Detection-Recovery Gap for Planted Dense Cycles

Cheng Mao CHENG.MAO@MATH.GATECH.EDU
School of Mathematics, Georgia Institute of Technology

Alexander S. Wein ASWEIN@UCDAVIS.EDU
Department of Mathematics, University of California, Davis

Shenduo Zhang SZHANG705@GATECH.EDU
School of Mathematics, Georgia Institute of Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Planted dense cycles are a type of latent structure that appears in many applications, such as small-world networks in social sciences and sequence assembly in computational biology. We consider a model where a dense cycle with expected bandwidth $n\tau$ and edge density p is planted in an Erdős–Rényi graph $G(n, q)$. We characterize the computational thresholds for the associated detection and recovery problems for the class of low-degree polynomial algorithms. In particular, a gap exists between the two thresholds in a certain regime of parameters. For example, if $n^{-3/4} \ll \tau \ll n^{-1/2}$ and $p = Cq = \Theta(1)$ for a constant $C > 1$, the detection problem is computationally easy while the recovery problem is hard for low-degree algorithms.

Keywords: Planted dense cycle, low-degree polynomial, computational lower bound, detection-recovery gap

1. Introduction

Recovering latent structures in networks is a broad class of problems that are essential both in theory and for applications in the social and biological sciences Watts (2004); Barabási (2012). In this work, we study the detection and recovery of a hidden cyclic structure in an observed network, a type of structure found in many real-world applications. For example, the celebrated *Watts–Strogatz small-world model* Watts and Strogatz (1998) assumes that n nodes have latent positions on a circle, and they have stronger connections with their k-nearest neighbors and weaker connections with all other nodes. Observing such a small-world network, the problem of interest is to recover the relative positions of the nodes—which nodes are k-nearest neighbors of each other—and hence the overall structure of the network. Since its proposal, the Watts–Strogatz model has been used extensively to study, for example, epidemic behavior Moore and Newman (2000), collaboration networks Uzzi and Spiro (2005), and brain networks Bassett and Bullmore (2006). More generally, the problem of recovering a one-dimensional embedding of n objects from pairwise similarities between them arises in a wider range of applications, including relative dating in archaeology Robinson (1951), *de novo* genome assembly in computational biology Lieberman-Aiden et al. (2009), and angular synchronization in tomography Singer (2011).

Despite the vast literature on related models and algorithms, the statistical and computational limits of this problem are not yet well-established in a rigorous framework. The information-theoretic thresholds for the Watts–Strogatz model are studied in Cai et al. (2017), but the upper bounds achieved by computationally efficient algorithms are far from the information-theoretic

thresholds. It is unknown whether these statistical-to-computational gaps are inherent, or whether they can be closed by other efficient algorithms. In the case where the bandwidth k is at most $n^{o(1)}$, sharp characterizations of recovery conditions are given in Bagaria et al. (2020); Ding et al. (2020a) under a more general model. Moreover, several other algorithms and analyses have been introduced for related models from the perspective of graphon estimation by Janssen and Smith (2022); Natik and Smith (2021); Giraud et al. (2021). However, none of the previous works have shown computational lower bounds against a class of efficient algorithms.

Moreover, the Watts–Strogatz small-world model can be seen as modeling a one-dimensional, noisy *random geometric graph* with latent locations on a circle. Random geometric graphs have long been studied in a variety of scientific fields; see, e.g., Penrose (2003) and a recent survey by Duchemin and De Castro (2022). In particular, detection or testing thresholds for high-dimensional, noiseless random geometric graphs were studied by Bubeck et al. (2016) and improved by Brennan et al. (2020); Liu et al. (2022). Variants of the model with edge noise have also been studied recently by Liu and Rácz (2021); Liu and Racz (2021). Recovery or reconstruction of the latent geometry from a random geometric graph has also been long studied in various models, especially using spectral techniques; see, e.g., Sussman et al. (2013); Araya Valdivia and Yohann (2019); Eldan et al. (2022). Despite the vast literature, the discrepancy between detection and recovery thresholds is yet to be understood in a single model.

In this work, we propose a variant of the Watts–Strogatz small-world model, which is a random graph with a *planted dense cycle*, and study the computational complexities of the associated detection and recovery problems in the framework of *low-degree polynomial algorithms*. This framework has proven to be successful at probing the computational complexity of detecting and estimating hidden structures in high-dimensional settings Hopkins and Steurer (2017); Kunisky et al. (2019); Schramm and Wein (2022) and is closely related to the sum-of-squares hierarchy Hopkins et al. (2017); Hopkins (2018). For problems such as planted clique, community detection, and sparse PCA, the conjectured hard regime where no polynomial-time algorithms are known to exist coincides with the regime where low-degree polynomials fail to solve the problem. For the planted dense cycle problem, we identify the regimes where low-degree polynomial algorithms fail to detect and recover the hidden cycle respectively. In particular, we show that the threshold for detection is drastically different from that for recovery, so there is a *detection-recovery gap* for this problem.

