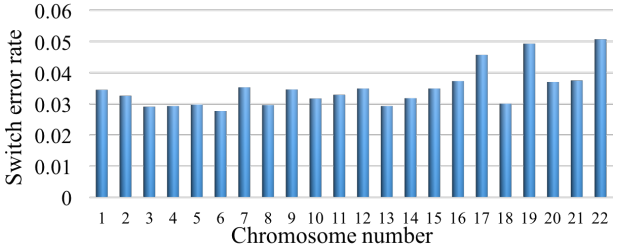


Figure 5. The switch error rates of Spectral-Stitching when run on the NA12878 data-set from 10x-genomics (10x Genomics, 2015).

Additionally, we have simulated reads generated by 10x-Genomics (10x, 2016) , which corresponds to the model in Section 4. Each measurement consists of multiple linked reads, which is generated by first randomly picking a segment of length 100 SNPs (called a *fragment*) on the line graph and then generating $\mathsf{Poisson}(9)$ number of linked reads uniformly located in this segment. The noise rate per read is $p = 0.01$. The empirical result is shown in Fig. 4(e). The information theoretic limit is calculated using Theorem 3, with L set to infinity (since the number of vertices involved in a measurement is quite large here).

To evaluate the performance of our algorithm on real data, we ran Spectral-Stitching for Chromosomes 1-22 on the NA12878 data-set made available by 10x-Genomics (10x Genomics, 2015). The nominal error rate per read is $p = 1\%$, and the average number of SNPs touched by each sample is $L \in [6, 7]$. The number of SNPs n ranges from 34240 to 191829, with the sample size m from 102633 to 574189. Here, we split all vertices into overlapping subsets of size $W = 100$. The performance is measured in terms of the *switch error rate*, that is, the fraction of positions where we need to switch the estimate to match the ground truth. The performance on Chromosomes 1-22 is reported in Fig. 5.

References

10x Genomics, 2016. URL http://www.10xgenomics.com. [Online; accessed 5-February-2016].

10x Genomics. NA12878 Loupe data-set, 2015. URL http://software.10xgenomics.com/files/samples/genome/NA12878_WGS.

Abbe, E. and Sandon, C. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

Abbe, E., Bandeira, A., Bracher, A., and Singer, A. Decoding binary node labels from censored measurements: Phase transition and efficient recovery. *IEEE Trans on Network Science and Engineering*, 1(1):10–22, 2015.

Abbe, Emmanuel and Wainwright, Martin. ISIT 2015 Tutorial: Information Theory and Machine Learning. 2015.

Abbe, Emmanuel, Bandeira, Afonso S, and Hall, Georgina. Exact recovery in the stochastic block model. *IEEE Trans. on Information Theory*, 62(1):471–487, 2016.

Bansal, N., Blum, A., and Chawla, S. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

Bhojanapalli, Srinadh and Jain, Prateek. Universal matrix completion. *International Conference on Machine Learning (ICML)*, pp. 1881–1889, 2014.

Browning, S. and Browning, B. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

Candes, E. J. and Recht, B. Exact Matrix Completion via Convex Optimization. *Foundations of Comp. Math.*, (6): 717–772, 2009.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of ACM*, 58(3):11:1–11:37, Jun 2011.

Chandrasekaran, Venkat, Sanghavi, Sujay, Parrilo, Pablo A, and Willsky, Alan S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chaudhuri, K., Chung, F., and Tsiatas, A. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012:1–23, 2012.

Chen, Jingchun and Yuan, Bo. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

Chen, Y. and Candes, E. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 739–747, 2015.

Chen, Y., Guibas, L., and Huang, Q. Near-Optimal Joint Object Matching via Convex Relaxation. *International Conference on Machine Learning (ICML)*, pp. 100–108, June 2014a.

Chen, Y., Suh, C., and Goldsmith, A. J. Information recovery from pairwise measurements: A Shannon-theoretic approach. *IEEE International Symposium on Information Theory, arXiv preprint arXiv:1504.01369*, 2015.

