Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Hidden Hamiltonian Cycle Recovery via Linear Programming

Vivek Bagaria, Jian Ding, David Tse, Yihong Wu, Jiaming Xu

To cite this article:
Vivek Bagaria, Jian Ding, David Tse, Yihong Wu, Jiaming Xu (2020) Hidden Hamiltonian Cycle Recovery via Linear Programming. Operations Research 68(1):53-70. https://doi.org/10.1287/opre.2019.1886

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 68, No. 1, January–February 2020, pp. 53–70
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Hidden Hamiltonian Cycle Recovery via Linear Programming

Vivek Bagaria,[a] Jian Ding,[b] David Tse,[a] Yihong Wu,[c] Jiaming Xu[d]

[a] Department of Electrical Engineering, Stanford University, Stanford, California 94305; [b] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [c] Department of Statistics and Data Science, Yale University, New Haven, Connecticut 06511; [d] Fuqua School of Business, Duke University, Durham, North Carolina 27708

Contact: vbagaria@stanford.edu (VB); dingjian@wharton.upenn.edu (JD); dntse@stanford.edu (DT); yihong.wu@yale.edu (YW); jiaming.xu868@duke.edu, http://orcid.org/0000-0001-6104-4742 (JX)

Received: August 31, 2018
Accepted: April 3, 2019
Published Online in Articles in Advance: January 2, 2020

Subject Classifications: networks/graphs: traveling salesman; programming: integer: relaxation; mathematics: combinatorics; probability: distributions
Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2019.1886

Abstract. We introduce the problem of hidden Hamiltonian cycle recovery, where there is an unknown Hamiltonian cycle in an n-vertex complete graph that needs to be inferred from noisy edge measurements. The measurements are independent and distributed according to P_n for edges in the cycle and Q_n otherwise. This formulation is motivated by a problem in genome assembly, where the goal is to order a set of contigs (genome subsequences) according to their positions on the genome using long-range linking measurements between the contigs. Computing the maximum likelihood estimate in this model reduces to a traveling salesman problem (TSP). Despite the NP-hardness of TSP, we show that a simple linear programming (LP) relaxation—namely, the fractional 2-factor (F2F) LP—recovers the hidden Hamiltonian cycle with high probability as $n \to \infty$ provided that $\alpha_n - \log n \to \infty$, where $\alpha_n \triangleq -2\log \int \sqrt{dP_n dQ_n}$ is the Rényi divergence of order $\frac{1}{2}$. This condition is information-theoretically optimal in the sense that, under mild distributional assumptions, $\alpha_n \geq (1 + o(1))\log n$ is necessary for any algorithm to succeed regardless of the computational cost. Departing from the usual proof techniques based on dual witness construction, the analysis relies on the combinatorial characterization (in particular, the half-integrality) of the extreme points of the F2F polytope. Represented as bicolored multigraphs, these extreme points are further decomposed into simpler "blossom-type" structures for the large deviation analysis and counting arguments. Evaluation of the algorithm on real data shows improvements over existing approaches.

Funding: V. Bagaria and D. Tse are supported by the Center for Science of Information, a National Science Foundation (NSF) Science and Technology Center grant [CCF-0939370], as well as the National Human Genome Research Institute of the National Institutes of Health [Award R01HG008164]. J. Ding was supported in part by the NSF [Grant DMS-1757479] and an Alfred Sloan fellowship. Y. Wu was supported in part by the NSF [Grants IIS-1447879 and CCF-1527105], the NSF CAREER program [Award CCF-1651588], and an Alfred Sloan fellowship. J. Xu was supported in part by a Simons-Berkeley Research Fellowship and the NSF [Grants CCF-1850743, IIS-1838124, and CCF-1856424].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2019.1886.

Keywords: traveling salesman problem • fractional 2-factor linear programming • polyhedral combinatorics • large deviations theory

1. Introduction

Given an input graph, the problem of finding a subgraph satisfying certain properties has diverse applications. MAX CUT, MAX CLIQUE, and the traveling salesman problem (TSP) are a few canonical examples. Traditionally, these problems have been studied in theoretical computer science from the worst-case perspective, and many such problems have been shown to be NP-hard. However, in machine learning applications, many such problems arise when an underlying *ground truth* subgraph needs to be recovered from the noisy measurement data represented by the entire graph. Canonical models to study such problems include planted partition models (Condon and Karp 2001) (such as planted clique; see Jerrum 1992) in community detection and planted ranking models (such as the Mallows model; see Mallows 1957) in rank aggregation. In these models, the planted or hidden subgraph represents the ground truth, and one is not necessarily interested in the worst-case instances but rather in only instances for which there is enough information in the data to recover the ground truth subgraph (i.e., when the amount of is are above the information limit). The key question is whether there exists an efficient recovery algorithm that can be successful all the way to the information limit.

In this paper, we pose and answer this question for a hidden Hamiltonian cycle recovery model.

Definition 1 (Hidden Hamiltonian Cycle Recovery)**.**

- *Given*: Let $n \geq 1$, and there are two distributions P_n and Q_n, parameterized by n.
- *Observation*: We observe a randomly weighted, undirected complete graph $G = ([n], E)$ with a hidden Hamiltonian cycle C^* such that every edge has an independent weight distributed as P_n if it is on C^* and as Q_n otherwise.
- *Inference Problem*: Recover the hidden Hamiltonian cycle C^* from the observed random graph.

Our problem is motivated from de novo genome assembly, the reconstruction of an organism's long sequence of A–G–C–T nucleotides from fragmented sequencing data. The first step of the standard assembly pipeline stitches together short, overlapping fragments (so-called shotgun reads) to form longer subsequences called contigs, of lengths typically tens to hundreds of thousands of nucleotides (Figure 1). Because of coverage gaps and other issues, these individual contigs cannot be extended to the whole genome. To get a more complete picture of the genome, the contigs need to be ordered according to their positions on the genome, a process called *scaffolding*. Recent advances in sequencing assays (Lieberman-Aiden et al. 2009, Putnam et al. 2016) aid this process by providing long-range linking information between these contigs in the form of randomly sampled *Hi-C reads*. These data can be summarized by a contact map (Figure 2), tabulating the counts of Hi-C reads linking each pair of contigs. The problem of ordering the contigs from the contact map data can be modeled by the hidden Hamiltonian cycle recovery problem, where the vertices of the graph are the contigs, the hidden Hamiltonian cycle is the true ordering of the contigs on the genome, and the weights on the graph are the counts of the Hi-C reads linking the contigs. Strictly speaking, this applies only to genomes that are circular. For genomes that are linear, the ordering of the contigs would correspond to a hidden Hamiltonian *path*. We show in Section EC.1 of the e-companion that our results extend to a hidden Hamiltonian path model as well. As can be seen in Figure 2(a), there is a much larger concentration of Hi-C reads between contigs adjacent on the genome than between faraway contigs. A first-order model is to choose $P_n = \text{Pois}(\lambda_n)$ and $Q_n = \text{Pois}(\mu_n)$, where λ_n is the average number of Hi-C reads between adjacent contigs and μ_n is the average number between nonadjacent contigs. The parameter $n\lambda_n + \frac{n(n-1)}{2}\mu_n$ increases with the coverage depth (the average number of Hi-C reads that include a given nucleotide) of the Hi-C reads and is part of the design of the sequencing experiment.

The hidden Hamiltonian cycle can be represented as an adjacency vector $x^* \in \{0,1\}^{\binom{n}{2}}$ such that $x^*_e = 1$ if edge e is on the Hamiltonian cycle and $x^*_e = 0$ otherwise. Let A denote the weighted adjacency matrix of G so that A_e is distributed according to P_n (respectively, Q_n) if $x^*_e = 1$ (respectively, 0). The maximum likelihood (ML) estimator for the hidden Hamiltonian cycle recovery problem is equivalent to solving the traveling salesman problem (TSP) on a transformed weighted graph, where each edge weight $w_e = \log \frac{dP_n}{dQ_n}(A_e)$ is the log likelihood ratio evaluated on the weights of the observed graph:

$$\widehat{x}_{\text{ML}} = \arg\max_x \langle w, x \rangle \quad \text{s.t.} \quad x \in \mathscr{X}(G), \tag{1}$$

where $\mathscr{X}(G)$ denotes the set of adjacency vectors of all possible Hamiltonian cycles in G. In the Poisson or Gaussian model where the log likelihood ratio is an affine function, we can simply take w to be A itself.

Figure 1. (Color online) Short Reads Are Assembled to Form Contigs, Which Are Then Ordered by Using Long-Range Linking Hi-C Reads

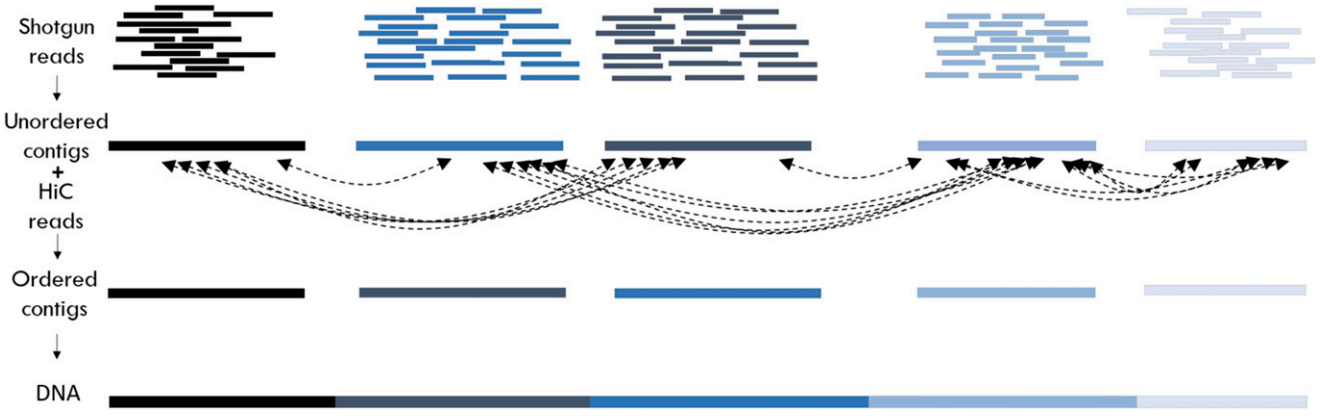
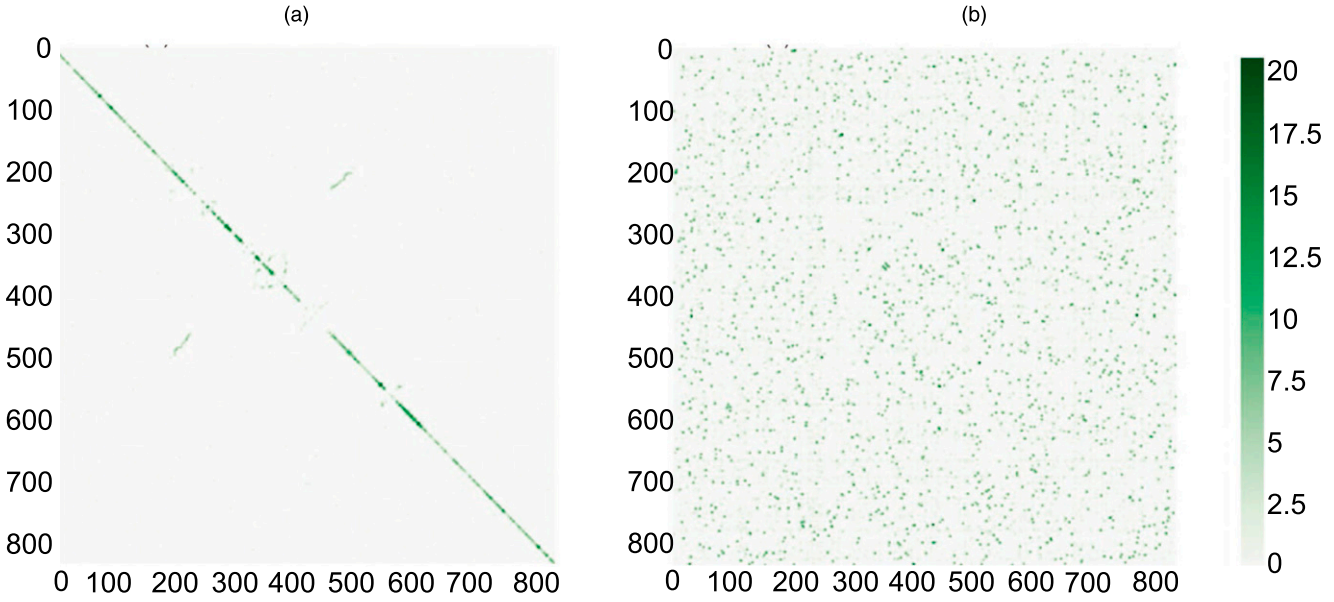