Notation Let $[n] := \{1, 2, \dots, n\}$ and $\binom{[n]}{2} := \{(i, j) : i, j \in [n],\ i < j\}$. We use the standard asymptotic notation $O(\cdot), o(\cdot), \Omega(\cdot), \omega(\cdot), \Theta(\cdot)$ as $n \to \infty$, and a tilde is added if the asymptotic relation holds up to a polylogarithmic factor in n.

Any subset $\alpha \subseteq \binom{[n]}{2}$ can be identified with the graph on vertex set $[n]$ induced by edges in α. Therefore, we can say "graph α" without ambiguity. Then $|\alpha|$ denotes the number of edges in the graph α. Let $V(\alpha) \subseteq [n]$ denote the vertex set of α, i.e., the set of vertices $v \in [n]$ that are non-isolated by the edges of α.

2. Models and main results

2.1. Planted dense cycles

We now formally introduce our models. For any $a, b \in [0, 1]$, define

$$\mathfrak{d}(a, b) := \min\{|a - b|, 1 - |a - b|\}.$$

In other words, $\mathfrak{d}(a, b)$ is the distance between a and b on a circle of circumference 1. Throughout the paper, we consider the setting where the number of vertices n grows, and other parameters p, q, and τ may depend on n.

Definition 1 (Model $\mathcal{P}$, Planted Dense Cycle) *Suppose that $0 \le q < p \le 1$ and $0 \le \tau \le 1/2$. Let $z \in [0,1]^n$ be a latent random vector whose entries $z_1, \dots, z_n$ are i.i.d. $\mathrm{Unif}([0,1])$ variables. We observe an undirected graph with adjacency matrix $A \in \mathbb{R}^{n \times n}$ whose edges, conditional on $z_1, \dots, z_n$, are independently sampled as follows: $A_{ij} \sim \mathrm{Bern}(p)$ if $\mathfrak{d}(z_i, z_j) \le \tau/2$ and $A_{ij} \sim \mathrm{Bern}(q)$ otherwise, where $(i,j) \in \binom{[n]}{2}$. We write $A \sim \mathcal{P}$.*

In short, a graph A from model $\mathcal{P}$ is a $G(n, q)$ Erdős–Rényi graph with a planted dense cycle that has edge density p and expected bandwidth $n\tau$. The location of the cycle is determined by the latent variable z. For comparison, the Watts–Strogatz model plants a dense cycle of bandwidth *exactly* $n\tau$; it also assumes that the average degree is matched to that in the noiseless case where $p = 1$ and $q = 0$, so $\tau = \tau p + (1 - \tau)q$ in Watts and Strogatz (1998); Cai et al. (2017). Moreover, the bandwidth $n\tau$ is typically much smaller than n in small-world networks, so we may assume $\tau \le 1/2$ throughout the paper to ease the presentation.

In addition, we use $\mathcal{Q}$ to denote an Erdős–Rényi graph model.

Definition 2 (Model $\mathcal{Q}$, Erdős–Rényi graph) *Suppose that $0 \le q < p \le 1$ and $0 \le \tau \le 1/2$. Let $r := \tau p + (1 - \tau)q$. We observe a $G(n, r)$ Erdős–Rényi graph with adjacency matrix $A \in \mathbb{R}^{n \times n}$. We write $A \sim \mathcal{Q}$.*

Note that the condition $r = \tau p + (1 - \tau)q$ is imposed so that the average degrees are matched in the two models $\mathcal{P}$ and $\mathcal{Q}$.

There are two problems associated with the model of planted dense cycle, detection and recovery. Detection of the planted cycle is formulated as a statistical hypothesis testing problem.

Problem 3 (Detection) *Observing the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of a graph, we test $A \sim \mathcal{P}$ against $A \sim \mathcal{Q}$.*

Recovery of the planted cycle is formulated as determining whether vertices i and j are neighbors in the cycle for $(i,j) \in \binom{[n]}{2}$, i.e., whether $\mathfrak{d}(z_i, z_j) \le \tau/2$. By symmetry, it suffices to consider the pair of vertices $(1, 2)$ and estimate $\mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$.

Problem 4 (Recovery) *Observing the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of a graph $A \sim \mathcal{P}$ with a planted cycle, we aim to recover $\chi := \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$.*

2.2. Overview of results

Our results fall within the framework of low-degree polynomial algorithms (see Kunisky et al. (2019)). Let $\mathbb{R}[A]_{\le D}$ denote the set of multivariate polynomials in the entries of A with degree at most D. The scaling of $D = D_n$ will be made precise later, but in general, when we speak of a "low-degree" polynomial, its degree is at most $D = n^{o(1)}$.

For the detection problem, we study the ability of such a polynomial to distinguish the two distributions $\mathcal{P}$ and $\mathcal{Q}$, in the following sense (Bandeira et al., 2022, Definition 1.6).

