Consistent recovery threshold of hidden nearest neighbor graphs

Jian Ding DINGJIAN@WHARTON.UPENN.EDU
Department of Statistics, The Wharton School, University of Pennsylvania

Yihong Wu YIHONG.WU@YALE.EDU
Department of Statistics and Data Science, Yale University

Jiaming Xu JIAMING.XU868@DUKE.EDU
The Fuqua School of Business, Duke University

Dana Yang XIAOQIAN.YANG@DUKE.EDU
The Fuqua School of Business, Duke University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Motivated by applications such as discovering strong ties in social networks and assembling genome subsequences in biology, we study the problem of recovering a hidden $2k$-nearest neighbor (NN) graph in an n-vertex complete graph, whose edge weights are independent and distributed according to P_n for edges in the hidden $2k$-NN graph and Q_n otherwise. The special case of Bernoulli distributions corresponds to a variant of the Watts-Strogatz small-world graph. We focus on two types of asymptotic recovery guarantees as $n \to \infty$: (1) exact recovery: all edges are classified correctly with probability tending to one; (2) almost exact recovery: the expected number of misclassified edges is $o(nk)$. We show that the maximum likelihood estimator achieves (1) exact recovery for $2 \le k \le n^{o(1)}$ if $\liminf \frac{2\alpha_n}{\log n} > 1$; (2) almost exact recovery for $1 \le k \le o\left(\frac{\log n}{\log\log n}\right)$ if $\liminf \frac{kD(P_n||Q_n)}{\log n} > 1$, where $\alpha_n \triangleq -2\log\int\sqrt{dP_n dQ_n}$ is the Rényi divergence of order $\frac{1}{2}$ and $D(P_n||Q_n)$ is the Kullback-Leibler divergence. Under mild distributional assumptions, these conditions are shown to be information-theoretically necessary for any algorithm to succeed. A key challenge in the analysis is the enumeration of $2k$-NN graphs that differ from the hidden one by a given number of edges. We also analyze several computationally efficient algorithms and provide sufficient conditions under which they achieve exact/almost exact recovery. In particular, we develop a polynomial-time algorithm that attains the threshold for exact recovery under the small-world model. [1]

Keywords: Nearest neighbor graphs, small-world graphs, Information-theoretic lower bounds

1. Introduction

The strong and weak ties are essential for information diffusion, social cohesion, and community organization in social networks (Granovetter, 1977). The strong ties between close friends are responsible for forming tightly-knit groups, while the weak ties between acquaintances are crucial for binding groups of strong ties together (Easley and Kleinberg, 2010). The celebrated Watts-Strogatz small-world graph (Watts and Strogatz, 1998) is a simple network model that exhibits both strong and weak ties. It posits that n nodes are located on a ring and starts with a $2k$-nearest neighbor (NN) graph of strong ties, where each node is connected to its $2k$ nearest neighbors (k on the left

[1] Extended abstract. Full version appears as [arXiv reference, v1911.08004].

and k on the right) on the ring. Then to generate weak ties, for every node, each of its strong ties is rewired with probability ϵ to a node chosen uniformly at random. As ϵ varies from 0 to 1, the graph interpolates between a ring lattice and an Erdős-Rényi random graph; for intermediate values of ϵ, the graph is a small-world network: highly clustered with many triangles, yet with a small diameter.

The Watts-Strogatz small-world graph and its variants, albeit simple, have been extensively studied and widely used in various disciplines to model real networks beyond social networks, such as academic collaboration network (Newman, 2001), metabolic networks (Wagner and Fell, 2001), brain networks (Bassett and Bullmore, 2006), and word co-occurrence networks in language modeling (Cancho and Solé, 2001; Motter et al., 2002). Most of the previous work focuses on studying the structures of small-world graphs (Newman and Watts, 1999) and their algorithmic consequences (Kleinberg, 2000; Moore and Newman, 2000; Saramäki and Kaski, 2005). However, in many practical applications, it is also of interest to distinguish strong ties from weak ones (Marsden and Campbell, 1984; Gilbert and Karahalios, 2009; Gilbert, 2012; Rotabi et al., 2017). For example, in Facebook (Marlow et al., 2009) or Twitter network (Huberman et al., 2008), identifying the close ties among a user's potentially hundreds of friends provides valuable information for marketing and ad placements. Even when additional link attribute information (such as the communication time in who-talks-to-whom networks (Onnela et al., 2007)) are available to be used to measure the strength of the tie, the task of discovering strong ties could still be challenging, as the link attributes are potentially noisy or only partially observed, obscuring the inherent tie strength. Therefore, it is of fundamental importance, in both theory and practice, to understand when and how we can infer strong ties from the noisy and partially observed network data. In this paper, we address this question in the following statistical model:

Definition 1 (Hidden $2k$-NN graph recovery)
Given: *$n \geq 1$, and two distributions P_n and Q_n, parametrized by n.*
Observation: *A randomly weighted, undirected complete graph w with a hidden $2k$-NN graph x^* on n vertices, such that the edge weights are independent, and for each edge e, the edge weight w_e is distributed as P_n if e is an edge in x^* and as Q_n otherwise.*
Inference Problem: *Recover the hidden $2k$-NN graph x^* from the observed random graph.*

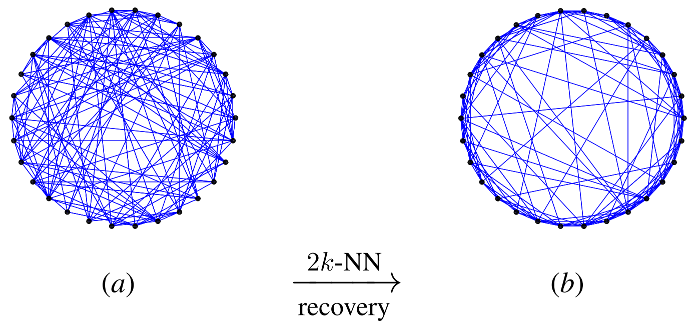

Figure 1: Left: An observed graph generated by the hidden $2k$-NN graph model with $n = 30$ vertices, $k = 4$, $P_n = \mathrm{Bern}(0.8)$, and $Q_n = \mathrm{Bern}(0.09)$; Right: the observed graph with vertices rearranged according to the latent $2k$-NN graph.