Figure 2. (Color online) (a) Contact Map Where the Rows (and Columns) Correspond to *Ordered* Contigs of Human Chromosome 1 (Putnam et al. 2016), and the Value at Entry (i, j) Corresponds to the Number of Hi-C Reads Between Contig i and Contig j; (b) Contact Map of the *Unordered* Matrix in (a), Where the Contigs Are Randomly Ordered (These Are the Data from Which the Ordering of the Contigs Is Inferred)



Solving TSP is NP-hard, and a natural approach is to look for a tractable relaxation. It is well known that TSP (1) can be cast as an integer linear program (ILP) (Schalekamp et al. 2013):

$$\widehat{x}_{\mathrm{TSP}} = \arg\max_{x} \langle w, x\rangle \tag{2}$$

$$\text{s.t. } x(\delta(v)) = 2, \tag{3}$$

$$x(\delta(S)) \geq 2,\ \forall S \subset [n],\ 3 \leq |S| \leq n-3, \tag{4}$$

$$x_e \in \{0,1\}, \tag{5}$$

where $\delta(S)$ denotes the set of all edges in G with exactly one endpoint in $S \subset [n]$, and $\delta(v) \triangleq \delta(\{v\})$; $x(\delta(S)) = \sum_{e\in\delta(S)} x_e$. In particular, (3) are called *degree* constraints, enforcing each vertex to have exactly two incident edges in the graph represented by the adjacency vector x, whereas (4) are *subtour elimination* constraints, eliminating solutions whose corresponding graph is a disjoint union of subtours of length less than n. Note that there are exponentially large numbers of subtour elimination constraints. If we drop the subtour elimination constraints as well as relax the integer constraints on x, we obtain the F2F linear programming (LP) relaxation (a 2-factor is a spanning subgraph consisting of disjoint cycles):

$$\widehat{x}_{\mathrm{F2F}} = \arg\max_{x} \langle w, x\rangle \tag{6}$$

$$\text{s.t. } x(\delta(v)) = 2,$$

$$x_e \in [0,1].$$

The main result of the paper is the following. We abbreviate P_n and Q_n as P and Q, respectively.

Theorem 1. *Define*

$$\alpha_n \triangleq -2\log \int \sqrt{\mathrm{d}P\mathrm{d}Q} \tag{7}$$

to be the Rényi divergence of order $\frac{1}{2}$ *between distributions* P *and* Q. *If*

$$\alpha_n - \log n \to +\infty, \tag{8}$$

then the optimal solution of the F2F LP (6) *satisfies* $\min_{x^*\in\mathcal{X}(G)} \cdot \mathbb{P}\{\widehat{x}_{\mathrm{F2F}} = x^*\} \to 1$ *as* $n \to \infty$.

The Rényi divergence of order $\rho > 0$ from P to Q is defined as (Rényi 1961)

$$D_\rho(P\|Q) \triangleq \frac{1}{\rho - 1}\log \int (dP)^\rho (dQ)^{1-\rho}. \tag{9}$$

It particular, for $\rho = 1/2$, it is related to the so-called Battacharyya distance $B(P,Q)$ via $D_{\frac{1}{2}}(P\|Q)=2B(P,Q)$. For Gaussian, Poisson, or Bernoulli weight distribution, the explicit expressions of α_n are given as follows:

$$\alpha_n = \begin{cases} \mu^2/4 & \text{if } P = \mathcal{N}(\mu,1),\ Q = \mathcal{N}(0,1) \\ \left(\sqrt{\lambda} - \sqrt{\mu}\right)^2 & \text{if } P = \mathrm{Pois}(\lambda),\ Q = \mathrm{Pois}(\mu) \\ -2\log\left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right) & \text{if } P = \mathrm{Bern}(p),\ Q = \mathrm{Bern}(q). \end{cases} \tag{10}$$

Although the relaxation from TSP to F2F LP is quite drastic, the resulting algorithm is, in fact, information-theoretically optimal for the hidden Hamiltonian cycle recovery problem. Specifically, under an assumption which can be easily verified for Poisson, Gaussian, or Bernoulli weight distribution, we show in Section 6 that if there exists any algorithm, efficient or not, which exactly recovers x^* with high probability, then it must hold that

$$\alpha_n \geq (1 + o(1)) \log n.$$

This necessary condition, together with sufficient condition (8), implies that the optimal recovery threshold is at

$$\liminf_{n\to\infty} \frac{\alpha_n}{\log n} = 1,$$

achieved by the F2F LP.

We discuss three consequences of Theorem 1. First, as a corollary of the integrality and the optimality of the F2F LP, it can be shown that the max-product belief propagation algorithm introduced in Bayati et al. (2011) can be used to solve the F2F LP exactly, which, for the Gaussian or Poisson weight distribution, requires $o(n^2 \log n)$ iterations (see Section 2 for details). Second, note that we do not require the edge weights to be real valued. Thus the formulation also encompasses the case of partial observation, by letting the weight of every edge in G takes on a special "erasure" symbol with some probability. See Section EC.6 in the e-companion for details. Third, from the optimality of the F2F LP, we show that for the Gaussian or Poisson weight distribution, with high probability, the hidden Hamiltonian tour can be recovered exactly in $O(n^3 \log^3 n)$ time using a packing LP solver (Allen-Zhu and Orecchia 2019) and a simple rounding scheme. See Section EC.8 in the e-companion for details.

In related work, a version of the hidden Hamiltonian cycle model was studied in Broder et al. (1994), where the observed graph is the superposition of a hidden Hamiltonian cycle and an Erdős–Rényi random graph with constant average degree d. Our measurement model is more general than the one in Broder et al. (1994), but more important, the goal in Broder et al. (1994) is not to recover the hidden Hamiltonian cycle but rather to find *any* Hamiltonian cycle in the observed graph, which may not coincide with the hidden one. (In fact, in the regime considered there, exact recovery of the hidden cycle is information-theoretically impossible. See Remark 1 for a justification.) The fractional 2-factor relaxation of TSP has been well studied in the worst case (Dantzig et al. 1954, Boyd and Carr 1999, Schalekamp et al. 2013). It has been shown that under the cost minimization formulation where the costs are symmetric and satisfy the triangle inequality, the *integral gap* of F2F is 4/3; here, the integrality gap is defined as the worst-case ratio of the cost of the optimal integral solution to the cost of the optimal relaxed solution. By contrast, our model does not make any metric assumption on the graph weights.

The rest of this paper is organized as follows. In Section 2, we describe a few other computationally efficient algorithms for the hidden Hamiltonian cycle problem and benchmark their performance against the information-theoretic limit. In Section 3, we discuss related work in more detail. Sections 4 and 5 are devoted to the proof of Theorem 1, and Section 6 characterizes the information-theoretic limit for the recovery problem. In Section EC.1 in the e-companion, we describe the closely related hidden Hamiltonian path problem and show that it can be reduced to and from the hidden Hamiltonian cycle problem both statistically and computationally. Empirical evaluation of various algorithms on both simulated and real DNA data sets are given in Section EC.2 in the e-companion.

2. Performance of Other Algorithms

It is striking to see that the simple F2F LP relaxation of the TSP achieves the optimal recovery threshold in the hidden Hamiltonian cycle model. A natural question to ask is whether there exists another efficient and perhaps even simpler estimator with provable optimality. We have considered various efficient algorithms and derived their performance guarantees. As summarized in Table 1, spectral algorithms is orderwise suboptimal; greedy methods including thresholding achieve the optimal scaling but not the sharp constant. Finally, max-product belief propagation also achieves the sharp threshold as a corollary of our result on the F2F LP. In Table 1,

$$\beta_n \triangleq -\frac{3}{2}\log \int (\mathrm{d}P)^{2/3}(\mathrm{d}Q)^{1/3} \tag{11}$$

is the $\frac{1}{3}$-Rényi divergence from Q to P (see (9)). By Jensen's and Hölder's inequality, for any distinct P and Q, we have

$$\frac{1}{2}\alpha_n < \beta_n < \alpha_n. \tag{12}$$

Table 1. Sufficient Conditions for Various Efficient Algorithms to Achieve Exact Recovery

Efficient algorithms	Performance guarantee
F2F LP Max-product BP	$\alpha_n - \log n \to +\infty$
Greedy merging	$\beta_n - \log n \to +\infty$
Simple thresholding Nearest neighbor	$\alpha_n - 2\log n \to +\infty$
Spectral methods	$\alpha_n \gg n^5$ (Gaussian)

For Gaussian weights with $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$, we have $\beta_n = \frac{1}{6}\mu^2 = \frac{2}{3}\alpha_n$. Simulation of these algorithms confirm these theoretical results. See Figure EC.1 in Section EC.2.1 of the e-companion.

2.1. Spectral Methods

Spectral algorithms are powerful methods for recovering the underlying structure in planted models based on the principal eigenvectors of the observed adjacency matrix A. Under planted models such as planted clique (Alon et al. 1998) or planted partition models (McSherry 2001), spectral algorithms and their variants have been shown to achieve either the optimal recovery thresholds (Massoulié 2013, Abbe and Sandon 2015, Bordenave et al. 2015) or the best possible performance within certain relaxation hierarchies (Deshpande and Montanari 2015, Meka et al. 2015, Barak et al. 2016). The rationale behind spectral algorithms is that the principal eigenvectors of $\mathbb{E}[A]$ contain information about underlying structures and the principal eigenvectors of A are close to those of $\mathbb{E}[A]$, provided that the spectral gap (the gap between the largest few eigenvalues and the rest of them) is much larger than the spectral norm of the perturbation $\|A - \mathbb{E}[A]\|$. In our setting, indeed, the principal eigenvectors of $\mathbb{E}[A]$ contain information about the ground truth Hamiltonian cycle C^*. To see this, let us consider the Gaussian case where $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$ as an illustrating example. Then the observed matrix can be expressed as

$$A = \mu C^* + Z,$$

where, with a slight abuse of notation, we use A to denote the weighted adjacency matrix of G and C^* to denote the adjacency matrix of the true Hamiltonian cycle; Z is a symmetric Gaussian matrix with zero diagonal and $Z_{ij} = Z_{ji}$ independently drawn from $\mathcal{N}(0, 1)$ for $i < j$. Because C^* is a circulant matrix, its eigenvalues and the corresponding eigenvectors can be explicitly derived via discrete Fourier transform. It turns out that the eigenvector corresponding to the second-largest eigenvalue of C^* contains perfect information about the true Hamiltonian cycle. Unfortunately, in contrast to the planted clique and planted partition models under which $\mathbb{E}[A]$ is low rank and has a large eigengap, here, C^* is full rank, and the gap between the second- and third-largest eigenvalue is on the order of $1/n^2$, which is much smaller than $\|Z\| = \Theta(\sqrt{n})$. Therefore, for spectral algorithms to succeed, a very high signal level $\mu^2 \gg n^5$ is required. This agrees with the empirical performance on simulated data and is highly suboptimal compared with the sufficient condition (8) of F2F LP: $\mu^2 - 4\log n \to \infty$.

2.2. Greedy Methods

To recover the hidden Hamiltonian cycle, we can also resort to greedy methods. It turns out that the following simple thresholding algorithm achieves the optimal recovery threshold (8) within a factor of 2: for each vertex, keep the two incident edges with the two largest weights and delete the other $n - 3$ edges. The resulting graph has degree at most 2. It can be shown that the resulting graph coincides with C^* with high probability provided that $\alpha_n - 2\log n \to +\infty$.

Another well-known greedy heuristic is the following nearest-neighbor algorithm. Start on an arbitrary vertex as the current vertex and find the edge with the largest weight connecting the current vertex and an unvisited vertex v; set the current vertex to v and mark v as visited. Repeat until all vertices have been visited. Let $v_1, \ldots, v_n$ denote the sequence of visited vertices and output the Hamiltonian cycle formed by $(v_1, \ldots, v_n, v_1)$. It can be shown (see Section EC.5 in the e-companion) that the resulting Hamiltonian cycle coincides with C^* with high probability provided that $\alpha_n - 2\log n \to +\infty$.