Definition 5 (Strong separation) *A polynomial $f = f_n \in \mathbb{R}[A]_{\le D}$ is said to* strongly separate *$\mathcal{P}$ and $\mathcal{Q}$ over A if*

$$\sqrt{\mathrm{Var}_{\mathcal{P}}[f(A)] \vee \mathrm{Var}_{\mathcal{Q}}[f(A)]} = o\left(|\mathbb{E}_{\mathcal{P}}[f(A)] - \mathbb{E}_{\mathcal{Q}}[f(A)]|\right)$$

as $n \to \infty$.

By Chebyshev's inequality, strong separation implies that, by thresholding the value $f(A)$, one can test between $A \sim \mathcal{P}$ and $A \sim \mathcal{Q}$ with both type I and type II errors of order $o(1)$.

For the recovery problem, recall that we aim to estimate $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$. The quantity of interest is the degree-D minimum mean squared error (see Schramm and Wein (2022))

$$\mathsf{MMSE}_{\le D} := \inf_{f \in \mathbb{R}[A]_{\le D}} \mathbb{E}_{\mathcal{P}}\left[(f(A) - \chi)^2\right].$$

It is equivalent to consider the degree-D maximum correlation

$$\mathsf{Corr}_{\le D} := \sup_{\substack{f \in \mathbb{R}[A]_{\le D}, \\ \mathbb{E}_{\mathcal{P}}[f(A)^2] \neq 0}} \frac{\mathbb{E}_{\mathcal{P}}[f(A) \cdot \chi]}{\sqrt{\mathbb{E}_{\mathcal{P}}[f(A)^2]}} \tag{1}$$

because of the following relation (Schramm and Wein, 2022, Fact 1.1)

$$\mathsf{MMSE}_{\le D} = \mathbb{E}_{\mathcal{P}}[\chi^2] - \mathsf{Corr}^2_{\le D}.$$

The trivial estimator $f(A) \equiv \mathbb{E}_{\mathcal{P}}[\chi]$ of χ achieves a correlation

$$\frac{\mathbb{E}_{\mathcal{P}}[f(A) \cdot \chi]}{\sqrt{\mathbb{E}_{\mathcal{P}}[f(A)^2]}} = \mathbb{E}_{\mathcal{P}}[\chi],$$

which motivates the following definition.

Definition 6 (Weak recovery) *A polynomial $f = f_n \in \mathbb{R}[A]_{\le D}$ is said to* weakly recover *an estimand χ given $A \sim \mathcal{P}$ if*

$$\frac{\mathbb{E}_{\mathcal{P}}[f(A) \cdot \chi]}{\sqrt{\mathbb{E}_{\mathcal{P}}[f(A)^2]}} = \omega\left(\mathbb{E}_{\mathcal{P}}[\chi]\right)$$

as $n \to \infty$.

Note that for the estimand $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$ in Problem 4, we have $\mathbb{E}_{\mathcal{P}}[\chi] = \tau$.

For both the detection and the recovery problem, we establish low-degree upper and lower bounds that match up to an n^{δ} factor for an arbitrarily small constant $\delta > 0$. Our main results are summarized in the following theorem.

Theorem 7 (Summary of the detection-recovery gap) *Suppose that $Cq \le p \le C'q$ for constants $C' > C > 1$. Fix any constant $\delta \in (0, 0.1)$. Suppose that $2/\delta \le D \le o\Big(\big(\frac{\log n}{\log\log n}\big)^2\Big)$ and $\tau \le (\log n)^{-8}$.*

- ***(Detection)*** *Consider Problem 3 and Definition 5. If $n^3p^3\tau^4 \le n^{-\delta}$, then no polynomial in $\mathbb{R}[A]_{\le D}$ strongly separates $\mathcal{P}$ and $\mathcal{Q}$. If $n^3p^3\tau^4 = \omega(1)$, then there is a polynomial in $\mathbb{R}[A]_{\le D}$ that strongly separates $\mathcal{P}$ and $\mathcal{Q}$.*
- ***(Recovery)*** *Consider Problem 4 and Definition 6. If $np\tau^2 \le n^{-\delta}$, then no polynomial in $\mathbb{R}[A]_{\le D}$ weakly recovers χ. If $np\tau^2 \ge n^{\delta}$, then there is a polynomial in $\mathbb{R}[A]_{\le D}$ that weakly recovers χ.*

Proof The four bounds are established in Theorems 8, 9, 12, and 13, respectively. It suffices to note that under the assumptions of the theorem, the conditions (4), (7), (9), and (17) are all satisfied. See also the discussion after each of the theorems. ■

By the above theorem, there is a gap between the detection threshold and the recovery threshold for planted dense cycles if we focus on low-degree polynomials. To better illustrate the detection-recovery gap, let us suppose $p = n^{-a}$ and $\tau = n^{-b}$ for constants $a, b \in (0, 1)$. Then the detection threshold is given by $3 - 3a - 4b = 0$, while the recovery threshold is given by $1 - a - 2b = 0$. We plot the phase diagram in Figure 1. In particular, in region B of the figure, detection is easy while recovery is hard.