Note that every $2k$-NN graph x can be described by a permutation σ on $[n]$ as follows: first, construct a Hamiltonian cycle $(\sigma(1), \sigma(2), \ldots, \sigma(n), \sigma(1))$, then connect pairs of vertices that are at distance at most k on the cycle (see Fig. 1 for graphical illustrations).

The $2k$-NN model encompasses partially observed networks as a special case. This can be accomplished by considering $P_n = \epsilon\delta_* + (1-\epsilon)P_n'$ and $Q_n = \epsilon\delta_* + (1-\epsilon)Q_n'$ where $*$ is a special symbol outside of the support of P_n' and Q_n' indicating those edge weights that are unobserved. When P_n and Q_n are Bernoulli distributions with corresponding success probabilities $p_n > q_n$, we arrive at a variant of the Watts-Strogatz small-world graph.

The problem of recovering a hidden NN graph is also motivated by *de novo genome assembly*, the reconstruction of an organism's long sequence of A, G, C, T nucleotides from fragmented sequencing data. The previous work (Bagaria et al., 2020) casts genome scaffolding as a hidden Hamiltonian cycle recovery problem, which is a special case of our model with $k = 1$. By considering $k > 1$, the general $2k$-NN graph model is a closer approximation to the real data. See the arXiv version for a more detailed discussion of this application.

Note that in the aforementioned applications we often have $k \ll n$; thus in this paper we focus on the regime of $k = n^{o(1)}$ and study the following two types of recovery guarantees. Let $x^* \in \{0,1\}^{\binom{n}{2}}$ denote the adjacency vector of the hidden $2k$-NN graph, where $x_e^* = 1$ for every edge e in the hidden $2k$-NN graph and $x_e^* = 0$ otherwise. Let $\mathcal{X}$ denote the collection of adjacency vectors of all $2k$-NN graphs with vertex set $[n]$.

Definition 2 (Exact recovery) *An estimator $\widehat{x} = \widehat{x}(w) \in \{0,1\}^{\binom{n}{2}}$ achieves exact recovery if, as $n \to \infty$,*

$$\sup_{x^* \in \mathcal{X}} \mathbb{P}\left\{\widehat{x} \neq x^*\right\} = o(1),$$

where w is distributed according to the hidden $2k$-NN graph model in Definition 1 with hidden $2k$-NN graph x^*.

Depending on the applications, we may not be able to reconstruct the hidden $2k$-NN graph x^* perfectly; instead, we may consider correctly estimating all but a small number of edges, which is required to be $o(nk)$, since a $2k$-NN graph contains kn edges. In particular, let $d(x^*, \widehat{x})$ be the Hamming distance $d(x^*, \widehat{x}) = \sum_e \mathbb{1}_{\{x_e^* \neq \widehat{x}_e\}}$.

Definition 3 (Almost exact recovery) *An estimator $\widehat{x} = \widehat{x}(w) \in \{0,1\}^{\binom{n}{2}}$ achieves almost exact recovery if, as $n \to \infty$,*

$$\sup_{x^* \in \mathcal{X}} \mathbb{E}\left[d(x^*, \widehat{x})\right] = o(nk).$$

Instead of using a permutation-based metric such as the Kendall tau distance, we choose the edge Hamming distance d for defining almost exact recovery, because for many practical application such as discovering strong ties in social networks, there is more value in recovering the edges rather than the permutation. Moreover, the edge sets arise naturally in the analysis of the maximum likelihood estimator $\widehat{x}_{\rm ML}$: the distribution of the log-likelihood ratio between a $2k$-NN graph x and the truth x^* only depends on the number of edges in x that differ from x^*. It is also worth noting that many computationally efficient algorithms, some of which to be discussed in Section 3.1, output an edge set instead of a permutation. One can further project the edge set to a $2k$-NN graph to recover the permutation, but it is unclear whether this projection can be done in polynomial time.

Intuitively, for a fixed network size n and a fixed number k of nearest neighbors, as the distributions P_n and Q_n get closer, the recovery problem becomes harder. This leads to an immediate question: *From an information-theoretic perspective, computational considerations aside, what are the fundamental limits of recovering the hidden $2k$-NN graph?* To answer this question, we derive necessary and sufficient conditions in terms of the model parameters (n, k, P_n, Q_n) under which the hidden $2k$-NN graph can be exactly or almost exactly recovered. These results serve as benchmarks for evaluating practical algorithms and aid us in understanding the performance limits of polynomial-time algorithms.

Specifically, we discover that the following two information measures characterize the sharp thresholds for exact and almost exact recovery, respectively. Define the Rényi divergence of order $1/2$:[2]

$$\alpha_n = -2 \log \int \sqrt{\mathrm{d}P_n \mathrm{d}Q_n}; \tag{1}$$

and the Kullback-Leibler divergence:

$$D(P_n \| Q_n) = \int \mathrm{d}P_n \log \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}.$$

Under some mild assumptions on P_n and Q_n, we show that the necessary and sufficient conditions are as follows:

- Exact recovery ($2 \leq k \leq n^{o(1)}$):

$$\liminf_{n\to\infty} \frac{2\alpha_n}{\log n} > 1; \tag{2}$$

- Almost exact recovery $\left(1 \leq k \leq o\left(\frac{\log n}{\log \log n}\right)\right)$:

$$\liminf_{n\to\infty} \frac{kD(P_n||Q_n)}{\log n} > 1. \tag{3}$$

The conditions for exact recovery and almost exact recovery are characterized by two different distance measures α_n and $D(P_n \| Q_n)$. This arises from large deviation analysis for different regimes of $d(x^*, \widehat{x})$. See Section 2.3 for a detailed explanation. For the special case of $k = 1$ (Hamiltonian cycle), the exact recovery condition was shown to be $\liminf_{n\to\infty} \frac{\alpha_n}{\log n} > 1$ (Bagaria et al., 2020). Comparing this with (2) for $k \geq 2$, we find that, somewhat surprisingly, the exact recovery threshold is halved when k increases from 1 to 2, and then stays unchanged as long as k remains $n^{o(1)}$. In contrast, the almost exact recovery threshold decreases inversely proportional to k over the range of $[1, o(\log n/ \log \log(n))]$. The sharp thresholds of exact recovery for $k \geq n^{\Omega(1)}$ and almost exact recovery for $k = \Omega(\log n/ \log \log n)$ remain open.