Finally, we consider a greedy merging algorithm proposed in Motahari et al. (2013): connect pairs of vertices with the largest edge weights until all vertices have degree 2. The output is a 2-factor, and it can be shown that the output 2-factor coincides with C^* with high probability provided that $\beta_n - \log n \to +\infty$, strictly improving on the performance guarantee of previous two greedy algorithms.

Notice that the aforementioned greedy algorithms only exploit local information and do not take into account the global cycle structure. Naturally, none of them achieves the optimal threshold (8). See Section EC.5 in the e-companion for further details.

2.3. Max-Product Belief Propagation

We can improve on the simple thresholding algorithm using an iterative message-passing algorithm known as max-product belief propagation. Specifically, at each time $t = 0, 1, \ldots, t_f$, each vertex i sends a real-valued message $m_{i\to j}(t)$ to each of its neighbors j. Messages are initialized by $m_{i\to j}(0) = w_e$ for all $e = (i, j)$. For $t \geq 1$, messages transmitted by vertex i in iteration t are updated based on messages received in iteration $t - 1$ recursively as follows:

$$m_{i\to j}(t) = w_e - \operatorname{2nd}\max_{\ell \neq j}\{m_{\ell\to i}(t-1)\},$$

where 2nd max denotes the second-largest value. At the end of the final iteration t_f, for every vertex, keep the two incident edges with the two largest received message values and delete the other $n - 3$ edges, and output the resulting graph. Note that belief propagation (BP) with one iteration $t_f = 1$ reduces to the simple thresholding algorithm.

The belief propagation algorithm is studied in Bayati et al. (2011) to find the b-factor with the maximum weight for $b \geq 1$; it is shown that if the fractional b-factor LP relaxation has no strictly fractional optimum solution, then the output of BP coincides with the optimal b-factor when $t_f \geq \lceil 2nw^*/\epsilon \rceil$, where w^* is the weight of the optimal b-factor and ϵ is the difference between the weight of the optimal b-factor and the second-largest weight of b-factors. Our optimality result of F2F implies that if $\alpha_n - \log n \to +\infty$, then with high probability, F2F has no fractional optimum solution, and the optimal 2-factor coincides with the ground truth x^*; Therefore, by combining our result with results of BP in Bayati et al. (2011), we immediately conclude that the output of BP coincides with x^* with high probability after t_f iterations, provided that $\alpha_n - \log n \to +\infty$. For both the Gaussian and Poisson models, with high probability, the number of iterations t_f of the BP algorithm is, in fact, $o(n^2 \log n)$, nearly linear in the problem size (see Section EC.9 in the e-companion for a justification).

3. Related Work

We discuss additional related work before presenting the proof of our main results. Because of the NP-hardness of TSP, researchers have imposed structural assumptions on the costs (weights) and devised efficient approximation algorithms. One natural assumption is the metric assumption under which the costs are symmetric ($c_{ij} = c_{ji}$ for all $i, j \in V$) and satisfy the triangle inequality ($c_{ik} \leq c_{ij} + c_{jk}$ for all $i, j, k \in V$). Metric TSP turns out to be still NP-hard, as shown by reduction from the NP-hard Hamiltonian cycle problem (Schrijver 2003, theorem 58.1). The best approximation algorithm for metric TSP currently known is Christofides' algorithm, which finds a Hamiltonian cycle of cost at most a factor of 3/2 times the cost of an optimal Hamiltonian cycle.

3.1. Integrality Gap of LP Relaxations of TSP

Various relaxations of TSP has also been extensively studied under the metric assumption. To measure the tightness of LP relaxations, a commonly used figure of merit is the *integrality gap*. As is the convention in the TSP literature, the optimization is formulated as a minimization problem with nonnegative costs. In general, the integrality gap is defined as the supremum of the ratio $\mathsf{OPT}/\mathsf{FRAC}$, over all instances of the problem, where FRAC denotes the objective value of the optimal fractional solution, and OPT denotes the objective value of the optimal integral solution (Chlamtác and Tulsiani 2012). Note that, by definition, the integrality gap is always at least 1. Dropping the integer constraints (5) in ILP formulation of TSP (2) leads to a LP relaxation known as *subtour LP* (Dantzig et al. 1954, Held and Karp 1970). The integrality gap of the subtour LP is known to be between 4/3 and 3/2. The integrality gap of fractional 2-factor LP (6) is shown in Boyd and Carr (1999) and Schalekamp et al. (2013) to be 4/3. In contrast to the previous worst-case approximation results on metric TSP, this paper focuses on a planted instance of TSP, where we impose probabilistic assumption on the costs (weights), and the goal is to recover the hidden Hamiltonian cycle. In particular, the metric assumption is not fulfilled in our hidden Hamiltonian cycle model, and hence the previous results do not apply. Our results imply that when $\alpha_n - \log n \to +\infty$, the optimal solution of F2F coincides with the optimal solution of TSP with probability tending to 1, where the probability is taken over the randomness of weights w in the hidden Hamiltonian cycle model. In other words, for "typical" instances of the hidden Hamiltonian cycle model, the optimal objective value of TSP is the same as that of F2F.

3.2. SDP Relaxations of TSP

Semidefinite programming (SDP) relaxations of the traveling salesman problem have also been extensively studied in the literature. A classical SDP relaxation of TSP due to Cvetković et al. (1999) is obtained by imposing an extra constraint on the second-largest eigenvalue of a Hamiltonian cycle in F2F LP (6). A more sophisticated SDP relaxation is derived in Zhao et al. (1998) by viewing the TSP as a quadratic assignment problem, from which one can obtain a simpler SDP relaxation of TSP based on association schemes (De Klerk et al. 2008). This SDP relaxation in De Klerk et al. (2008) is shown to dominate that of Cvetković et al. (1999). Because all these SDP relaxations are tighter than the F2F LP, our results immediately imply that the optimal solutions of these SDP relaxations coincide with the true Hamiltonian cycle x^* with high probability provided $\alpha_n - \log n \to +\infty$.

3.3. Data Seriation

The problem of recovering a hidden Hamiltonian cycle (path) in a weighted complete graph falls into a general problem known as data seriation (Kendall 1971) or data stringing (Chen et al. 2011). In particular, we are given a similarity matrix Y for n objects, and we are interested in seriating or stringing the data by ordering the n objects so that similar objects i and j are near each other. Data seriation has diverse applications ranging from data visualization and DNA sequencing to functional data analysis (Chen et al. 2011) and archaeological dating (Robinson 1951). Most previous work on data seriation focuses on the noiseless case (Robinson 1951, Kendall 1971), where there is an unknown ordering of n objects so that if object j is closer than object k to object i in the ordering, then $Y_{ij} \geq Y_{ik}$ (i.e., the similarity between i and j is

always no less than the similarity between i and k). Such a matrix Y is called *Robinson matrix*. It is shown in Atkins et al. (1998) that one can recover the underlying true ordering of objects up to a global shift by component-wisely sorting the second eigenvector of the Laplacian matrix associated with Y if Y is a Robinson matrix. The data seriation problem has also been formulated as a quadratic assignment problem and convex relaxations are derived (see Fogel et al. 2013, Lim and Wright 2014, and references therein).

One interesting generalization of the hidden Hamiltonian cycle model is to extend the hidden structure from a cycle to k-regular graph for general $k \geq 2$ (e.g., nearest-neighbor graphs), which can potentially better fit the genome assembly data. The underlying k-regular graph can represent the hidden geometric structure, and the observed graph can be viewed as a realization of the Watts–Strogatz small-world graph (Watts and Strogatz 1998) if the weight distribution is Bernoulli. Recent work (Cai et al. 2017) has studied the problem of detecting and recovering the underlying k-regular graph under the small-world graph model and derived conditions for reliable detection and recovery; however, the information limit and the optimal algorithm remain open.

Finally, we mention that the Rényi divergence of order 1/2 also plays a key role in determining the exact recovery threshold for community detection under the stochastic block models (Abbe and Sandon 2015, Jog and Loh 2015, Mossel et al. 2015, Abbe et al. 2016, Zhang and Zhou 2016).

4. Proof Techniques and a Simpler Result

The proof of the main result, Theorem 1, is quite involved. In this section, we will discuss the high-level ideas and the difference with the conventional proof using dual certificates. As a warm-up, we also prove a weaker version of the result on the 2-factor ILP. The proof of the full result is given in Section 5.

4.1. Proof Techniques

A standard technique for analyzing convex relaxations is the *dual certificate argument*, which amounts to constructing the dual variables so that the desired Karush-Kuhn-Tucker conditions are satisfied for the primal variable corresponding to the ground truth. This type of argument has been widely used, for instance, for proving the optimality of SDP relaxations for community detection under stochastic block models (Abbe et al. 2016; Hajek et al. 2016a, b, c; Agarwal et al. 2017; Perry and Wein 2017; Bandeira 2018). However, for the F2F LP (6), we were only able to find explicit constructions of dual certificates that attain the optimal threshold within a factor of 2. Instead, the proof of Theorem 1 is by means of a direct *primal argument*, which shows that with high probability no other vertices of the F2F polytope has a higher objective value than that of the ground truth. Nevertheless, it is still instructive to describe this dual construction before explaining the ideas of the primal proof.

To certify the optimality of x^* for F2F LP, it reduces to constructing a dual variable $u \in \mathbb{R}^n$ (corresponding to the degree constraints) such that for every edge (i, j),

$$u_i + u_j \leq w_{ij}, \quad \text{if } x^*_{ij} = 1, \tag{13}$$

$$u_i + u_j \geq w_{ij}, \quad \text{if } x^*_{ij} = 0. \tag{14}$$

A simple choice of u is

$$u_i = \frac{1}{2}\min_j \{w_{ij} : x^*_{ij} = 1\}. \tag{15}$$

Then (13) is fulfilled automatically, and (14) can be shown (see Section EC.4 in the e-companion) to hold with high probability provided that

$$\beta_n - \log n \to +\infty. \tag{16}$$

where β_n is the $\frac{1}{3}$-Rényi divergence defined in (11). Because $\alpha_n/2 < \beta_n < \alpha_n$ by (12), this construction shows that F2F achieves the optimal recovery threshold by at most a multiplicative factor of 2. For specific distributions, this factor-of-two gap can be further improved (e.g., to $\frac{3}{2}$ for Gaussian weights), for which we have $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$ and $\beta = \frac{1}{6}\mu^2 = \frac{2}{3}\alpha$. However, this certificate does not get us all the way to the information limit (8).

Departing from the usual dual certificate argument, our proof of the optimality of F2F relaxation relies on delicate primal analysis. In particular, we show that $\langle w, x - x^* \rangle < 0$ for any vertex (extremal point) of the F2F polytope $x \neq x^*$ with high probability. It is known that the F2F polytope is not integral in the sense that some of its vertices is fractional. Fortunately, it turns out that for any vertex x, its fractional entry x_e must be 1/2. Thanks to this *half-integrality* property, we can encode the difference $y \triangleq 2(x - x^*)$ as a bicolored multigraph G_y with a total weight $w(G_y) = 2\langle w, x - x^* \rangle$. Finally, we bound $w(G_y)$ via a *divide-and-conquer* argument by first decomposing G_y into an edge-disjoint union of graphs in a family with simpler structures and then proving that for every graph H in this family, its total weight $w(H)$ is negative with high probability under condition (8). Our decomposition of G_y heavily exploits the fact that G_y is a balanced multigraph in the sense that every vertex has an equal number of incident red edges and blue edges, and the classical graph-theoretic result that every connected balanced multigraph has an Eulerian circuit with edges alternating in colors.

4.2. 2-Factor Integer Linear Programming Relaxation

The 2-factor (2F) integer linear programming relaxation of the TSP is

$$\widehat{x}_{2F} = \arg\max_{x} \langle w, x\rangle \quad \text{s.t. } x(\delta(v)) = 2,\ x_e \in \{0,1\}. \tag{17}$$

The 2-factor ILP is the same as the F2F LP (6) except that the x_e's have integrality constraints and is therefore a tighter relaxation of the original TSP than F2F LP. As a warm-up for the optimality proof of F2F LP, we provide a much simpler proof, showing that the optimal solution of the 2F ILP coincides with the true cycle x^* with high probability, under the same condition that $\alpha_n - \log n \to +\infty$. We note that although it is not an LP, the 2F ILP is shown in Letchford et al. (2008) to be solvable in $O(n^2 m \log(n^2/m)) = O(n^4)$ time using a variant of the blossom algorithm (Edmonds 1965a, b), where $n = |V(G)|$ and $m = |E(G)|$. See Section EC.7 in the e-companion for detailed discussions on the time complexity of 2F ILP.