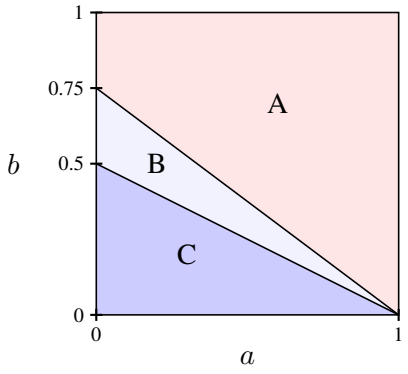

Figure 1: The detection-recovery gap for planted dense cycles with $p = n^{-a}$ and $\tau = n^{-b}$. Detection is hard in region A, and easy in regions B and C. Recovery is hard in regions A and B, and easy in region C.

In Theorem 7, we have assumed that p and q are of the same order, which is a standard simplification in the literature for related problems (see, e.g., Hajek et al. (2015)). In fact, for three of the four bounds (Theorems 8, 9, and 12, except the recovery upper bound), the edge density p in the cycle can be much higher than the edge density q outside the cycle; for the detection and recovery lower bounds (Theorems 8 and 12), p can also approach q in the sense that $p-q$ is of a smaller order than p or q. The latter regime is also addressed in a related but different context of computational lower bounds by Brennan et al. (2019).

For the recovery upper bound, our proposed statistic and analysis yield stronger results than weak recovery if we consider efficient algorithms beyond low-degree polynomials. Namely, we produce an estimator $\hat{\chi} \in \{0, 1\}$ that recovers $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$ with high probability, and also a consistent estimator of the underlying random geometric graph. See Theorem 14 and Corollary 15.

Technical contributions It is worth noting that none of the four bounds follow trivially from existing work. For the detection lower bound, while the framework of low-degree polynomials is well-understood (see, e.g., Hopkins (2018); Bandeira et al. (2022)), we provide a new application to random geometric graphs. The analysis also prompts us to study the signed triangle count proposed by Bubeck et al. (2016) in the noisy case, proving the detection upper bound. For the recovery lower bound, we generalize the technique developed by Schramm and Wein (2022) for the planted clique problem to a general binary observation model, and then apply it to our problem. Finally, the most technical part of this work is the recovery upper bound, where we provide a delicate analysis of self-avoiding walks between two vertices in the observed graph. The same statistic has been used by Hopkins and Steurer (2017) for community detection, but we perform a new analysis of certain probabilistic and combinatorial properties of self-avoiding walks in random geometric graphs.

Open problems While we have characterized the detection and recovery thresholds for low-degree polynomial algorithms, the information-theoretic thresholds for both problems remain largely open. Most existing results in the literature of small-world graphs or random geometric graphs focus on different regimes and are not comparable to our results. For example, Ding et al. (2020a) consider the regime $\tau n = n^{o(1)}$ and Liu and Rácz (2021) assume a constant p. One possible exception is the work by Cai et al. (2017), which assumes a bandwidth exactly τn instead of τn in expectation. Ignoring this difference, their results can be compared to ours for p, q, and τ all of the order n^{-a}, i.e., on the diagonal $a = b$ in Figure 1. One of their results states that detection is information-theoretically possible if $a < 1/2$. Consequently, the information-theoretic threshold for detection would be inside region A in Figure 1, and there would be a statistical-to-computational gap for the detection problem. However, since the comparison between Cai et al. (2017) and our work is not fully rigorous, we leave the study of information-theoretic thresholds to future work.

Another interesting question left open by our work is what the detection and recovery thresholds are for higher-dimensional geometry. For example, the latent locations $z_1, \dots, z_n$ may be distributed on the unit sphere $\mathcal{S}^{d-1}$ in $\mathbb{R}^d$ for $d \geq 3$, rather than on a circle. We believe many of the results in this work extend to the case of a fixed d, but if d grows with n, then the problem becomes significantly more difficult and novel ideas are required.

In the sequel, we present low-degree lower bounds before upper bounds for both the detection and the recovery problem. The rationale behind this nonstandard order of presentation is in fact an important advantage of the low-degree framework: The proof of a low-degree lower bound will naturally suggest an efficient algorithm that potentially achieves the matching upper bound.

3. The detection problem

As discussed in the introduction, the planted dense cycle model is a one-dimensional random geometric graph model. Detection of geometry in random graphs has been studied, and a canonical algorithm for this task is counting *signed triangles* proposed by Bubeck et al. (2016). On the other hand, computational lower bounds for random geometric graphs are not well-understood even in the one-dimensional case. We first present the low-degree lower bounds, whose proof suggests that the statistic of signed triangles has the best distinguishing power. Then we give a self-contained analysis of signed triangles in our case.