For the Bernoulli distribution (in other words, unweighted graphs) with $P_n = \mathrm{Bern}(p)$ and $Q_n = \mathrm{Bern}(q)$, we have the explicit expressions of

$$\alpha_n = -2 \log \left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right) \quad \text{and} \quad D(P_n \| Q_n) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

[2] It is also related to the so-called Battacharyya distance $B(P_n, Q_n)$ via $\alpha_n = 2B(P_n, Q_n)$.

As an interesting special case, consider the parametrization

$$p = 1 - \epsilon + \frac{2\epsilon k}{n-1} \quad \text{and} \quad q = \frac{2\epsilon k}{n-1}, \tag{4}$$

so that the mean number of edges in the observed graph stays at nk for all $\epsilon \in [0,1]$. This can be viewed as an approximate version of the Watts-Strogatz small-world graph, in which we start with a $2k$-NN graph, then rewire each edge with probability ϵ independently at random. In this case, our main results specialize to:

- Exact recovery is possible if and only if

$$\epsilon = o(1/n) \quad \text{for } k = 1; \quad \epsilon = o(1/\sqrt{n}) \quad \text{for } 2 \le k \le n^{o(1)}. \tag{5}$$

- Almost exact recovery is possible if and only if

$$k(1-\epsilon) \ge 1 + o(1) \quad \text{for } 1 \le k \le o(\log n / \log\log n). \tag{6}$$

In the related work (Cai et al., 2017), a similar case of Bernoulli distributions has been studied.[3] It is shown in Cai et al. (2017) that exact recovery is impossible if $1-\epsilon = o\left(\sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k} \frac{1}{\log \frac{n \log n}{k^2}}\right)$. In particular, this impossibility result requires $\epsilon \to 1$, which is highly suboptimal compared to the sharp exact recovery condition (5). It is also shown in Cai et al. (2017) that almost exact recovery can be achieved efficiently via thresholding on the number of common neighbors when $1 - \epsilon = \omega\left((\frac{\log n}{n})^{1/4} \vee (\frac{\log n}{k})^{1/2} \right)$ and via spectral ordering when $1 - \epsilon = \omega\left(\frac{n^{3.5}}{k^4}\right)$; these sufficient conditions, however, are very far from being optimal.

Finally, we remark that our sharp exact and almost exact recovery thresholds are achieved by the maximum likelihood estimator (MLE) for the hidden $2k$-NN graph problem, which is computationally intractable in the worst case. For the special case $k = 1$, the exact recovery threshold is shown to be achieved efficiently in polynomial-time via a linear programming (LP) relaxation of the MLE (namely, the fractional 2-factor LP) (Bagaria et al., 2020). For $k \ge 2$, however, it remains open whether the exact recovery threshold or the almost exact recovery threshold can be achieved efficiently in polynomial-time. In this work we analyze several computationally efficient algorithms and provide sufficient conditions under which they achieve exact/almost exact recovery. Moreover, under the small-world model where the edge weights are distributed as Bernoulli, we give a polynomial-time algorithm that attains the threshold for exact recovery.

The paper is organized as follows. In Section 2 we present our main results on the sharp thresholds for exact and almost recovery, and give a sketched proof. In Section 3 we analyze the computationally efficient recovery algorithms. Detailed proofs of all the results are included in the arXiv version Ding et al. (2019).

[3] To be precise, the previous work Cai et al. (2017) considers Bernoulli distributions under a slightly different parameterization: $p = 1 - \epsilon + \frac{2\epsilon^2 k}{n-1}$ and $q = \frac{2\epsilon k}{n-1}$. In addition to exact recovery and approximate recovery, a hypothesis testing problem between the small-world graph and Erdős-Rényi random graph is studied.

2. Sharp thresholds for recovery of hidden $2k$-NN graphs

2.1. Exact recovery

The maximum likelihood estimator for the hidden $2k$-NN graph problem is equivalent to finding the *max-weighted* $2k$*-NN subgraph* with weights given by the log likelihood ratios. Specifically, assuming that dP_n/dQ_n is well-defined, for each edge e, let $L_e = \log \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(w_e)$. Then the MLE is the solution to the following combinatorial optimization problem:

$$\widehat{x}_{\mathrm{ML}} = \arg\max_{x\in\mathcal{X}} \ \langle L, x\rangle\,, \tag{7}$$

where we recall that $\mathcal{X}$ denotes the collection of adjacency vectors of all $2k$-NN graphs on $[n]$. When $k = 1$, (7) reduces to the *max-weighted Hamiltonian cycle* problem. Note that in the Poisson, Gaussian or Bernoulli model where the log likelihood ratio is an affine function of the edge weight, we can simply replace L in (7) by the edge weights w.

Recall that $\alpha_n = -2\log\int\sqrt{dP_n dQ_n}$. We show that if $2 \le k \le n^{o(1)}$, then the condition $\liminf_{n\to\infty}(2\alpha_n/\log n) > 1$ is sufficient for $\widehat{x}_{\mathrm{ML}}$ to achieve exact recovery. This condition is also necessary, with the following additional assumption:

Assumption 1 (Bagaria et al. (2020), Assumption 1) *Let* $X = \log\frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(\omega_x)$ *for some* $\omega_x \sim P_n$ *and* $Y = \log\frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(\omega_y)$ *for some* $\omega_y \sim Q_n$*. Assume that*

$$\sup_{\tau\in\mathbb{R}}\left(\log\mathbb{P}\left\{Y\ge\tau\right\} + \log\mathbb{P}\left\{X\le\tau\right\}\right) \ge -(1+o(1))\alpha_n + o(\log n).$$

Remark 4 (Generality of Assumption 1) *Via Chernoff's inequality, it can be shown that (derivation in Bagaria et al. (2020), p67)* $\sup_{\tau\in\mathbb{R}}\left(\log\mathbb{P}\left\{Y\ge\tau\right\} + \log\mathbb{P}\left\{X\le\tau\right\}\right) \le -\alpha_n$*. We rely on this Chernoff's inequality in our large deviation analysis to establish the sufficient condition for* $\widehat{x}_{ML}$ *to achieve exact recovery. Assumption 1 essentially ensures that the Chernoff's inequality is asymptotically tight, so we can invert the large deviation analysis to show that the sufficient condition is also almost necessary. It was shown in (Bagaria et al., 2020, Lemma 6) that Assumption 1 is very general and is fulfilled by a wide class of weight distributions including Poisson, Gaussian and Bernoulli distributions.*

The following is our main result regarding exact recovery.