Let x denote the adjacency vector of a given 2-factor. To prove that x^* is the unique optimal solution to the 2F ILP, it suffices to show $\langle w, x - x^*\rangle < 0$ for the adjacency vector of any 2-factor $x \neq x^*$. To capture the difference between x and x^*, we define $y \in \{0, \pm 1\}^{\binom{n}{2}}$ by

$$y = x - x^*. \tag{18}$$

Define a simple graph G_y with bicolored edge whose adjacency matrix is $|y|$ with isolated vertices removed and each edge is colored red if $y_e = -1$ and blue if $y_e = +1$ (see Figure 3 for an example). Furthermore, for a given bicolored graph B, we define its weight as

$$w(B) \triangleq \sum_{\text{blue } e \in E(B)} w_e - \sum_{\text{red } e \in E(B)} w_e.$$

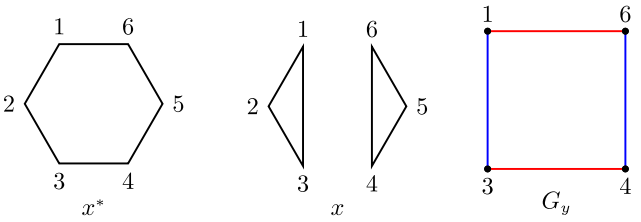
Then $w(G_y) = \langle w, x - x^*\rangle$.

A bicolored graph is *balanced* if for every vertex the number of red incident edges is equal to the number of blue incident edges. Because $x(\delta(v)) = 2$ and $x^*(\delta(v)) = 2$ for every vertex v, it follows that $y(\delta(v)) = 0$, and thus G_y is balanced. Define

$$\mathcal{B} = \{B : B \text{ is a simple, connected, and balanced bicolored graph}\},$$
$$\mathcal{B}^* = \{B \in \mathcal{B} : V(B) \subset [n], x^*_e = 1 \text{ for every red edge } e \in E(B)\},$$

where $V(B)$ and $E(B)$ denote the vertex set and edge set of B, respectively.

Figure 3. (Color online) The Ground Truth x^* Is a Cycle of Length 6, x Is a Feasible Solution to (17) Corresponding to Two Disjoint Triangles, and the Graph G_y for $y = x - x^*$ Is an Alternating 4-Cycle



Let $B_1, \ldots, B_m$ denote the connected components of G_y. Because each connected component of G_y is balanced, it follows that $B_i \in \mathcal{B}^*$, and

$$w(G_y) = \sum_{i=1}^{m} w(B_i).$$

Hence, to show $w(G_y) < 0$ for all possible G_y, it reduces to proving that $w(B) < 0$ for all $B \in \mathcal{B}^*$.

Fix an even integer $\ell \geq 4$, and let

$$\mathcal{B}^*_\ell = \{B \in \mathcal{B}^* : |E(B)| = \ell\}.$$

Fix any $B \in \mathcal{B}^*_\ell$. By the balancedness, B has $\ell/2$ red edges and $\ell/2$ blue edges. Hence,

$$w(B) \overset{d}{=} \sum_{i=1}^{\ell/2} Y_i - \sum_{i=1}^{\ell/2} X_i,$$

where X_i's and Y_i's are independent sequences of random variables such that X_i's are independent and identically distributed (i.i.d.) copies of $\log(\mathrm{d}P/\mathrm{d}Q)$ under distribution P, and Y_i's are i.i.d. copies of $\log(\mathrm{d}P/\mathrm{d}Q)$ under distribution Q; the notation $\overset{d}{=}$ denotes equality in distribution. It follows from Chernoff's inequality (see the large-deviation bound (EC.2) in Section EC.3 in the e-companion) that

$$\mathbb{P}\{w(B) \geq 0\} = \mathbb{P}\left\{\sum_{i=1}^{\ell/2} Y_i - \sum_{i=1}^{\ell/2} X_i \geq 0\right\} \leq \exp(-\alpha_n \ell/2). \tag{19}$$

Next we claim that there are at most $(2n)^{\ell/2}$ different graphs B in $\mathcal{B}^*_\ell$. The proof of the claim is deferred to the end of this section. Combining the union bound with (19) gives that

$$\mathbb{P}\left\{\max_{B \in \mathcal{B}^*_\ell} w(B) \geq 0\right\} \leq |\mathcal{B}^*_\ell| \exp(-\alpha_n \ell/2) \leq \exp\{-(\alpha_n - \log(2n))\ell/2\}.$$

Taking another union bound over all integers $\ell \geq 4$, we get the desired result:

$$\begin{aligned}\mathbb{P}\left\{\max_{B\in\mathcal{B}^*} w(B) \geq 0\right\} &\leq \sum_{\ell=4}^{\infty} \mathbb{P}\left\{\max_{B\in\mathcal{B}^*_\ell} w(B) \geq 0\right\} \\ &\leq \sum_{\ell=4}^{\infty} \exp\{-(\alpha_n - \log(2n))\ell/2\} \\ &\leq \frac{\exp\{-2(\alpha_n - \log(2n))\}}{1-\exp\{-(\alpha_n - \log(2n))/2\}} \overset{(8)}{\to} 0.\end{aligned}$$

We are left to show that $|\mathcal{B}^*_\ell| \leq (2n)^{\ell/2}$. This follows from the following classical graph-theoretic result that every connected balanced multigraph G has an *alternating* Eulerian circuit—that is, the edges in the circuit alternate in color.

Lemma 1. *Every connected balanced bicolored multigraph G has an alternating Eulerian circuit.*

The lemma is proved in Kotzig (1968, theorem 1) in a more general form (see also Pevzner 1995, corollary 1). For completeness, we provide a short proof in Section EC.10.1 in the e-companion.

In view of Lemma 1, for every $B \in \mathcal{B}^*_\ell$, it must have a Eulerian circuit T given by the sequence of $(v_0, v_1, \ldots, v_{\ell-1}, v_\ell = v_0)$ of vertices (vertices may repeat) such that $v_i \in [n]$, and (v_i, v_{i+1}) is a red edge for even i and blue edge for odd i in B. Let $\mathcal{T}$ denote the set of all possible such Eulerian circuits. Moreover, every Eulerian circuit $T \in \mathcal{T}$ uniquely determines a $B \in \mathcal{B}^*_\ell$, because the vertex set $V(B)$ is the union of vertices v_i's, and the colored edge set $E(B)$ is the union of colored edges (v_i, v_{i+1})'s in T. Hence, $|\mathcal{B}^*_\ell| \leq |\mathcal{T}|$. To enumerate all possible $T \in \mathcal{T}$, it suffices to enumerate all the possible labelings of vertices in T. Recall that, by definition, for every red edge $e = (v_i, v_{i+1})$ in T, $x^*_e = 1$. Thus the two endpoints v_i and v_{i+1} must be neighbors in the true cycle corresponding to x^*, and hence once the vertex labeling of v_i is fixed, there are at most two different choices for the vertex labeling of v_{i+1}. Therefore, we enumerate all possible Eulerian circuits $T \in \mathcal{T}$ by sequentially choose the vertex labeling of v_i from $i = 0$ to $i = \ell - 1$. Given the vertex labelings of $(v_0, \ldots, v_{i-1})$, the number of choices of the vertex labeling of v_i is at most n for even i and 2 for odd i. Hence, $|\mathcal{T}| \leq (2n)^{\ell/2}$, which further implies that $|\mathcal{B}^*_\ell| \leq (2n)^{\ell/2}$.

5. Proof of Theorem 1

In this section, we prove that the optimal solution of the fractional 2-factor coincides with x^* with high probability, provided that $\alpha_n - \log n \to \infty$. This is the bulk of the paper.

5.1. Graph Notations

We describe several key graph-theoretic notations used in the proof. We start with multigraphs. Formally, a multigraph G is an ordered pair (V, E) with a vertex set $V = V(G)$ and an edge *multiset* $E = E(G)$ consisting of subsets of $V(G)$ of size 2. Note that, by definition, multigraphs do not have self-loops. A multiedge is a set of edges in $E(G)$ with the same endpoints. The multiplicity of an edge is its multiplicity as an element in $E(G)$. We call a multiedge single and double if its edge multiplicity is 1 and 2, respectively. Note that a double edge (u, v) refers to the set of two edges connecting vertices u and v. We say a multigraph G is bicolored if every distinct element in $E(G)$ is colored in either red or blue, and the repeated copies of an element all have the same color.

For two multigraphs G and H on the same set of vertices, we define $G - H$ to be the multigraph induced by the edge multiset $E(G) \setminus E(H)$. The union of multigraphs G and H is the multigraph $G \cup H$ with vertex set $V(G) \cup V(H)$ and edge multiset $E(G) \cup E(H)$. Note that here the union of multisets is defined so that the multiplicity of each of the elements adds up. For example, $\{a, a, b\} \cup \{a, b, c\} = \{a, a, a, b, b, c\}$. By definition, the multiplicity of an element in $E(G) \cup E(H)$ is the sum of its multiplicity in $E(G)$ and $E(H)$. When $E(G) \cap E(H) = \emptyset$, $G \cup H$ is called an edge-disjoint union. When $V(G) \cap V(H) = \emptyset$, $G \cup H$ is called an vertex-disjoint union.

A *walk* in a multigraph G is a sequence $(v_0, v_1, \ldots, v_m)$ of vertices (which may repeat) such that $(v_{i-1}, v_i) \in E(G)$ for $1 \leq i \leq m$. A *trail* in a multigraph G is a walk $(v_0, v_1, \ldots, v_m)$ such that for all $1 \leq i \leq m$, the number of times that edge (v_{i-1}, v_i) appears in the walk is no more than its edge multiplicity in $E(G)$. A trail is *closed* if the starting and ending vertices are the same. A *circuit* is a closed trail. An *Eulerian trail* in a multigraph G is a trail $(v_0, v_1, \ldots, v_m)$ such that for every $e \in E(G)$, the number of times that it appears in the trail coincides with its edge multiplicity in $E(G)$. An *Eulerian circuit* is a closed Eulerian trail. A *path* is a trail with no repeated vertex. A *cycle* consists a path plus an edge from its last vertex to the first.

5.2. Proof Outline

Let $x^* = (x^*_e)$ denote the adjacency vector of the hidden Hamiltonian cycle (ground truth). The feasible set of the F2F LP (6) is the *F2F polytope*:

$$Q \triangleq \left\{x \in [0,1]^{\binom{n}{2}} : x(\delta(v)) = 2, \forall v \in [n]\right\}. \tag{20}$$

To prove that x^* is the unique optimal solution to the F2F LP with high probability, it suffices to show that $\langle w, x - x^*\rangle < 0$ holds with high probability for any vertex (extremal point) of the F2F polytope x other than x^*. It turns out that the vertices of the F2F polytope Q have the following simple characterization (Balinski 1965, Boyd and Carr 1999, Schalekamp et al. 2013).

First of all, for any vertex x, its fractional entry must be a half-integer; that is,

$$x_e \in \{0, 1/2, 1\}, \quad \forall e. \tag{21}$$

Furthermore, if we define the *support graph* of x as the graph with vertex set $[n]$ and edge set $\{e : x_e \neq 0\}$, then each connected component of the support graph of x must be one of the following two cases: it is either

1. a cycle of at least three vertices with $x_e = 1$ for all edges e in the cycle, or
2. consisting of an even number of odd-sized cycles with $x_e = 1/2$ for all edges e in the cycles that are connected by paths of edges e with $x_e = 1$. In this case, if we remove the edges in the odd cycles, the resulting graph is a spanning disjoint set of paths formed by edges e with $x_e = 1$.

See Figure 4(a) for a graphical representation when $n = 6$. It turns out that, among the aforementioned characterizations of the vertices of the F2F polytope, our analysis of the LP relaxation uses only the half-integrality property (21).