3.1. Lower bound

The standard procedure for proving low-degree lower bounds consists in analyzing the distinguishing power of an orthonormal basis of functions of the observations under model $\mathcal{Q}$. Towards this end, for $(i,j) \in \binom{[n]}{2}$, define

$$\bar{A}_{ij} := \frac{A_{ij} - r}{\sqrt{r(1-r)}}. \tag{2}$$

For $\alpha \subseteq \binom{[n]}{2}$, define

$$\phi_\alpha(A) := \prod_{(i,j)\in\alpha} \bar{A}_{ij},$$

and let $\phi_\varnothing(A) \equiv 1$. Then $\{\phi_\alpha\}_{\alpha \subseteq \binom{[n]}{2}}$ is an orthonormal basis for functions on the hypercube $\{0,1\}^{\binom{[n]}{2}}$ under $\mathcal{Q}$. Moreover, since r is the average edge density in both models $\mathcal{P}$ and $\mathcal{Q}$, the larger p is compared to r, the larger signal we have at each edge. Hence we define a quantity

$$\mu := \frac{p-r}{\sqrt{r(1-r)}} \tag{3}$$

that can be understood as the signal-to-noise ratio of model $\mathcal{P}$. We have the following theorem.

Theorem 8 (Detection lower bound) *Consider Problem 3. Fix any constant $\delta \in (0, 0.1)$. No polynomial $f \in \mathbb{R}[A]_{\le D}$ strongly separates $\mathcal{P}$ and $\mathcal{Q}$ in the sense of Definition 5, if*

$$n^3\tau^4\mu^6 \le n^{-\delta}, \qquad \mu = \tilde{O}(1), \qquad D = o\left(\left(\frac{\log n}{\log\log n}\right)^2\right). \tag{4}$$

To clarify, $\tilde{O}(1)$ in (4) does not stand for a specific bound but rather allows $\mu = \mu_n$ to be *any* sequence that scales as $\tilde{O}(1)$, and similarly for the condition on D. In addition, to ease the presentation, we have assumed the conditions in (4) that are stronger than what is required by the proof: It suffices to assume $n^3\tau^4\mu^6 \le n^{-o(1)}$ for an appropriately defined $o(1)$ quantity, and the degree D can be polylogarithmic in n or even $n^{o(1)}$ if μ is sufficiently small.

The proof of the above theorem is deferred to Section A.1. We now provide a proof sketch. To show that no polynomial of degree at most D strongly separates $\mathcal{P}$ and $\mathcal{Q}$, it suffices to prove that the "advantage"

$$\mathsf{Adv}_{\le D} := \sup_{\substack{f\in\mathbb{R}[A]_{\le D},\\ \mathbb{E}_\mathcal{Q}[f(A)^2]\neq 0}} \frac{\mathbb{E}_\mathcal{P}[f(A)]}{\sqrt{\mathbb{E}_\mathcal{Q}[f(A)^2]}}$$

is $O(1)$; see (Bandeira et al., 2022, Proposition 6.2). Furthermore, it is known (Hopkins, 2018, Section 2.3) that

$$\mathsf{Adv}^2_{\le D} = \sum_{\alpha\subseteq\binom{[n]}{2}\,:\,|\alpha|\le D} (\mathbb{E}_\mathcal{P}[\phi_\alpha(A)])^2. \tag{5}$$

The rest of the proof consists in controlling all the summands in (5), which is done in Section A.1. This eventually leads to Proposition 20, from which Theorem 8 easily follows.

To further clarify the intuition behind the sum in (5), for each subgraph $\alpha \subseteq \binom{[n]}{2}$, we can understand the quantity $\mathbb{E}_\mathcal{P}[\phi_\alpha(A)]$ as the "power" of the statistic $\phi_\alpha(A)$ in distinguishing $\mathcal{P}$ from

$\mathcal{Q}$. The lower bound requires that the total distinguishing power, as a sum of $(\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)])^2$ over all low-degree α, is bounded. On the other hand, if $\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)]$ is large for a particular choice of α, then the corresponding statistic $\phi_\alpha(A)$ can be used for testing between $\mathcal{P}$ and $\mathcal{Q}$. A careful study of $\mathbb{E}[\phi_\alpha(A)]$ in Proposition 20 suggests that the bottleneck case is when the graph α is a triangle. Therefore, it is natural to consider signed triangles for the upper bound.

3.2. Upper bound

While signed triangles have been analyzed for random geometric graphs in previous works such as Bubeck et al. (2016); Brennan et al. (2020); Liu and Rácz (2021); Liu et al. (2022), none of these results apply in our case. For example, the setup closest to ours can be found in Liu and Rácz (2021), where high-dimensional random geometric graphs are studied but the probability p has to be fixed. Therefore, we present a self-contained analysis of the signed triangle statistic

$$S_3(A) := \sum_{H \in \binom{[n]}{3}} \prod_{(i,j) \in \binom{H}{2}} \bar{A}_{ij}. \tag{6}$$

Note that if $\bar{A}_{ij}$ were replaced by A_{ij} in the above definition, then $S_3(A)$ would be the number of triangles in the graph A. Hence $S_3(A)$ is a standardized version of triangle count.