Theorem 5 (Exact recovery) *Let* $k \ge 2$*.*

- *Suppose*

$$\alpha_n - \frac{1}{2}(\log n + 17\log k) \to +\infty. \tag{8}$$

  Then the MLE (7) achieves exact recovery: $\mathbb{P}\left\{\widehat{x}_{\mathrm{ML}} \ne x^*\right\} \to 0$*. In particular, this holds if* $k = n^{o(1)}$ *and*

$$\liminf_{n\to\infty}\frac{2\alpha_n}{\log n} > 1.$$

- *Conversely, assume that* $k < n/12$ *and Assumption 1 holds. If exact recovery is possible, then*

$$\liminf_{n\to\infty}\frac{2\alpha_n}{\log n} \ge 1.$$

When $k = 1$, as shown in Bagaria et al. (2020) the sharp threshold for exact recovery is $\liminf_{n\to\infty} \frac{\alpha_n}{\log n} > 1$, which is stronger than the condition in Theorem 5 by a factor of 2. In other words, from $k = 1$ to $k \geq 2$ there is a strict decrease in the required level of signal. A simple explanation is that the hidden $2k$-NN graph x^* contains more edges when $k \geq 2$, and the elevated weights on these edges provide extra signal for determining the latent permutation σ^*. However, this extra information ceases to help as k increases from 2 to $n^{o(1)}$, which can be attributed to the following fact: when we swap any pair of adjacent vertices on σ^*, we always get a $2k$-NN graph x which differ from x^* by 4 edges, regardless of how large k is. In fact for all $2 \leq k \leq n^{o(1)}$, the bottleneck for exact recovery is formed by such swaps, resulting in the k-independent necessary condition $\liminf_{n\to\infty} \frac{2\alpha_n}{\log n} \geq 1$ (see Section 2.2 in the arXiv version for details).

2.2. Almost exact recovery

In this section, we present our main results for almost exact recovery. Let X_i's and Y_i's denote i.i.d. copies of the log-likelihood ratio $\log \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}$ under distributions P_n and Q_n respectively, with log MGFs $\psi_P(\lambda)$ and $\psi_Q(\lambda)$. Denote the Legendre transforms of the log MGFs as

$$\begin{aligned} E_Q(\tau) &= \psi_Q^*(\tau) \triangleq \sup_{\lambda\in\mathbb{R}} \lambda\tau - \psi_Q(\lambda), \\ E_P(\tau) &= \psi_P^*(\tau) \triangleq \sup_{\lambda\in\mathbb{R}} \lambda\tau - \psi_P(\lambda) = \sup_{\lambda\in\mathbb{R}} \lambda\tau - \psi_Q(1+\lambda) = E_Q(\tau) - \tau. \end{aligned} \tag{9}$$

Then Chernoff bound gives that for all $\tau \in [-D(Q_n\|P_n), D(P_n\|Q_n)]$ and $\Delta \geq 1$,

$$\mathbb{P}\left\{\sum_{i=1}^{\Delta} X_i \leq \Delta\tau\right\} \leq e^{-\Delta E_P(\tau)}, \quad \mathbb{P}\left\{\sum_{i=1}^{\Delta} Y_i \geq \Delta\tau\right\} \leq e^{-\Delta E_Q(\tau)}. \tag{10}$$

Note that E_P and E_Q are convex and monotone functions, such that as τ increases from $-D(Q_n\|P_n)$ to $D(P_n\|Q_n)$, $E_Q(\tau)$ increases from 0 to $D(P_n\|Q_n)$ and $E_P(\tau)$ decreases from $D(Q_n\|P_n)$ to 0. The following assumption postulates a quadratic lower bound of E_P at the boundary:

Assumption 2 *There exists an absolute constant $c > 0$, such that for all $\eta \in [0, 1]$,*

$$E_P((1-\eta)D(P_n\|Q_n)) \geq c\eta^2 D(P_n\|Q_n). \tag{11}$$

Remark 6 (Generality of Assumption 2) *Note that $E_P(\tau)$ is convex with minimum 0 and curvature (second-order derivative) $1/Var_P(\log(dP_n/dQ_n))$ at $\tau = D(P_n\|Q_n)$. In view of Taylor expansion of $E_P(\tau)$ at $\tau = D(P_n\|Q_n)$, Assumption 2 essentially ensures that $E_P(\tau)$ is bounded from below by a quadratic parabola with curvature at least $\Omega(1/D(P_n\|Q_n))$, giving us the desired stability on the error exponent in the large deviation analysis of log-likelihood ratios at τ close to $D(P_n\|Q_n)$. When the weight distributions are Gaussian, $E_P(\tau)$ is exactly a quadratic parabola with curvature $1/(2D(P_n\|Q_n))$ at $\tau = D(P_n\|Q_n)$. Thus Assumption 2 holds. It can also be shown that Assumption 2 is satisfied whenever the distribution of $\log(dP_n/dQ_n)$ under P_n is sub-Gaussian with proxy variance $O(D(P_n\|Q_n))$ (see Hajek et al. (2017), Section 3).*