To capture the difference between a given vertex x and the true solution x^*, we use the following multigraph representation: define $y \in \mathbb{R}^{\binom{n}{2}}$ by

$$y = 2(x - x^*), \tag{22}$$

with $y_e = 2(x_e - x_e^*) \in \{0, \pm 1, \pm 2\}$ (see Figure 4(b)). Define a multigraph G_y whose adjacency matrix is $|y|$ with isolated vertices removed and each edge e is colored red if $y_e < 0$ or blue if $y_e > 0$. In particular, the edge multiplicity of G_y is at most 2. Compared with (18), the extra factor of 2 in (22) is to ensure that y is still integral; as a consequence, G_y may be a multigraph with multiplicity 2 instead of a simple graph. For any given bicolored multigraph F, we define its weight as

$$w(F) \triangleq \sum_{\text{blue } e \in E(F)} w_e - \sum_{\text{red } e \in E(F)} w_e,$$

where the summation above includes all repeated copies of e in $E(F)$. Then $w(G_y) = \langle w, x - x^* \rangle$. Hence, to prove that x^* is the unique optimal solution to the LP program with high probability, it reduces to showing that $w(G_y) < 0$ for all possible G_y constructed from the extremal point $x \neq x^*$ with high probability.

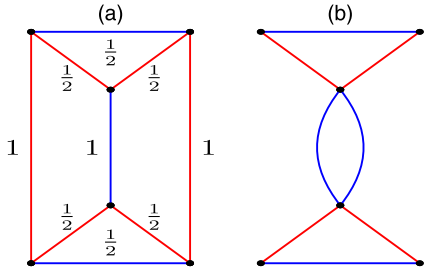
Instead of first calculating the probability of $w(G_y) \leq 0$ and then taking a union bound on all possible G_y, our proof crucially relies on a decomposition of G_y into some suitably defined simpler graphs. In the next subsection, we will describe a family $\mathcal{F}^*$ of graphs and show that every possible G_y can be decomposed as a union of graphs in $\mathcal{F}^*$: $G_y = \cup_{i=1}^m F_i$ with $F_i \in \mathcal{F}^*$ for each $1 \leq i \leq m$. Because the multiplicity of e in $E(G_y)$ is equal to the sum of multiplicities of e in $E(F_i)$ over $i \in [m]$, it follows that

$$w(G_y) = \sum_{i=1}^m w(F_i).$$

Therefore, to show that $w(G_y) < 0$ for all possible G_y with high probability, it suffices to show $w(F) < 0$ for all graphs F in family $\mathcal{F}^*$.

We remark that in the analysis of the 2F ILP in Section 4.2, we have $y = x - x^*$ as opposed to $y = 2(x - x^*)$, and thus G_y is a balanced simple graph. Consequently, we can simply decompose G_y into its connected components, which are connected, balanced simple graphs. By contrast, here, the decomposition of G_y as a multigraph is much more sophisticated because of the existence of double edges. In particular, the weight of a double edge in F appears twice in $w(F)$, and hence its variance is twice the total variance of two independent edge weights. For this reason, to control the deviation of $w(F)$ from its mean, it is essential to account for the contribution of double edges and single edges separately, which, in turn, requires us to separate the double edges from single edges in our decomposition.

Figure 4. (Color online) (a) The Support Graph of a Fractional Vertex x of the F2F Polytope with $n = 6$ (the Edges in the Support Graph of x^* Are Highlighted in Red); (b) the Multigraph Representation of $y = 2(x - x^*)$



5.3. Edge Decomposition

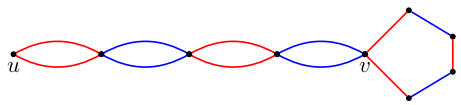
Our decomposition of G_y relies on the notion of balanced multigraph and alternating Eulerian circuit. A bicolored multigraph is *balanced* if for every vertex the number of red incident edges is equal to the number of blue incident edges. Because $x(\delta(v)) = 2$ and $x^*(\delta(v)) = 2$ for every vertex v, it follows that $y(\delta(v)) = 0$, and thus G_y is balanced. As a result, the vertices in G_y all have even degrees (in fact, either 2, 4, 6, or 8). Therefore each connected component of G_y has an Eulerian circuit. Recall that Eulerian circuit is alternating if the edges in the Eulerian circuit alternate in color. In view of Lemma 1, each connected component of G_y has an Eulerian circuit. In the remainder, we suppress the subscript y in G_y whenever the context is clear.

Next, we describe a family $\mathcal{F}$ of graphs and show that G is a union of graphs in this family. First, we

need to introduce a few notations. For any pair of two vertices u, v in graph G, vertex identification (also known as vertex contraction) produces a graph by removing all edges between u, v and replacing u, v with a single vertex w incident to all edges formerly incident to either u or v. When u and v are adjacent (i.e., sharing two endpoints of edge e), vertex identification specializes to the edge contraction of e, and the resulting graph is denoted by $G \cdot e$; visually, e shrinks to a vertex. Note that edge contraction may introduce multiedges. We define a *stem* as a path $(v_0, v_1, \ldots, v_{k-1})$ for some k distinct vertices such that (v_{i-1}, v_i) is a double edge for all $1 \le i \le k$ and the double edges alternate in color. The two endpoints v_0 and v_{k-1} of the stem are identified as the tips of the stem. We say a tip of the stem is red if it is incident to the red double edge; otherwise, we say it is blue. Given a stem and an even cycle C_0 consisting of only single edges of alternating colors, we define the following *blossoming* procedure to connect the stem with C_0: first contract any single blue (red) edge in C_0 to a vertex v and attached to v the stem by identifying v with a blue (red) tip of the stem. The resulting graph known as a *flower* has an alternating circuit, and the contracted C_0 is called a *blossom*. The tip of the stem not incident to the blossom is called the tip of the flower. We say a flower is red (blue) if its tip is red (blue). For example, a red flower is shown in Figure 5. Similar notions of the stem, flower, and blossom were introduced in Edmonds (1965b) in the context of simple graphs.

Then we introduce a family $\mathcal{U}$ of balanced graphs. We start with an even cycle G_0 in alternating colors. At each step $t \ge 1$, construct a new balanced graph G_t from G_{t-1} as follows. Fix any cycle consisting of at least four edges in G_{t-1}. In this cycle, pick any edge and apply the following *flowering* procedure: contract the red (blue) edge to one vertex w and attach to w a flower by identifying w with the root of a blue (red) flower. Because we allow contracting an edge incident to a stem, it is possible to have a vertex with multiple stems attached. Let $\mathcal{U}$ denote the collection of all graphs obtained from applying the flowering procedure recursively for finitely many times. In particular, $\mathcal{U}$ includes all even cycles in alternating colors. For example, the graph in Figure 6 is in $\mathcal{U}$, which is obtained by starting with a 10-cycle and applying the flowering procedure four times. By construction, any graph $H \in \mathcal{U}$ must contain an even $\ell \ge 4$ number of single edges.

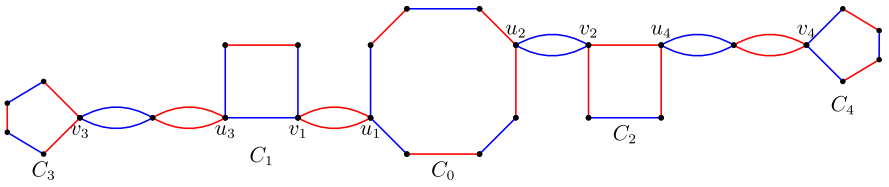
Figure 5. (Color online) A Red Flower Consisting of a Stem of Four Alternating Double Edges Followed by a Blossom of Five Single Edges



Alternatively, note that each graph in the family $\mathcal{U}$ can be viewed as cycles interconnected by stems. Thus, we can represent G_t using a tree T_t, whose nodes correspond to even cycles and links correspond to stems (see Figure 7 for a tree representation of the graph in Figure 6). Next we describe this enumeration scheme in detail. Every node in the tree represents a cycle in alternating colors, with a mark ℓ being the length of the cycle. The cycle corresponding to the root node is assumed to have a fixed ordering of edges, and the root node has an extra mark that is 1 if the color of the first edge in the corresponding cycle is blue and 0 otherwise. Every link (u, v) represents a stem consisting of only double edges of alternating colors with mark (k, i), where k is the length of the stem, and i is the index of the contracted edge of the parent vertex u. We start with tree T_0 with a single root node corresponding to $G_0 = C_0$, with the mark being the length of C_0. At each step $t \ge 1$, we view the flowering procedure as growing to a new tree T_t as follows. For any vertex u in T_{t-1} that corresponds to an alternating cycle C, a new vertex v that corresponds to an alternating cycle C', and a stem, we connect u and v with an edge corresponding to the stem. The edge is marked with (k, i), where k is the length of the stem, and i is the index of the contracted edge in C. Then the edges in the alternating cycle C' are indexed by $1, 2, \ldots$ by starting from the contracted edge in C' and traversing C' in a clockwise direction.

Finally, we need to introduce the notion of homomorphism between two bicolored multigraphs, H and F. There exist multiple definitions of homomorphism between multigraphs; here, we follow the convention in Lovász (2012, section 5.2.1). Let A and B denote two multisets. Let A' and B' denote the set of distinct elements in A and B, respectively. We say $\psi : A \to B$ is bijective if $\psi : A' \to B'$ is bijective, and for every element $a \in A'$, the multiplicity of a in A is the same as the multiplicity of $\psi(a)$ in B. For example, if $A = \{a, a, b, c\}$ and $B = \{x, x, y, z\}$, let $\psi(a) = x$, $\psi(b) = y$, and $\psi(c) = z$; then $\psi : A \to B$ is bijective. A node-and-edge homomorphism $H \to F$ is a vertex map $\phi : V(H) \to V(F)$ and *bijective* edge map $\psi : E(H) \to E(F)$ pair such that if $e \in E(H)$ connects i and j, then $\psi(e)$ connects $\phi(i)$ and $\phi(j)$ and has the same color as e. We say H is homomorphic to F if such a node-to-edge homomorphism exists. By construction, an edge e is incident to u in H if and only if $\psi(e)$ is incident to $\phi(u)$ in F. Therefore, if H is homomorphic to F, then they are either both balanced or both unbalanced. Moreover, because ψ is bijective, $H \to F$ is edge-multiplicity preserving; that is, the multiplicity of $\psi(e)$ in $E(F)$ is the same as the multiplicity of e in $E(H)$. Hence, the number of double (single) edges in H and F is the

Figure 6. (Color online) Example of Graph in Family $\mathcal{U}$



same. Furthermore, note that for two node-and-edge homomorphisms $(\phi,\psi): H \to F$ and $(\phi,\psi'): H \to F'$ with the same vertex map ϕ, it holds that $F = F'$. Hence, when the context is clear, we simply write $\phi: H \to F$ or $\phi(H) = F$ by suppressing the underlying edge map.

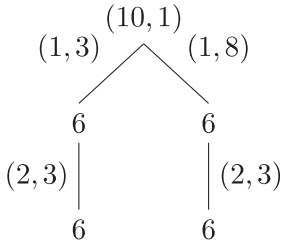
Let $\mathcal{F}$ denote the collection of all graphs F such that $H \to F$ for some $H \in \mathcal{U}$. In particular, $\mathcal{F} \supseteq \mathcal{U}$, and this inclusion is as strict as the example in Figure 8 shows. The next lemma shows that $\mathcal{F}$ includes all connected balanced simple graphs. This result serves as the base case of the induction proof of the decomposition lemma.

Lemma 2. *An alternating cycle is homomorphic to any connected balanced simple graph G with an equal number of edges. In particular, $G \in \mathcal{F}$.*

Proof. By Lemma 1, G has an Eulerian circuit $T = (v_0, v_1, \ldots, v_{m-1}, v_m = v_0)$ of alternating colors where m is the total number of edges in G and vertices v_i's may repeat. Let C denote any alternating cycle with m edges. We write $C = (u_0, u_1, \ldots, u_{m-1}, u_m = u_0)$ such that the edge (u_0, u_1) has the same color as (v_0, v_1). Then we define a vertex and edge map pair (ϕ,ψ) from C to G such that $\phi(u_i) = v_i$ and $\psi((u_i, u_{i+1})) = (v_i, v_{i+1})$ for all $0 \le i \le m-1$. Because both G and C are simple graphs, $\psi: E(C) \to E(G)$ is bijective. Hence, $(\phi,\psi): C \to G$ is a node-and-edge homomorphism, and the conclusion follows. □

In contrast to connected balanced simple graphs, if a balanced graph G contains double edges, then certainly no alternating cycle is homomorphic to G. What's more, it is possible that G is not homomorphic to any graph in the class $\mathcal{U}$; that is, G may not belong to $\mathcal{F}$. See Figure 9 for such an example. Nevertheless, the next lemma shows that if G has an edge multiplicity of at most 2, then it can be decomposed as a *union of elements* in $\mathcal{F}$.