Theorem 9 (Detection upper bound) *Consider Problem 3. Suppose that $p \geq Cq$ for a constant $C > 1$. The degree-3 polynomial $S_3(A)$ defined in (6) strongly separates $\mathcal{P}$ and $\mathcal{Q}$ in the sense of Definition 5, if*

$$n^3\tau^4p^6/r^3 = \omega(1), \qquad n^3\tau^2p^3 = \omega(1). \tag{7}$$

The proof of the theorem is deferred to Section A.2. In short, we control the two expectations $\mathbb{E}_{\mathcal{P}}[f(A)]$, $\mathbb{E}_{\mathcal{Q}}[f(A)]$ and the two variances $\mathrm{Var}_{\mathcal{P}}[f(A)]$, $\mathrm{Var}_{\mathcal{Q}}[f(A)]$ in Propositions 21, 23, and 24, which together result in Theorem 9. We have again chosen simplicity over generality for the statement of the above theorem by assuming $p \geq Cq$ for $C > 1$. A more general condition can be obtained from a refined comparison between the bounds in Propositions 21 and 24. The two conditions in (7) can be interpreted as follows. First, as we see in the proof, $\mu = \Theta(p/r^{1/2})$, so the first condition in (7) matches the first condition in (4) up to an n^δ factor; they together give the detection threshold stated in Theorem 7. Next, for any three vertices, the probability that they are neighbors in the planted cycle and form a triangle in A is $\Theta(\tau^2p^3)$; as a result, there are $\Theta(n^3\tau^2p^3)$ triangles in the planted cycle on average. Therefore, the second condition in (7) is a minimal condition guaranteeing the existence of triangles in the planted cycle in the first place. Further, note that if p and q are of the same order, then $n^3\tau^2p^3 = \Omega(n^3\tau^4p^6/r^3)$, so the second condition in (7) is subsumed by the first condition.

4. The recovery problem

Similar to the previous section, we start with the low-degree lower bound, whose proof suggests an optimal efficient algorithm. Then we analyze the algorithm to establish the matching upper bound. The optimal statistic for recovery turns out to be a signed count of self-avoiding walks between vertices 1 and 2, a statistic that has been used for related problems such as community detection in Hopkins and Steurer (2017) and spiked matrix models in Ding et al. (2020b).

4.1. Lower bound

A general strategy for proving low-degree lower bounds for estimation problems was proposed by Schramm and Wein (2022). We provide a lower bound in Proposition 11 that extends the one in (Schramm and Wein, 2022, Section 3.5) for the planted clique problem. Let us start with a general recovery problem with binary observations.

Definition 10 *For an integer $N \geq 1$, let $B_1, \dots, B_N$ be i.i.d. $\mathrm{Bern}(q)$ variables. Consider a latent random subset $W \subseteq [N]$ from an arbitrary prior over subsets of $[N]$. Conditional on W, we define the observation $A \in \mathbb{R}^N$ as follows. If $i \notin W$, then let $A_i := B_i$. If $i \in W$, then sample an independent $A_i \sim \mathrm{Bern}(p)$.*

Given A from the above model, we aim to estimate $\chi := \mathbb{1}\{1 \in W\}$. For a positive integer D, define

$$\mathsf{Corr}_{\leq D} := \sup_{\substack{f \in \mathbb{R}[A]_{\leq D}, \\ \mathbb{E}[f(A)^2] \neq 0}} \frac{\mathbb{E}[f(A) \cdot \chi]}{\sqrt{\mathbb{E}[f(A)^2]}}$$

as in (1). Let

$$\lambda := \frac{p - q}{\sqrt{q(1-q)}}, \tag{8}$$

which is a signal-to-noise ratio analogous to μ in (3) for the detection problem (here in the recovery problem, model $\mathcal{Q}$ is irrelevant, so r is replaced by q in the definition of λ). The following result is proved in Section A.3.

Proposition 11 *Assume the model in Definition 10. For $\beta \subseteq \alpha \subseteq [N]$, let*

$$P_{\alpha\beta} := \mathbb{P}\{\alpha \setminus W = \beta\}.$$

Suppose $P_{\alpha\alpha} > 0$ for all $\alpha \subseteq [N]$. Then we have

$$\mathsf{Corr}^2_{\leq D} \leq \sum_{\alpha \subseteq [N] \,:\, |\alpha| \leq D} \rho_\alpha^2 \, \lambda^{2|\alpha|},$$

where ρ_α is defined recursively by $\rho_\varnothing := \mathbb{P}\{1 \in W\}$ and

$$\rho_\alpha := \frac{1}{P_{\alpha\alpha}} \left(\mathbb{P}\{\alpha \cup \{1\} \subseteq W\} - \sum_{\beta \subsetneq \alpha} \rho_\beta \, P_{\alpha\beta} \right).$$

We now return to the problem of planted dense cycle and present the following result.