Chen, Yudong, Jalali, A., Sanghavi, S., and Caramanis, C. Low-Rank Matrix Recovery From Errors and Erasures. *IEEE Trans on Info Theory*, 59(7):4324–4337, 2013.

Chen, Yudong, Lim, Shiau H, and Xu, Huan. Weighted graph clustering with non-uniform uncertainties. In *International Conference on Machine Learning (ICML)*, pp. 1566–1574, 2014b.

Chen, Yudong, Sanghavi, Sujay, and Xu, Huan. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014c.

Chen, Yuxin, Kamath, Govinda, Suh, Changho, and Tse, David. Community recovery in graphs with locality (full version). 2016. URL http://arxiv.org/abs/1602.03828.

Chin, Peter, Rao, Anup, and Vu, Van. Stochastic Block Model and Community Detection in the Sparse Graphs: A spectral algorithm with optimal rate of recovery. *arXiv preprint arXiv:1501.05021*, 2015.

Coja-Oghlan, Amin. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

Condon, Anne and Karp, Richard M. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. Wiley-interscience, 2006.

Das, Shreepriya and Vikalo, Haris. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC genomics*, 16(1):1, 2015.

Donmez, N. and Brudno, M. Hapsembler: an assembler for highly polymorphic genomes. In *Research in Computational Molecular Biology*, pp. 38–52. Springer, 2011.

Durrett, R. *Random graph dynamics*, volume 200. Cambridge university press Cambridge, 2007.

Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

Girvan, M. and Newman, M. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

Globerson, Amir, Roughgarden, Tim, Sontag, David, and Yildirim, Cafer. How Hard is Inference for Structured Prediction? In *ICML*, pp. 2181–2190, 2015.

Hajek, Bruce, Wu, Yihong, and Xu, Jiaming. Achieving Exact Cluster Recovery Threshold via Semidefinite Programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015a.

Hajek, Bruce, Wu, Yihong, and Xu, Jiaming. Exact recovery threshold in the binary censored block model. In *Information Theory Workshop*, pp. 99–103, 2015b.

Holland, Paul W, Laskey, Kathryn Blackmond, and Leinhardt, Samuel. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Illumina. Data processing of Nextera mate pair reads on Illumina sequencing platform. *Technical Note: Sequencing*, 2012. URL http://www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Symposium on Theory of Computing (STOC)*, pp. 665–674, 2013.

Jalali, Ali, Chen, Yudong, Sanghavi, Sujay, and Xu, Huan. Clustering Partially Observed Graphs via Convex Optimization. In *International Conference on Machine Learning (ICML)*, pp. 1001–1008, June 2011.

Javanmard, Adel, Montanari, Andrea, and Ricci-Tersenghi, Federico. Phase Transitions in Semidefinite Relaxations. *arXiv preprint arXiv:1511.08769*, 2015.

Jog, Varun and Loh, Po-Ling. Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

Kamath, G., Sasoglu, E., and Tse, D. Optimal Haplotype Assembly from High-Throughput Mate-Pair Reads. *IEEE International Symposium on Information Theory*, pp. 914–918, June 2015.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 99:2057–2078, 2010a.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE Trans on Info Theory*, (6):2980–2998, 2010b.

Mossel, Elchanan, Neeman, Joe, and Sly, Allan. Belief propagation, robust reconstruction, and optimal recovery of block models. *arXiv preprint arXiv:1309.1380*, 2013.

Porter, Mason A, Onnela, Jukka-Pekka, and Mucha, Peter J. Communities in networks. *Notices of the AMS*, 56 (9):1082–1097, 2009.

Shi, Jianbo and Malik, Jitendra. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Si, H., Vikalo, H., and Vishwanath, S. Haplotype assembly: An information theoretic view. In *IEEE Information Theory Workshop*, pp. 182–186, 2014.

Swamy, C. Correlation clustering: maximizing agreements via semidefinite programming. In *Symposium on Discrete Algorithms (SODA)*, pp. 526–527, 2004.