Figure 7. Tree Representation of Graph in Figure 6



Lemma 3 (Decomposition). *Every balanced multigraph G with an edge multiplicity of at most 2 can be decomposed as a union of elements in $\mathcal{F}$.*

The proof is given in the e-companion to this paper.

For $k \ge 0$ and $\ell \ge 3$, we define $\mathcal{U}_{k,\ell} \subset \mathcal{U}$ as the bicolored balanced multigraphs $H \in \mathcal{U}$ with k *double edges* and ℓ *single edges*. The following lemma upper bounds the number of unlabeled graphs in $\mathcal{U}_{k,\ell}$.

Lemma 4 (Enumeration of Isomorphism Classes). *Let $k \ge 0$ and even $\ell \ge 4$. Then the number of unlabeled graphs in $\mathcal{U}_{k,\ell}$ is at most $17^k 4^\ell$.*

The proof is given in the e-companion to this paper. Define

$$\mathcal{F}^* = \{F \in \mathcal{F} : V(F) \subset [n] \text{ and } |E(F)| \le 4n \text{ and for every red edge } e \in E(F), x^*_e = 1\}.$$

The constraint that $|E(F)| \le 4n$ is because, for any vertex x of the F2F polytope, the multigraph G_y obtained from $y = 2(x - x^*)$ has a maximal degree of at most 8. Given $H \in \mathcal{U}$, we say a homomorphism $\phi: H \to F$ is *compatible with* x^* if $\phi(H) \in \mathcal{F}^*$. Denote by Φ^*_H the set of all homomorphisms $\phi: H \to F$ that are compatible with x^*. Then

$$\mathcal{F}^* = \{\phi(H) : \phi \in \Phi^*_H, H \in \mathcal{U}\}.$$

In the following, we upper bound the number of elements in Φ^*_H for a given $H \in \mathcal{U}$. We need to set up a few notations. Let H_d and H_s denote the subgraph of H induced by all the double edges and all the single edges, respectively. Then we have an edge-disjoint union $H = H_d \cup H_s$. For a vertex map $\phi \in \Phi^*_H$, let ϕ_d and ϕ_s denote ϕ restricted to $V(H_d)$ and $V(H_s)$, respectively. Note that $\phi_d(v) = \phi_s(v)$ for all $v \in V(H_d) \cap V(H_s)$. We write $\phi = (\phi_s, \phi_d)$.

Lemma 5 (Enumeration of Homomorphisms). *Let $k \ge 0$, and let $\ell \ge 4$ be an even integer. Fix a bicolored balanced multigraph $H \in \mathcal{U}_{k,\ell}$.*

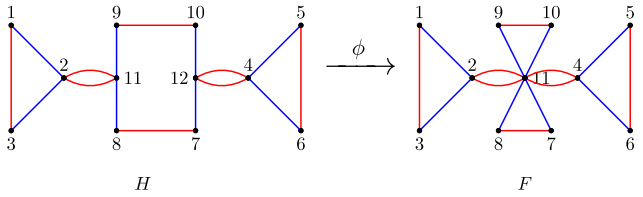
- *There exists an integer $0 \le r \le \ell/2$ such that*

$$\log|\Phi^*_{H_d}| \le \frac{1}{2}(k + r)\log(2n), \tag{23}$$

where

$$\Phi^*_{H_d} \triangleq \{\phi_d : \exists \phi_s, \text{ s.t. } (\phi_d, \phi_s) \in \Phi^*_H\}.$$

Figure 8. (Color online) Example of a Graph F in $\mathcal{F}$ but Not in $\mathcal{U}$



Note. Here, F is homomorphic to $H \in \mathcal{U}$, where the homomorphism ϕ maps both vertices 11 and 12 to 11.

- *For any fixed vertex map* $\phi_d : V(H_d) \to [n]$,

$$\log\left|\{\phi_s : (\phi_s, \phi_d) \in \Phi_H^*\}\right| \le (\ell/2 - r)\log n + (\ell/2 + k)\log 2. \tag{24}$$

The proof is given in the e-companion to this paper.

5.4. Proof of Theorem 1

We prove that if

$$\alpha - \log n \ge 16 \log 17, \tag{25}$$

then

$$\min_{x^*} \mathbb{P}\{\hat{x}_{\mathrm{F2F}} = x^*\} \ge 1 - 8\exp\bigl(-(\alpha - \log n)/8\bigr). \tag{26}$$

In fact, we will prove a stronger statement:

$$\min_{x^*} \mathbb{P}\bigl\{\langle w, x - x^*\rangle \le -(\alpha - \log n)/2,\ \forall \text{ extremal point } x \ne x^*\bigr\} \ge 1 - 8\exp\bigl(-(\alpha - \log n)/8\bigr). \tag{27}$$

Then Theorem 1 readily follows by taking $\alpha - \log n \to +\infty$.

For any extremal point x of F2F polytope, letting $y = 2(x - x^*)$, by Lemma 3,

$$G_y = \bigcup_{i=1}^{m} F_i, \quad F_i \in \mathcal{F}$$

for each $1 \le i \le m$ and some finite m. Note that for each red edge e in G_y, $x_e^* = 1$. Therefore, $F_i \in \mathcal{F}^*$. Thus, to prove (27), it suffices to show

$$\mathbb{P}\Bigl\{\max_{F \in \mathcal{F}^*} w(F) \le (\alpha - \log n)/2\Bigr\} \ge 1 - 8\exp\bigl(-(\alpha - \log n)/8\bigr). \tag{28}$$

Fix $k \ge 0$ and $\ell \ge 4$; define

$$\mathcal{F}_{k,\ell}^* = \bigl\{F \in \mathcal{F}^* : E(F) \text{ consists of } k \text{ double edges and } \ell \text{ single edges}\bigr\}.$$

Then

$$\mathcal{F}_{k,\ell}^* = \{\phi(H) : H \in \mathcal{U}_{k,\ell} \text{ and } \Phi \in \Phi_H^*\}, \tag{29}$$

and

$$\mathcal{F}^* = \bigcup_{k \ge 0} \bigcup_{\ell \ge 4} \mathcal{F}_{k,\ell}^*.$$

In view of (29), we have

$$\max_{F \in \mathcal{F}_{k,\ell}^*} w(F) = \max_{H \in \mathcal{U}_{k,\ell}} \max_{\phi \in \Phi_H^*} w(\phi(H)). \tag{30}$$

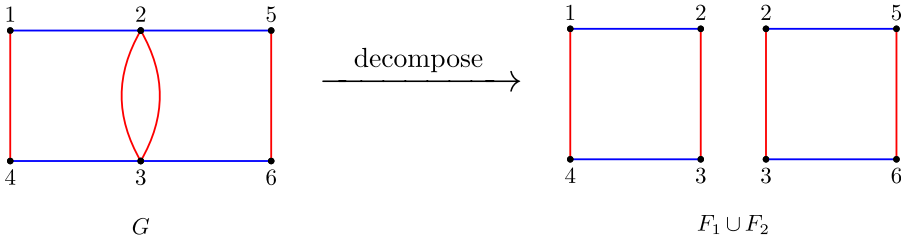
We first show a high probability bound to the inner maximum for a given $H \in \mathcal{U}_{k,\ell}$. Because maximizing over ϕ is equivalent to first maximizing over ϕ_d and then maximizing over ϕ_s for a fixed ϕ_d, it follows that

$$\max_{\phi \in \Phi_H^*} w(\phi(H)) = \max_{\phi_d \in \Phi_{H_d}^*} \left(w\bigl(\phi_d(H_d)\bigr) + \max_{\phi_s : (\phi_d, \phi_s) \in \Phi_H^*} w\bigl(\phi_s(H_s)\bigr) \right).$$

Recall that X_i's and Y_i's are two independent sequences of random variables, where X_i's are i.i.d. copies of $\log(\mathrm{d}P/\mathrm{d}Q)$ under distribution P and Y_i's are i.i.d. copies of $\log(\mathrm{d}P/\mathrm{d}Q)$ under distribution Q. Recall that ℓ_r and ℓ_b (respectively, k_r and k_b) denote the number of red and blue single (respectively, double) edges in H, respectively. Let $\delta = k_b - k_r = (\ell_r - \ell_b)/2$. Then $\delta \le \min\{k, \ell/2\}$. In view of (EC.23) and (EC.24) in the e-companion, for a fixed ϕ_d,

$$w\bigl(\phi_d(H_d)\bigr) \overset{d}{=} 2\left(\sum_{i=1}^{(k+\delta)/2} Y_i - \sum_{i=1}^{(k-\delta)/2} X_i \right),$$

Figure 9. (Color online) Example of a Graph G That Is Not in $\mathcal{F}$ but Can Be Decomposed as $F = F_1 \cup F_2$ with $F_1, F_2 \in \mathcal{F}$



and for a fixed ϕ_s,

$$w(\phi_s(H_s)) \stackrel{d}{=} \sum_{i=1}^{\ell/2-\delta} Y_i - \sum_{i=1}^{\ell/2+\delta} X_i,$$

where $\stackrel{d}{=}$ denotes equality in distribution. Moreover, for a fixed ϕ_d, $w(\phi_d(H_d))$ is the sum of the weights on double edges, which is independent of the collection of $w(\phi_s(H_s))$ ranging over all possible ϕ_s such that $(\phi_s, \phi_d) \in \Phi_H^*$.

Recall from Lemma 5 that there exists an integer $0 \le r \le \ell/2$ such that

$$\log|\Phi_{H_d}^*| \le \frac{1}{2}(k+r)\log(2n),$$
$$\log|\{\phi_s : (\phi_s, \phi_d) \in \Phi_H^*\}| \le (\ell/2 - r)\log n + (k+\ell/2)\log 2.$$

Invoking the large deviation bound Lemma EC.1 in Section EC.3 in the e-companion with $s = (k-\delta)/2$, $t = \ell/2 - \delta$, $u = \delta$, and $v = r - \delta$, and noting that

$$s + u + v/2 = (k+r)/2$$
$$t - v = \ell/2 - r,$$
$$s + t + u - v/2 = (k + \ell - r)/2 \ge k/2 + \ell/4,$$

and (25), we get

$$\mathbb{P}\left\{\max_{\phi \in \Phi_H^*} w(\phi(H)) \ge -(\alpha - \log n)(k/4 + \ell/8)\right\} \le 5\exp\left(-(\alpha - \log n)(k/8 + \ell/16)\right). \tag{31}$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left\{\max_{F \in \mathcal{F}_{k,\ell}^*} w(F) \ge -(\alpha - \log n)(k/4 + \ell/8)\right\} \\
&\stackrel{(a)}{\le} 5|\mathcal{U}_{k,\ell}|\exp\left(-(\alpha - \log n)(k/8 + \ell/16)\right) \\
&\stackrel{(b)}{\le} 5 \times 17^k 4^\ell \exp\left(-(\alpha - \log n)(k/8 + \ell/16)\right) \\
&\stackrel{(c)}{\le} 5\exp\left(-(\alpha - \log n)(k/16 + \ell/32)\right),
\end{aligned}$$

where (a) follows from union bound, (b) follows from Lemma 4, and (c) holds because $\alpha - \log n \ge 16\log 17$ by assumption (25). Taking another union bound over $k \ge 0$ and $\ell \ge 4$, we get

$$\begin{aligned}
&\mathbb{P}\left\{\max_{F \in \mathcal{F}^*} w(F) \ge -(\alpha - \log n)/2\right\} \\
&= \mathbb{P}\left\{\max_{k \ge 0, \ell \ge 4} \max_{F \in \mathcal{F}_{k,\ell}^*} w(F) \ge -(\alpha - \log n)/2\right\} \\
&\le \sum_{k \ge 0}\sum_{\ell \ge 4} \mathbb{P}\left\{\max_{F \in \mathcal{F}_{k,\ell}^*} w(F) \ge -(\alpha - \log n)(k/4 + \ell/8)\right\} \\
&\le 5\sum_{k \ge 0}\sum_{\ell \ge 4} \exp\left(-(\alpha - \log n)(k/16 + \ell/32)\right) \\
&\le \frac{5}{1 - e^{-(\alpha - \log n)/8}} \frac{e^{-(\alpha - \log n)/8}}{1 - e^{-(\alpha - \log n)/32}} \\
&\le \frac{5}{1 - 1/17} \frac{e^{-(\alpha - \log n)/8}}{1 - 1/4} \le 8e^{-(\alpha - \log n)/8}.
\end{aligned}$$

Therefore, we arrive at the desired (28), completing the proof of Theorem 1.