Theorem 7 (Almost exact recovery) *If Assumption 2 holds, $k\log k = o(\log n)$ and*

$$\liminf_{n\to\infty} \frac{kD(P_n\|Q_n)}{\log n} > 1, \tag{12}$$

then the MLE (7) achieves almost exact recovery. Conversely, if $k = O(\log n)$ and almost exact recovery is possible, then

$$\liminf_{n\to\infty} \frac{kD(P_n\|Q_n)}{\log n} \geq 1. \tag{13}$$

Theorem 7 should be compared with the exact recovery threshold $\liminf(2\alpha_n/\log n) > 1$ for $2 \leq k \leq n^{o(1)}$; the latter is always stronger, since

$$\alpha_n = -2\log\int\sqrt{dP_n dQ_n} = -2\log\mathbb{E}_{P_n}\sqrt{\frac{dQ_n}{dP_n}} \leq -2\mathbb{E}_{P_n}\log\sqrt{\frac{dQ_n}{dP_n}} = D(P_n\|Q_n),$$

by Jensen's inequality. Unlike exact recovery, the almost exact recovery threshold is inversely proportional to k. Intuitively, this is because almost exact recovery only requires one to distinguish the latent $2k$-NN graph x^* from those $2k$-NN graphs that differ from x^* by $\Omega(kn)$ edges; in contrast, as we show in Section 2.2 in the arXiv version, the condition for exact recovery arises from eliminating those solutions differing from x^* by four edges.

2.3. Proof sketch of the results on recovery thresholds

To prove Theorem 5 and Theorem 7, we need to introduce the notion of *difference graph*, which encodes the difference between a proposed $2k$-NN graph and the ground truth. Given $x, x^* \in \{0,1\}^{\binom{n}{2}}$, let $G = G(x)$ be a bi-colored simple graph on $[n]$ whose adjacency vector is $x - x^* \in \{0,\pm1\}^{\binom{n}{2}}$, in the sense that each pair (i,j) is connected by a blue (resp. red) edge if $x_{ij} - x^*_{ij} = 1$ (resp. -1). See Fig. 2 for an example. By definition, red edges in $G(x)$ are true edges in x^* that are missed by the proposed solution x, and blue edges correspond to spurious edges that are absent in the ground truth.

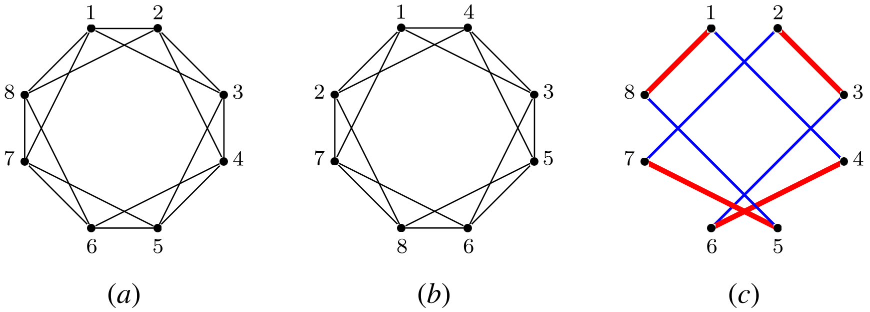

Figure 2: An example for a difference graph G. From left to right: (a) The $2k$-NN graph x^* corresponding to the Hamiltonian cycle $(1,2,3,4,5,6,7,8,1)$; (b) The $2k$-NN graph x corresponding to the Hamiltonian cycle $(1,4,3,5,6,8,7,2,1)$. (c) The difference graph G formed by $x - x^*$. The red (thick) edges stand for edges that in x^* but not x, while the blue (thin) edges are in x but not x^*.

A key property of difference graphs is the following: Since $2k$-NN graphs are $2k$-regular, the difference graph G is *balanced*, in the sense that for each vertex, its red degree (the number of incident red edges) coincides with its blue degree. Consequently, G has equal number of red edges and blue edges, and the number of red (or blue) edges measures the closeness of x to x^*. Denote

$$\mathcal{X}_\Delta = \{x \in \mathcal{X} : d(x,x^*) = 2\Delta\} = \{x \in \mathcal{X} : G(x) \text{ contains exactly } \Delta \text{ red edges}\}. \tag{14}$$

In particular, $\{\mathcal{X}_\Delta : \Delta \geq 0\}$ partitions the feasible set $\mathcal{X}$. The analysis of the MLE relies crucially on bounding the size of $\mathcal{X}_\Delta$. To see this, note that by the definition of the MLE,

$$\mathbb{P}\left\{\widehat{x}_{\rm ML} \neq x^*\right\} \leq \mathbb{P}\left\{\exists x \neq x^* : \langle L, x - x^* \rangle \geq 0\right\} \leq \sum_{\Delta \geq 2} \mathbb{P}\left\{\exists x \in \mathcal{X}_\Delta : \langle L, x - x^* \rangle \geq 0\right\}. \quad (15)$$

Thus the MLE achieves exact recovery if the right-hand side of (15) is of order $o(1)$. Similarly, for the MLE to achieve almost exact recovery we need

$$\mathbb{P}\left\{d(\widehat{x}_{\rm ML}, x^*) \geq 2\epsilon_n nk\right\} \leq \sum_{\Delta \geq \epsilon_n nk} \mathbb{P}\left\{\exists x \in \mathcal{X}_\Delta : \langle L, x - x^* \rangle \geq 0\right\} \quad (16)$$

to be of order $o(1)$ for some sequence ϵ_n going to 0. Naturally we would want to bound the summands of (15) and (16) via a union bound, which calls for an upper bound on the size of $\mathcal{X}_\Delta$.