Theorem 12 (Recovery lower bound) *Consider Problem 4. Fix any constant $\delta \in (0, 0.1)$. No polynomial $f \in \mathbb{R}[A]_{\leq D}$ weakly recovers $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \leq \tau/2\}$ given $A \sim \mathcal{P}$ in the sense of Definition 6, if*

$$n\tau^2\lambda^2 \leq n^{-\delta}, \qquad \lambda = O(1), \qquad D = o\left(\left(\frac{\log n}{\log\log n} \right)^2 \right), \qquad \tau D^4 \leq 0.1. \tag{9}$$

Let us discuss the conditions in (9), which are analogous to those in (4). First, the main condition is $n\tau^2\lambda^2 \le n^{-\delta}$, which can be weakened to $n\tau^2\lambda^2 \le n^{-o(1)}$ for an appropriately defined $o(1)$ quantity by a closer inspection of the above proof. Also, the degree D can be polylogarithmic in n or even $n^{o(1)}$ if λ is sufficiently small. Finally, the technical condition $\tau D^4 \le 0.1$ is inactive if $n\lambda^2 \ge 1$; even if it is active, the condition is mild because the interesting regime of small-world networks is where the bandwidth $n\tau$ is much smaller than the total number of vertices n.

Theorem 12 is proved in Section A.3 and we now provide a sketch. To apply Proposition 11, we note that model $\mathcal{P}$ in Definition 1 is a special case of the model in Definition 10. Namely, let $N = \binom{n}{2}$, use an index pair $(i,j) \in \binom{[n]}{2}$ instead of a single index, and let

$$W = \left\{(i,j) \in \binom{[n]}{2} \,:\, \mathfrak{d}(z_i, z_j) \le \tau/2\right\}. \tag{10}$$

In addition, we have

$$\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\} = \mathbb{1}\{(1,2) \in W\}.$$

Proposition 11 then implies that

$$\mathsf{Corr}^2_{\le D} \le \sum_{\alpha \subseteq \binom{[n]}{2} \,:\, |\alpha| \le D} \rho_\alpha^2 \, \lambda^{2|\alpha|}, \tag{11}$$

where ρ_α is defined recursively by $\rho_\varnothing = \mathbb{P}\{\mathfrak{d}(z_1, z_2) \le \tau/2\} = \tau$, and

$$\rho_\alpha = \frac{1}{P_{\alpha\alpha}}\left(\mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} - \sum_{\beta \subsetneq \alpha} \rho_\beta \, P_{\alpha\beta}\right). \tag{12}$$

Then the bulk of the proof consists in bounding ρ_α^2 for each α using the above recursion. This is done in Section A.3, eventually leading to the bounds on $\mathsf{Corr}^2_{\le D}$ in Proposition 29. Theorem 12 then follows as a consequence.

The recursive definition (12) is similar to that for joint cumulants of the random variables χ and $(A_{ij})_{(i,j)\in\alpha}$ (see Schramm and Wein (2022)). Intuitively, for each $\alpha \subseteq \binom{[n]}{2}$, the cumulant-like quantity ρ_α measures the amount of "information" $(A_{ij})_{(i,j)\in\alpha}$ contains about the estimand χ. The above lower bound controls the total amount of information that all subgraphs with at most D edges have about χ. On the other hand, if ρ_α is large for a particular choice of α, then the corresponding subgraph $(A_{ij})_{(i,j)\in\alpha}$ may be useful for recovering χ. The analysis of ρ_α in Proposition 29 turns out to suggest that we should consider self-avoiding walks between vertices 1 and 2, which we study in the next subsection for the upper bound.

4.2. Upper bound

Similar to $\bar{A}$ in (2), we consider a standardized version $\tilde{A}$ of the observed graph, defined by

$$\tilde{A}_{ij} := \frac{A_{ij} - q}{\sqrt{q(1-q)}} \tag{13}$$

for $(i,j) \in \binom{[n]}{2}$. Compared to (2), the parameter r is replaced by q in (13) because model $\mathcal{Q}$ is irrelevant for the recovery problem. Moreover, for $\alpha \subset \binom{[n]}{2}$, define

$$\tilde{A}_\alpha := \prod_{(i,j)\in\alpha} \tilde{A}_{ij}. \tag{14}$$

As discussed above, the proof of the recovery lower bound suggests that self-avoiding walks between vertices 1 and 2 are informative about $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$, which motivates us to consider the following. Fix an integer $\ell \ge 1$. Let SAW_ℓ be the set of all length-$(\ell+1)$ self-avoiding walks from vertex 1 and to vertex 2, i.e.,

$$\mathsf{SAW}_\ell := \Big\{ \{(1, i_1), (i_1, i_2), (i_2, i_3), \dots, (i_{\ell-1}, i_\ell), (i_\ell, 2)\} : i_1, \dots, i_\ell, 1, 2 \text{ are all distinct} \Big\}. \quad (15)$$

Define the signed count of SAW_ℓ in the observed graph A as

$$T(A) := \sum_{\alpha \in \mathsf{SAW}_\ell} \tilde{A}_\alpha = \sum_{i_1 \ne \cdots \ne i_\ell \ne 1 \ne 2} \tilde{A}_{1 i_1} \tilde{A}_{i_1 i_2} \tilde{A}_{i_2 i_3} \cdots \tilde{A}_{i_{\ell-1} i_\ell} \tilde{A}_{i_\ell 2}. \quad (16)$$

As discussed above, this statistic has appeared in, e.g., Hopkins and Steurer (2017) for community detection. The following theorem shows that the statistic $T(A)$ achieves weak recovery of χ, and its proof can be found in Section A.4.