6. Information-Theoretic Necessary Conditions

We first present a general necessary condition needed for *any* algorithm to succeed in recovering the hidden Hamiltonian cycle with high probability. Recall that X and Y are two independent random variables distributed as the log likelihood ratio $\log(\mathrm{d}P/\mathrm{d}Q)$ under P and Q, respectively.

Theorem 2 (Information-Theoretic Conditions). *If there exists a sequence of estimators $\widehat{x}$ such that $\min_{x^* \in \mathcal{X}(G)} \mathbb{P}\{\widehat{x} = x^*\} \to 1$ as $n \to \infty$, then*

$$\sup_{\tau \in \mathbb{R}}\{\log \mathbb{P}\{X \le \tau\} + \log \mathbb{P}\{Y \ge \tau\}\} + \log n \le O(1). \tag{32}$$

Next, we state a regularity assumption on P and Q under which it immediately follows from Theorem 2 that $\alpha_n \ge (1 + o(1))\log n$ is necessary information theoretically, thereby establishing the optimality of F2F LP.

Assumption 1. *It holds that*

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

Corollary 1. *Suppose Assumption 1 holds. If there exists a sequence of estimators $\widehat{x}$ such that $\min_{x^* \in \mathcal{X}(G)} \mathbb{P}\{\widehat{x} = x^*\} \to 1$ as $n \to \infty$, then*

$$\alpha_n \ge (1 + o(1))\log n. \tag{33}$$

Assumption 1 is very general and fulfilled when the weight distributions are either Poisson, Gaussian, or Bernoulli, as the following result shows.

Lemma 6. *Assumption 1 holds in the Gaussian case with $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$, the Poisson case with $P = \mathrm{Pois}(\lambda)$ and $Q = \mathrm{Pois}(\mu)$ for $\lambda \ge \mu$ such that*

$$\log(\lambda\mu) = o\left((\sqrt{\lambda} - \sqrt{\mu})^2\right) + o(\log n),$$

and the Bernoulli case with $P = \mathrm{Bern}(p)$ and $Q = \mathrm{Bern}(q)$ for $p \ge q$.

The proof is given in the e-companion to this paper.

Let us now explain the intuition behind Assumption 1: Denote the log moment-generating function of X and Y as

$$\psi_P(\theta) = \log \mathbb{E}[e^{\theta X}], \quad \psi_Q(\theta) = \log \mathbb{E}[e^{\theta Y}] = \psi_P(\theta - 1). \tag{34}$$

Denote the Legendre transform of ψ_P and ψ_Q as

$$E_P(\tau) = \sup_{\theta \ge 0}\{-\theta\tau - \psi_P(-\theta)\}, E_Q(\tau) = \sup_{\theta \ge 0}\{\theta\tau - \psi_Q(\theta)\}. \tag{35}$$

Then Chernoff's inequality gives the following large deviation bounds: for any $\tau \in \mathbb{R}$,

$$\mathbb{P}\{X \le \tau\} \le \exp(-E_P(\tau)), \quad \mathbb{P}\{Y \ge \tau\} \le \exp\big(-E_Q(\tau)\big), \tag{36}$$

Therefore,

$$\sup_{\tau\in\mathbb{R}}\big\{\log \mathbb{P}\{X \le \tau\} + \log \mathbb{P}\{Y \ge \tau\}\big\} \le -\inf_{\tau\in\mathbb{R}}\big\{E_P(\tau) + E_Q(\tau)\big\}.$$

The infimum on the right-hand side is, in fact, equal to α_n. Indeed,

$$\begin{aligned}
&\inf_{\tau\in\mathbb{R}} E_P(\tau) + E_Q(\tau)\\
&\quad= \inf_{\tau\in\mathbb{R}} \sup_{\theta_1,\theta_2\ge 0} \{-\theta_1\tau - \psi_P(-\theta_1) + \theta_2\tau - \psi_Q(\theta_2)\}\\
&\quad\ge \sup_{\theta_1,\theta_2\ge 0} \left\{\inf_{\tau\in\mathbb{R}}(\theta_2-\theta_1)\tau - \psi_P(-\theta_1) - \psi_Q(\theta_2)\right\}\\
&\quad= \sup_{\theta\ge 0}\{-\psi_P(-\theta) - \psi_Q(\theta)\}\\
&\quad= -\psi_P(-1/2) - \psi_Q(1/2) = -2\log\int\sqrt{\mathrm{d}P\mathrm{d}Q} = \alpha_n,
\end{aligned}$$

and the infimum over τ is, in fact, achieved by

$$\tau^* = \psi_P'(-1/2) = \psi_Q'(1/2),$$

so that $E_P(\tau^*) + E_Q(\tau^*) = \alpha_n$. Hence,

$$\sup_{\tau\in\mathbb{R}}\big\{\log \mathbb{P}\{X \le \tau\} + \log \mathbb{P}\{Y \ge \tau\}\big\} \le -E_P(\tau^*) - E_Q(\tau^*) = -\alpha_n.$$

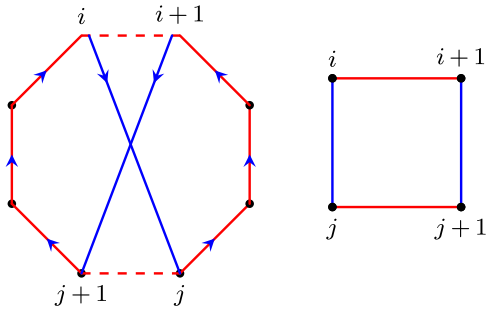
Therefore, the point of Assumption 1 is to require that the large deviation exponents in Chernoff's inequalities (36) are asymptotically tight, so that we can reverse the Chernoff bound in the lower bound proof.

6.1. Proof of Theorem 2

To lower bound the worst-case probability of error, consider the Bayesian setting where the hidden Hamiltonian cycle x^* is drawn uniformly at random from all possible Hamiltonian cycles of G. Because the prior distribution of x^* is uniform, the ML estimator minimizes the error probability among all estimators. Thus, without loss of generality, we can assume that the estimator $\widehat{x}$ used is $\widehat{x}_{\mathrm{ML}}$ and the true Hamiltonian cycle x^* is given by $(1, 2, \ldots, n, 1)$. Hence, by assumption, $\mathbb{P}\{\widehat{x}_{\mathrm{ML}} = x^*\} \to 1$.

Recall that the ML estimator is equivalent to finding a Hamiltonian cycle of the maximum weight. Given a Hamiltonian cycle x, define the simple graph G_x with a bicolored edge whose adjacency matrix is $|x - x^*|$, and each edge is colored in red if $(x - x^*)_e = -1$ and in blue if $(x - x^*)_e = +1$. Also, each edge e has a weight $w_e(x - x^*)_e$, and hence $w(G_x) = \langle w, x - x^*\rangle$. Note that if G_x is a 4-cycle of alternating colors given by $(i, i+1, j+1, j, i)$, then x corresponds to a Hamiltonian cycle constructed by deleting edges $(i, i+1), (j, j+1)$ in C^* and adding edges $(i, j), (i+1, j+1)$ (see Figure 10 for an illustration). Let $\mathcal{D}$ denote the set of all possible 4-cycles of alternating colors given by $(i, i+1, j+1, j)$. Then $|\mathcal{D}| = n(n-3)/2$, because for a given i, j have $(n-3)$ choices except $i-1, i, i+1$.

Figure 10. (Color online) The Cycle $(1, 2, \ldots, i, j, j-1, \ldots, i+1, j+1, j+2, \ldots, n)$ and the Corresponding Graph G_x as a 4-Cycle



Define

$$S = \sum_{D\in\mathcal{D}} \mathbf{1}_{\{w(D)\ge 0\}}.$$

If $S > 0$, then there exists a Hamiltonian cycle $x \ne x^*$ whose weight is at least as large as the weight of C^*; hence the likelihood function has at least two maximizers, which in turn implies the probability of exact recovery by ML estimator is at most 1/2. Therefore, $\frac{1}{2}\mathbb{P}\{S > 0\} \le \mathbb{P}\{\text{ML fails}\} = o(1)$. As a consequence, $\mathbb{P}\{S = 0\} \to 1$.

To explain the intuition, suppose $w(D)$ are mutually independent for all $D \in \mathcal{D}$. Then

$$\begin{aligned}
\mathbb{P}\{S = 0\} &= \mathbb{P}\{\forall D \in \mathcal{D}, w(D) < 0\}\\
&= \prod_{D\in\mathcal{D}} \mathbb{P}\{w(D) < 0\}\\
&\overset{(a)}{=} (1 - \mathbb{P}\{Y_1 + Y_2 - X_1 - X_2 \ge 0\})^{|\mathcal{D}|}\\
&\le \exp(-|\mathcal{D}|\mathbb{P}\{Y_1 + Y_2 - X_1 - X_2 \ge 0\}),
\end{aligned} \tag{37}$$

where (a) holds because $w(D)$ has the same distribution as $Y_1 + Y_2 - X_1 - X_2$, and the inequality in the last line holds in view of $1 - x \le e^{-x}$. In view of $\mathbb{P}\{S = 0\} \to 1$, it follows Equation (37) that

$$\log|\mathcal{D}| + \log \mathbb{P}\{Y_1 + Y_2 - X_1 - X_2 \ge 0\} \to -\infty. \tag{38}$$

Furthermore, for any $\tau \in \mathbb{R}$, we have

$$\begin{aligned}
&\log \mathbb{P}\{Y_1 + Y_2 - X_1 - X_2 \ge 0\}\\
&\quad\ge \log(\mathbb{P}\{Y_1 \ge \tau\}\mathbb{P}\{Y_2 \ge \tau\}\mathbb{P}\{X_1 \le \tau\}\mathbb{P}\{X_2 \le \tau\})\\
&\quad= 2\log \mathbb{P}\{Y \ge \tau\} + 2\log \mathbb{P}\{X \le \tau\}.
\end{aligned} \tag{39}$$

Combining Equations (38) and (39) and recalling that $|\mathcal{D}| = n(n-3)/2$, we immediately get that

$$\log \mathbb{P}\{Y \ge \tau\} + \log \mathbb{P}\{X \le \tau\} + \log n \to -\infty. \tag{40}$$

Taking the supremum over $\tau \in \mathbb{R}$ of the Equation (40) yields the desired (32).

However, $w(D)$ and $w(D')$ are dependent if D and D' share edges. To deal with this dependency, we focus on a subset of $\mathcal{D}$. In particular, for any $\tau \in \mathbb{R}$, define

$$I = \{\text{odd } i : w_{i,i+1} \le \tau\}$$

and

$$J = \{(i,j) \in I \times I : i \ne j, w_{i,j} + w_{i+1,j+1} \ge 2\tau\}.$$

Then for any $(i,j) \in J$, the alternating 4-cycle given by $(i, i+1, j+1, j, i)$ belongs to $\mathcal{D}$ and has a non-positive weight. Hence, $|J| \le S$, and thus $\mathbb{P}\{|J| = 0\} \ge \mathbb{P}\{S = 0\} \to 1$.