Following similar arguments as in (Bagaria et al., 2020, Sec. 4.2), we can prove a simple bound

$$|\mathcal{X}_\Delta| \leq (4kn)^\Delta, \quad (17)$$

resulting in a condition that is suboptimal compared to the desired (8) by a factor of 2 when $k \geq 2$. To achieve the sharp threshold for exact recovery, (17) can be significantly improved by a delicate combinatorial argument:

$$|\mathcal{X}_\Delta| \leq 2\left(Ck^{17}n\right)^{\Delta/2} \quad (18)$$

for some universal constant C. The full proof of (18) is provided in Section 2.3 of the arXiv version, and constitutes the most crucial part of the argument. The key idea is to count the red edge sets and blue edge sets separately. For a $2k$-NN graph x, let $E_{\rm red}(x)$ denote the set of red edges in $G(x)$. Define

$$\mathcal{E}_{\rm red}(\Delta) = \left\{E_{\rm red}(x) : x \in \mathcal{X}_\Delta\right\}, \quad \mathcal{X}\left(E_{\rm red}\right) = \left\{x \in \mathcal{X}_\Delta : E_{\rm red}(x) = E_{\rm red}\right\}.$$

To count the members in $\mathcal{X}_\Delta$, we first enumerate the set of red edges; then for a fixed set of red edges we enumerate the $2k$-NN graphs that are compatible with it. In particular, we show that for all $\Delta \geq 2$ and $E_{\rm red} \in \mathcal{E}_{\rm red}(\Delta)$,

$$|\mathcal{E}_{\rm red}(\Delta)| \leq (96k^2)^\Delta \binom{kn}{\lfloor \Delta/2 \rfloor}, \quad |\mathcal{X}(E_{\rm red})| \leq 2(32k^3)^{2\Delta}\Delta^{\Delta/k}. \quad (19)$$

The desired bound (18) immediately follows.

Proof of the sufficiency part for exact recovery: For each $x \in \mathcal{X}_\Delta$, the law of $\langle L, x - x^* \rangle$ only depends on Δ: $\langle L, x - x^* \rangle \overset{d}{=} \sum_{i \leq \Delta}(Y_i - X_i)$, where X_i's and Y_i's are i.i.d. copies of $\log \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}$ under P_n and Q_n, respectively, and $\overset{d}{=}$ denotes equal in distribution. By the Chernoff bound, $\mathbb{P}\{\sum_{i \leq \Delta}(Y_i - X_i)\} \leq \exp(-\alpha_n \Delta)$, where α_n is defined in (1). Using (18) and the union bound we have

$$\mathbb{P}\left\{\exists x \in \mathcal{X}_\Delta : \langle L, x - x^* \rangle \geq 0\right\} \leq 2\left(Ck^{17}n\right)^{\Delta/2} \exp(-\alpha_n \Delta). \quad (20)$$

Substituting (20) into (15) yields that $\mathbb{P}\{\widehat{x}_{\rm ML} \neq x^*\} \to 0$ provided that $\alpha_n - \frac{1}{2}\log(Ck^{17}n) \to +\infty$.

Proof of the sufficiency part for almost exact recovery: For almost exact recovery we only need to rule out $\widehat{x}_{ML} \in \mathcal{X}_\Delta$ for some $\Delta \geq \epsilon_n nk$. In this range, there is a large difference between

$|\mathcal{E}_{\text{red}}(\Delta)|$ and $|\mathcal{X}_\Delta|$. Indeed from (19), there may be up to $2(32k^3)^{2\Delta}\Delta^{\Delta/k}$ members of $\mathcal{X}_\Delta$ with the same set of red edges. Hence for large Δ, it is more advantageous to separate the contributions from the red edges and blue edges. We have for all $\tau \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}\left\{\exists x \in \mathcal{X}_\Delta : \langle L, x - x^* \rangle \geq 0\right\} &\leq |\mathcal{E}_{\text{red}}(\Delta)|\, \mathbb{P}\left\{\sum_{i \leq \Delta} X_i \leq \Delta\tau\right\} + |\mathcal{X}_\Delta|\, \mathbb{P}\left\{\sum_{i \leq \Delta} Y_i \geq \Delta\tau\right\} \\ &\overset{(10)}{\leq} |\mathcal{E}_{\text{red}}(\Delta)|\, e^{-\Delta E_P(\tau)} + |\mathcal{X}_\Delta|\, e^{-\Delta E_Q(\tau)}. \end{aligned} \tag{21}$$

To balance out the two terms in (21), the exponential tilting parameter τ is chosen so that $E_Q(\tau)$ is large. Given that $E_Q(\tau)$ is an increasing function on $[-D(Q_n\|P_n), D(P_n\|Q_n)]$, we choose τ close to $D(P_n\|Q_n)$. Under the assumption $\liminf(kD(P_n\|Q_n)/\log n) > 1$, we have $kD(P_n\|Q_n)(1-\eta) \geq \log n$ for some $\eta \in (0,1)$. Set $\tau = (1-\eta)D(P_n\|Q_n)$, $\epsilon_n = 1/(kD(P_n\|Q_n))$, and use the bounds in (19) to further upper bound both terms in (21) by $\exp(-\Delta\Omega(D(P_n\|Q_n)))$. It follows from (16) that

$$\mathbb{P}\left\{d(\widehat{x}_{\text{ML}}, x^*) \geq 2\epsilon_n nk\right\} \leq 2 \sum_{\Delta \geq \epsilon_n nk} \exp\left(-\Delta\Omega(D(P_n\|Q_n))\right) = \exp(-\Omega(n)) = o(1).$$

Proof techniques for the necessary conditions: The bottleneck for exact recovery happens at $\Delta = 2$. The set $\mathcal{X}_\Delta$ consists of difference graphs formed by reversing the order of two adjacent vertices on the permutation σ^*. In total there are n such difference graphs and they are close to being independent. As a result the union bound in (20) is almost tight for the $\Delta = 2$ summand, so that the sufficient condition for exact recovery turns out to also be almost necessary. For almost exact recovery, the necessary condition follows from a mutual information and rate-distortion argument. See Section 2.2 and Section 3.2 in the arXiv version for the full proofs.

3. Efficient recovery algorithms

3.1. Efficient recovery algorithms for the general model

For simplicity we focus on the hidden $2k$-NN graph model with Gaussian weight distributions $P_n = \mathcal{N}(\mu_n, 1)$ and $Q_n = \mathcal{N}(0,1)$ for $\mu_n > 0$. Analysis in this section can be extended to general weight distributions. From Theorems 5 and 7, under the Gaussian model, the sharp thresholds for exact recovery (for $2 \leq k \leq n^{o(1)}$) and almost exact recovery (for $1 \leq k \leq o(\log n/\log\log n)$) are

$$\liminf_{n\to\infty} \frac{\mu_n^2}{2\log n} > 1, \quad \text{and} \quad \liminf_{n\to\infty} \frac{k\mu_n^2}{2\log n} > 1, \tag{22}$$

respectively. Since the log likelihood ratio is given by $L_e = \log \frac{\mathrm{d}P_n}{\mathrm{d}Q_n}(w_e) = \mu_n w_e - \mu_n^2$, the MLE (7) simplifies to

$$\widehat{x}_{\text{ML}} = \operatorname{argmax}_{x \in \mathcal{X}} \langle w, x \rangle. \tag{23}$$

In the special case of $k = 1$, this reduces to the max-weighted Hamiltonian cycle problem. The previous work (Bagaria et al., 2020) analyzes its 2-factor integer linear program (ILP) relaxation and fractional 2-factor linear program (LP) relaxation, and show that they achieve the sharp exact recovery threshold $\liminf_{n\to\infty} \frac{\mu_n^2}{4\log n} > 1$. This motivates us to consider the ILP and LP relaxation for general k.

$2k$-factor ILP relaxation By relaxing the $2k$-NN graph constraint in the MLE (23) to a degree constraint, we arrive at the following $2k$-factor ILP:

$$\begin{aligned} \widehat{x}_{2k\mathrm{F}} = \operatorname{argmax}_x \; & \langle w, x\rangle \\ \text{s.t.} \; & \sum_{v\sim u} x_{(u,v)} = 2k, \;\; \forall u, \\ & x_e \in \{0,1\}, \;\; \forall e \end{aligned} \tag{24}$$

where the first constraint enforces that every vertex has degree $2k$. It is known that for constant k, the ILP (24) can be solvable in $O(n^4)$ time (Letchford et al., 2008).

To analyze $\widehat{x}_{2k\mathrm{F}}$, note that each feasible solution x to the ILP is a $2k$-regular graph. Therefore, the difference graph $x - x^*$ is still balanced and the simple bound (17) continues to hold: $|\mathcal{Y}_\Delta| \le (4kn)^\Delta$, where $\mathcal{Y}_\Delta$ is the collection of $2k$-regular graphs x such that the difference graph $x - x^*$ contains exactly Δ red edges. Moreover, for $x \in \mathcal{Y}_\Delta$, $\langle w, x - x^*\rangle \sim \mathcal{N}(-\Delta\mu_n, 2\Delta)$. Hence from the union bound

$$\mathbb{P}\{\widehat{x}_{2k\mathrm{F}} \neq x^*\} \le \sum_{\Delta\ge 1} (4kn)^\Delta \exp\left(-\frac{\Delta\mu_n^2}{4}\right) = \sum_{\Delta\ge 1} \exp\left(-\Delta\left(\frac{\mu_n^2}{4} - \log(4kn)\right)\right).$$

We conclude that when $2 \le k \le n^{o(1)}$, $\widehat{x}_{2k\mathrm{F}}$ achieves exact recovery if $\liminf_{n\to\infty} \mu_n^2/(4\log n) > 1$, which is only off by a multiplicative factor of 2 compared to the sharp threshold (22).

LP relaxation By further relaxing the integer constraint in $\widehat{x}_{2k\mathrm{F}}$, we arrive at the following LP:

$$\begin{aligned} \widehat{x}_{\mathrm{LP}} = \operatorname{argmax}_x \; & \langle w, x\rangle \\ \text{s.t.} \; & \sum_{v\sim u} x_{(u,v)} = 2k, \;\; \forall u, \\ & x_e \in [0,1], \;\; \forall e. \end{aligned}$$

We claim that even though $\widehat{x}_{\mathrm{LP}}$ is a relaxation of $\widehat{x}_{2k\mathrm{F}}$, it still achieves exact recovery for $2 \le k \le n^{o(1)}$ when $\liminf_{n\to\infty} \mu_n^2/(4\log n) > 1$. That is because firstly, the feasible set of the LP is a fractional $2k$-factor polytope, the entries of whose extreme points are all half-integrals by the determinant analysis analysis in (Balinski, 1965, p 280). That is, $(\widehat{x}_{\mathrm{LP}})_e \in \{0, 1/2, 1\}$ for all e. Moreover, for a half-integral $2k$-factor graph x, the difference graph $x - x^*$ can be represented by a balanced multigraph with edge multiplicity at most 2 (we refer the reader to Bagaria et al. (2020) for details). The rest of the proof follows exactly from the proof of (Bagaria et al., 2020, Theorem 1).

To sum up, both $\widehat{x}_{2k\mathrm{F}}$ and $\widehat{x}_{\mathrm{LP}}$ achieve exact recovery under the condition $\liminf_{n\to\infty} \mu_n^2/(4\log n) > 1$. Whether they can achieve almost exact recovery under weaker conditions remains open.

Simple thresholding To partially address the problem of almost exact recovery, we consider a naïve thresholding estimator $\widehat{x}_{\mathrm{TH}}$ given by

$$\widehat{x}_{\mathrm{TH}}(e) = \mathbb{1}\left\{w_e > \sqrt{(2+\epsilon_n)\log n}\right\},$$

where the sequence ϵ_n will be later specified. For each edge e in the true $2k$-NN graph x^*, $w_e \sim \mathcal{N}(\mu_n, 1)$ and thus

$$\mathbb{P}\{\widehat{x}_{\mathrm{TH}}(e) = 0\} \le \exp\left(-(\mu_n - \sqrt{(2+\epsilon_n)\log n})^2/2\right);$$

Similarly for edge e not in x^*,

$$\mathbb{P}\{\widehat{x}_{\mathrm{TH}}(e) = 1\} \le \exp(-(\sqrt{(2+\epsilon_n)\log n})^2/2) = n^{-(1+\epsilon_n/2)}.$$

Recall that $d(\widehat{x}_{\mathrm{TH}}, x^*) = \sum_e \mathbb{1}\{\widehat{x}_{\mathrm{TH}}(e) \neq x^*(e)\}$. We have

$$\mathbb{E}\left[d\left(\widehat{x}_{\mathrm{TH}}, x^*\right)\right] \le kn \exp\left(-(\mu_n - \sqrt{(2+\epsilon_n)\log n})^2/2\right) + n^2 \cdot n^{-(1+\epsilon_n/2)}.$$

Hence for $\mathbb{E}(d(\widehat{x}_{\mathrm{TH}}, x^*))$ to be of order $o(nk)$, it suffices to choose ϵ_n such that $\mu_n - \sqrt{(2+\epsilon_n)\log n} = \omega(1)$ and $\epsilon_n \log n = -2\log k + \omega(1)$. Such an ϵ_n sequence exists as long as $\mu_n = \sqrt{2\log n - 2\log k} + \omega(1)$. In other words, the estimator $\widehat{x}_{\mathrm{TH}}$ achieves almost exact recovery under the condition $\mu_n = \sqrt{2\log(n/k)} + \omega(1)$, which is optimal for $k = 1$ in view of (22).