Note that $w_{i,i+1}$ has the same distribution as X. Thus for any $\tau \in \mathbb{R}$,

$$\mathbb{P}\{w_{i,i+1} \le \tau\} = \mathbb{P}\{X \le \tau\} \triangleq p.$$

Also, $w_{i,i+1}$ are mutually independent for different i. Thus $|I| \sim \text{Binom}(\lceil n/2 \rceil, p)$. By Chernoff's bound for binomial distribution,

$$\mathbb{P}\{|I| \le np/4\} \le \exp(-np/8). \tag{41}$$

Thus,

$$\begin{aligned}\mathbb{P}\{|J| = 0\} &\le \mathbb{P}\{|J| = 0, |I| > np/4\} + \mathbb{P}\{|I| \le np/4\} \\ &\le \mathbb{P}\{|J| = 0 \mid |I| > np/4\} + \exp(-np/8). \end{aligned} \tag{42}$$

Let $q \triangleq \mathbb{P}\{Y \ge \tau\}$. Then $\mathbb{P}\{Y_1 + Y_2 \ge 2\tau\} \ge \mathbb{P}\{Y_1 \ge \tau\} \cdot \mathbb{P}\{Y_2 \ge \tau\} = q^2$, and hence

$$\begin{aligned}&\mathbb{P}\{|J| = 0 \mid |I| > np/4\} \\ &\quad = \mathbb{P}\{\forall i < j \in I, w_{(i,j)} + w_{(i+1,j+1)} < 2\tau \mid |I| > np/4\} \\ &\quad \overset{(a)}{\le} (1 - q^2)^{\binom{np/4}{2}} \le e^{-q^2\binom{np/4}{2}},\end{aligned} \tag{43}$$

where (a) holds because conditional on I, $w_{(i,j)} + w_{(i+1,j+1)}$ are i.i.d. copies of $Y_1 + Y_2$. Combining Equations (40)–(42) yields

$$\mathbb{P}\{|J| = 0\} \le e^{-q^2\binom{np/4}{2}} + e^{-np/8}. \tag{44}$$

Recall that $\mathbb{P}\{|J| = 0\} \to 1$. It follows that

$$e^{-q^2\binom{np/4}{2}} + e^{-np/8} \ge 1 + o(1). \tag{45}$$

Hence, $\log n + \log p + \log q \le O(1)$, or equivalently,

$$\log \mathbb{P}\{X \le \tau\} + \log \mathbb{P}\{Y \ge \tau\} + \log n \le O(1). \tag{46}$$

Taking the supremum over $\tau \in \mathbb{R}$ of Equation (46) yields the desired (32).

Remark 1. In passing, we remark that in the Bernoulli case where $p = 1$ and $q = d/n$ for a fixed constant d, exact recovery of the hidden cycle is information-theoretically impossible. To see this, suppose the hidden Hamiltonian cycle is given by sequence of vertices $(1, 2, \ldots, n, 1)$. If $q = \Omega(1/n)$, then with a nonvanishing probability there exist $1 \le i \le n-4$ and $i+2 \le j \le n-2$ such that (i,j) and $(i+1, j+1)$ are edges in G. Thus we have a new Hamiltonian cycle by deleting edges $(i, i+1)$ and $(j, j+1)$ in the hidden one and adding edges (i,j) and $(i+1, j+1)$, leading to the impossibility of exact recovery. See Figure 10 for an illustration.

Acknowledgments

D. Tse and J. Xu thank the support of the Simons Institute, where this collaboration began. J. Xu also thanks Mohit Tawarmalani for inspiring discussions on fractional 2-factor LP relaxation of TSP. Y. Wu is grateful to Dan Spielman for helpful comments and to David Pollard and Dana Yang for an extremely thorough reading of the manuscript and various corrections.

References

Abbe E, Sandon C (2015) Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *Proc. 2015 IEEE 56th Annual Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 670–688.

Abbe E, Bandeira AS, Hall G (2016) Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory* 62(1):471–487.

Agarwal N, Bandeira AS, Koiliaris K, Kolla A (2017) Multisection in the stochastic block model using semidefinite programming. Boche H, Caire G, Calderbank R, März M, Kutyniok G, Mathar R, eds. *Compressed Sensing and Its Applications*, Applied and Numerical Harmonic Analysis (Birkhäuser, Cham, Switzerland), 125–162.

Allen-Zhu Z, Orecchia L (2019) Nearly linear-time packing and covering LP solvers. *Math. Programming* 175(1–2):307–353.

Alon N, Krivelevich M, Sudakov B (1998) Finding a large hidden clique in a random graph. *Random Structures Algorithms* 13(3–4): 457–466.

Atkins JE, Boman EG, Hendrickson B (1998) A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Comput.* 28(1):297–310.

Balinski ML (1965) Integer programming: Methods, uses, computations. *Management Sci.* 12(3):253–313.

Bandeira A (2018) Random Laplacian matrices and convex relaxations. *Foundations Comput. Math.* 18(2):345–379.

Barak B, Hopkins SB, Kelner JA, Kothari P, Moitra A, Potechin A (2016) A nearly tight sum-of-squares lower bound for the planted clique problem. *Proc. IEEE 57th Annual Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 428–437.

Bayati M, Borgs C, Chayes J, Zecchina R (2011) Belief propagation for weighted b-matchings on arbitrary graphs and its relation to linear programs with integer solutions. *SIAM J. Discrete Math.* 25(2):989–1011.

Bordenave C, Lelarge M, Massoulié L (2015) Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs. *Proc. 2015 IEEE 56th Annual Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 1347–1357.

Boyd S, Carr R (1999) A new bound for the ratio between the 2-matching problem and its linear programming relaxation. *Math. Programming* 86(3):499–514.

Broder AZ, Frieze AM, Shamir E (1994) Finding hidden Hamiltonian cycles. *Random Structures Algorithms* 5(3):395–410.

Cai T, Liang T, Rakhlin A (2017) On detection and structural reconstruction of small-world random networks. *IEEE Trans. Network Sci. Engrg.* 4(3):165–176.

Chen K, Chen K, Muller HG, Wang JL (2011) Stringing high-dimensional data for functional analysis. *J. Amer. Statist. Assoc.* 106(493):275–284.

Chlamtác E, Tulsiani M (2012) Convex relaxations and integrality gaps. Anjos M, Lasserre J, eds. *Handbook on Semidefinite, Conic and Polynomial Optimization* (Springer, Boston), 139–169.

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

Cvetković D, Čangalović M, Kovačević-Vujčić V (1999) Semidefinite programming methods for the symmetric traveling salesman problem. Cornuéjols G, Burkard RE, Woeginger GJ, eds. *Proc. Internat. Conf. Integer Programming Combin. Optim.* (Springer, Berlin), 126–136.

Dantzig G, Fulkerson R, Johnson S (1954) Solution of a large-scale traveling-salesman problem. *J. Oper. Res. Soc. Amer.* 2(4): 393–410.

De Klerk E, Pasechnik DV, Sotirov R (2008) On semidefinite programming relaxations of the traveling salesman problem. *SIAM J. Optim.* 19(4):1559–1573.

Deshpande Y, Montanari A (2015) Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. Grünwald P, Hazan E, Kale S, eds. *Proc. 28th Conf. Learn. Theory*, vol. 40 (PMLR, Paris), 523–562.

Edmonds J (1965a) Maximum matching and a polyhedron with 0, 1-vertices. *J. Res. Natl. Bureau Standards B* 69B(55–56):125–130.

Edmonds J (1965b) Paths, trees, and flowers. *Canadian J. Math.* 17(3): 449–467.

Fogel F, Jenatton R, Bach F, d'Aspremont A (2013) Convex relaxations for permutation problems. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates, Red Hook, NY), 1016–1024.

Hajek B, Wu Y, Xu J (2016a) Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inform. Theory* 62(5):2788–2797.

Hajek B, Wu Y, Xu J (2016b) Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inform. Theory* 62(10):5918–5937.

Hajek B, Wu Y, Xu J (2016c) Semidefinite programs for exact recovery of a hidden community. Feldman V, Rakhlin A, Shamir O, eds. *Proc. 29th Conf. Learn. Theory*, vol. 49 (PMLR, New York), 1051–1095.

Held M, Karp RM (1970) The traveling-salesman problem and minimum spanning trees. *Oper. Res.* 18(6):1138–1162.

Jerrum M (1992) Large cliques elude the Metropolis process. *Random Structures Algorithms* 3(4):347–359.

Jog V, Loh P-L (2015) Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence. Preprint arXiv 1509.06418, submitted September 21, https://arxiv.org/abs/1509.06418.

Kendall DG (1971) Abundance matrices and seriation in archaeology. *Probab. Theory Related Fields* 17(2):104–112.

Kotzig A (1968) Moves without forbidden transitions in a graph. *Matematický časopis* 18(1):76–80.

Letchford AN, Reinelt G, Theis DO (2008) Odd minimum cut sets and *b*-matchings revisited. *SIAM J. Discrete Math.* 22(4):1480–1487.

Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289–293.

Lim CH, Wright S (2014) Beyond the Birkhoff polytope: Convex relaxations for vector permutation problems. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Red Hook, NY), 2168–2176.

Lovász L (2012) *Large Networks and Graph Limits*, Colloquium Publications, vol. 60 (American Mathematical Society, Providence, RI).

Mallows CL (1957) Non-null ranking models. I. *Biometrika* 44(1/2): 114–130.

Massoulié L (2013) Community detection thresholds and the weak Ramanujan property. *Proc. 46th Annual ACM Sympos. Theory Comput.* (ACM, New York), 694–703.

McSherry F (2001) Spectral partitioning of random graphs. *Proc. 42nd IEEE Sympos. Foundations Comput. Sci.* (IEEE Computer Society, Washington, DC), 529–537.

Meka R, Potechin A, Wigderson A (2015) Sum-of-squares lower bounds for planted clique. *Proc. 47th Annual ACM Sympos. Theory Comput.* (ACM, New York), 87–96.

Mossel E, Neeman J, Sly A (2015) Consistency thresholds for the planted bisection model. *Proc. 47th Annual ACM Sympos. Theory Comput.* (ACM, New York), 69–75.

Motahari AS, Bresler G, Tse D (2013) Information theory of DNA shotgun sequencing. *IEEE Trans. Inform. Theory* 59(10): 6273–6289.

Perry W, Wein A (2017) A semidefinite program for unbalanced multisection in the stochastic block model. *Proc. 2017 International Conference on Sampling Theory and Applications (SampTA)* (IEEE, Piscataway, NJ), 64–67.

Pevzner PA (1995) DNA physical mapping and alternating Eulerian cycles in colored graphs. *Algorithmica* 13(1–2):77–105.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 26(3):342–350.

Rényi A (1961) On measures of entropy and information. *Proc. 4th Berkeley Sympos. Math. Statist. Probab.*, vol. 1: *Contributions Theory Statist.* (University of California Press, Berkeley), 547–561.

Robinson WS (1951) A method for chronologically ordering archaeological deposits. *Amer. Antiquity* 16(4):293–301.

Schalekamp F, Williamson DP, van Zuylen A (2013) 2-matchings, the traveling salesman problem, and the subtour LP: A proof of the Boyd-Carr conjecture. *Math. Oper. Res.* 39(2):403–417.

Schrijver A (2003) *Combinatorial Optimization: Polyhedra and Efficiency*, Algorithms and Combinatorics, vol. 24 (Springer Science & Business Media, New York).

Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. *Nature* 393(6684):440–442.

Zhang AY, Zhou HH (2016) Minimax rates of community detection in stochastic block models. *Ann. Statist.* 44(5):2252–2280.

Zhao Q, Karisch SE, Rendl F, Wolkowicz H (1998) Semidefinite programming relaxations for the quadratic assignment problem. *J. Combin. Optim.* 2(1):71–109.

Vivek Bagaria is a doctoral student in the Electrical Engineering Department at Stanford. His research interests include data science, algorithms, machine learning, and blockchains.

Jian Ding is an associate professor in the Wharton School of Business at University of Pennsylvania. His research interest is on probability theory, with a focus on interactions with statistical physics and theoretical computer science. He received a National Science Foundation CAREER award and an Alfred Sloan Foundation fellowship in 2015 and the Rollo Davidson Prize in 2017.

David Tse is the Thomas Kailath and Guanghan Xu Professor in the School of Engineering at Stanford University. He is a member of the U.S. National Academy of Engineering. He received the 2017 Claude E. Shannon Award from the Information Theory Society and the 2019 IEEE Richard W. Hamming Medal. His research interests are in information theory, computational genomics, machine learning, and blockchains.

Yihong Wu is an assistant professor in the Department of Statistics and Data Science at Yale University. He received a Sloan fellowship in mathematics in 2018 and a National Science Foundation CAREER award in 2017. He is broadly interested in the theoretical and algorithmic aspects of high-dimensional statistics, information theory, and optimization.

Jiaming Xu is an assistant professor in the Fuqua School of Business at Duke University. His research interests include data science, high-dimensional statistical inference, information theory, convex and nonconvex optimization, queueing theory, and game theory. He received a Simons-Berkeley Fellowship in 2016.