It is worth pointing out that $\widehat{x}_{\mathrm{TH}}$ may not be a valid $2k$-NN graph. One can of course consider the modified estimator by projecting $\widehat{x}$ to the set of $2k$-NN graphs; however, it is unclear whether this can be done in polynomial time. It is an interesting open problem whether a computationally efficient $2k$-NN graph estimator can be obtained from $\widehat{x}_{\mathrm{TH}}$ and still inherits the almost exact recovery guarantee $\mu_n = \sqrt{2\log(n/k)} + \omega(1)$.

In passing, we remark that although spectral methods have been successfully used to recover the hidden structures based on the principal eigenvectors of the observed graph for a variety of problems such as clustering and community detection, the spectral methods are highly suboptimal in our model when $k = n^{o(1)}$, as the adjacency matrix of the $2k$-NN graph is full-rank and has a vanishing eigen-gap (See Section 4.1 in the arXiv version for details).

3.2. Achieving the sharp threshold for exact recovery under the small-world model

In this section we introduce a polynomial-time algorithm that attains the exact recovery threshold under the Watts-Strogatz small-world model. Recall that $P_n = \mathrm{Bern}(p)$ and $Q_n = \mathrm{Bern}(q)$, where

$$p = 1 - \epsilon + \frac{2\epsilon k}{n-1} \quad \text{and} \quad q = \frac{2\epsilon k}{n-1}. \tag{25}$$

The observed graph $w \in \{0,1\}^{\binom{n}{2}}$ can be viewed as a noisy version of the true $2k$-NN graph x^*. By Theorem 5, for $2 \le k \le n^{o(1)}$, the sharp threshold for exact recovery is $\liminf(-2\log\epsilon/\log n) > 1$, i.e., $\epsilon \le n^{-\frac{1}{2}-\Omega(1)}$. We give a polynomial-time algorithm that succeeds under this condition. Note that to exactly recover x^*, it suffices to recover the corresponding Hamiltonian cycle identified by a permutation σ^*. To recover σ^*, the algorithm works by first determining the neighborhood of one vertex and their ordering on the Hamiltonian cycle, and then sequentially finding the remaining vertices to complete the cycle in a greedy manner. See Section 4.2 in the arXiv version for a detailed description of the algorithm and the proof that it achieves the sharp threshold for exact recovery.

Acknowledgments

J. Ding is supported in part by the NSF Grant DMS-1757479. Y. Wu is supported in part by the NSF Grant CCF-1900507, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. J. Xu is supported by the NSF Grants IIS-1838124, CCF-1850743, and CCF-1856424. D. Yang is supported by the NSF Grant CCF-1850743. We thank David Pollard for his comments and suggestions.

References

Vivek Bagaria, Jian Ding, David Tse, Yihong Wu, and Jiaming Xu. Hidden hamiltonian cycle recovery via linear programming. *Operations Research*, 2020.

Michel Louis Balinski. Integer programming: methods, uses, computations. *Management science*, 12(3):253–313, 1965.

Danielle Smith Bassett and ED Bullmore. Small-world brain networks. *The neuroscientist*, 12(6): 512–523, 2006.

T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. On detection and structural reconstruction of small-world random networks. *IEEE Transactions on Network Science and Engineering*, 4(3): 165–176, 2017.

Ramon Ferrer I Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.

Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. Consistent recovery threshold of hidden nearest neighbor graphs. *arXiv preprint arXiv:1911.08004*, 2019.

David Easley and Jon Kleinberg. *Networks, crowds, and markets*. Cambridge university press, 2010.

Eric Gilbert. Predicting tie strength in a new medium. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1047–1056. ACM, 2012.

Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. *IEEE Trans. Information Theory*, 63(8):4729–4745, September 2017. arXiv 1509.07859.

Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2008. ISSN 13960466. doi: 10.5210/fm.v14i1.2317. URL https://firstmonday.org/ojs/index.php/fm/article/view/2317.

Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, pages 163–170, New York, NY, USA, 2000. ACM. ISBN 1-58113-184-4. doi: 10.1145/335305.335325. URL http://doi.acm.org/10.1145/335305.335325.

Adam N Letchford, Gerhard Reinelt, and Dirk Oliver Theis. Odd minimum cut sets and b-matchings revisited. *SIAM Journal on Discrete Mathematics*, 22(4):1480–1487, 2008.

Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn. Maintained relationships on Facebook. http://overstated.net/2009/03/09/maintained-relationships-on-facebook, March 2009. Accessed: 2019-09-22.

Peter V Marsden and Karen E Campbell. Measuring tie strength. *Social forces*, 63(2):482–501, 1984.

Cristopher Moore and Mark EJ Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678, 2000.

Adilson E Motter, Alessandro PS De Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102, 2002.

Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

Mark EJ Newman and Duncan J Watts. Scaling and percolation in the small-world network model. *Physical review E*, 60(6):7332, 1999.

J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

Rahmtin Rotabi, Krishna Kamath, Jon Kleinberg, and Aneesh Sharma. Detecting strong ties using network motifs. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 983–992. International World Wide Web Conferences Steering Committee, 2017.

Jari Saramäki and Kimmo Kaski. Modelling development of epidemics with dynamic small-world networks. *Journal of Theoretical Biology*, 234(3):413–421, 2005.

Andreas Wagner and David A Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810, 2001.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.