Theorem 13 (Recovery upper bound) *Consider Problem 4. Suppose that $Cq \le p \le C'q$ for constants $C' > C > 1$. For any constant $\delta \in (0, 0.1)$, fix an integer $\ell > 1/\delta$. The degree-$(\ell+1)$ polynomial $T(A)$ defined in (16) weakly recovers $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$ given $A \sim \mathcal{P}$ in the sense of Definition 6, if*

$$n\tau^2 p \ge n^\delta, \qquad \tau = o(1). \quad (17)$$

If p and q are of the same order, then we have $\lambda = \frac{p-q}{\sqrt{q(1-q)}} = \Theta(p^{1/2})$. Therefore, the main condition $n\tau^2 p \ge n^\delta$ in (17) matches the first condition in (9) up to an n^δ factor; they together give the recovery threshold stated in Theorem 7. We can also obtain a more general condition for the upper bound using Propositions 31 and 38, but the condition is not tight in the regime where $p/q \ge n^c$ for a constant $c > 0$. Proving a tight condition requires more technical work beyond the scope of this paper. Moreover, as we have explained, the condition $\tau = o(1)$ in (17) is natural because the bandwidth is usually much smaller than the total number of vertices.

We have focused on weak recovery in the sense of Definition 6 and established the detection-recovery gap in the framework of low-degree polynomials. Let us now consider the more practical problem of exactly recovering the indicator $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$ with high probability using a polynomial-time algorithm. Towards this end, fix a quantity $\epsilon \in (0, \tau/2)$ and define

$$\kappa = \kappa(\epsilon) := \frac{\mathbb{E}\big[T \,|\, \mathfrak{d}(z_1, z_2) = \frac{\tau}{2}\big] + \mathbb{E}\big[T \,|\, \mathfrak{d}(z_1, z_2) = \frac{\tau}{2} + \epsilon\big]}{2}. \quad (18)$$

By Proposition 31, the quantity κ can be computed explicitly. We then threshold the statistic $T(A)$ in (16) at κ to obtain the estimator

$$\hat{\chi} := \mathbb{1}\{T(A) \ge \kappa\}.$$

The following result is a consequence of our analysis of the statistic $T(A)$, and its proof is deferred to the end of Section A.4.

Theorem 14 *In the setting of Theorem 13, we additionally assume $\ell > 3/\delta$ and set $\epsilon := \tau n^{-\delta/4}$. Then the estimator $\hat{\chi} = \mathbb{1}\{T(A) \ge \kappa\}$ of $\chi = \mathbb{1}\{\mathfrak{d}(z_1, z_2) \le \tau/2\}$ satisfies*

$$\mathbb{E}\big[(\hat{\chi} - \chi)^2\big] = \mathbb{P}\{\hat{\chi} \ne \chi\} \le C_\ell \, \tau n^{-\delta/2}$$

for a constant $C_\ell > 0$ depending only on ℓ.

Since $\mathfrak{d}(z_1, z_2) \le \tau/2$ with probability τ, a trivial estimator $\tilde{\chi} \equiv 0$ makes an error with probability τ. Therefore, the above error probability $O(\tau n^{-\delta/2})$ is small.

An immediate consequence of the above result is that we can estimate the underlying random geometric graph consistently. To be more precise, we denote the adjacency matrix of the geometric graph by $X \in \{0,1\}^{n\times n}$, which is defined by $X_{ij} := \mathbb{1}\{\mathfrak{d}(z_i, z_j) \le \tau/2\}$.

Corollary 15 *In the setting of Theorem 14, there is an estimator $\hat{X} \in \{0,1\}^{n\times n}$ of the random geometric graph $X \in \{0,1\}^{n\times n}$ such that*

$$\mathbb{E}[\|\hat{X} - X\|_F^2] \le C_\ell\, \tau n^{2-\delta/2}$$

for a constant $C_\ell > 0$ depending only on ℓ.

This result follows immediately from Theorem 14, because by symmetry, it suffices to estimate each edge $X_{ij} = \mathbb{1}\{\mathfrak{d}(z_i, z_j) \le \tau/2\}$ in the same way as we did for $(i, j) = (1, 2)$.

Acknowledgments

C.M. was supported in part by NSF grants DMS-2053333 and DMS-2210734. Part of this work was done while A.S.W. was with the Algorithms and Randomness Center at Georgia Tech, supported by NSF grants CCF-2007443 and CCF-2106444. S.Z. was supported in part by NSF grant DMS-2053333. We thank Guy Bresler, Will Perkins, and Jiaming Xu for helpful discussions.

References

Ernesto Araya Valdivia and De Castro Yohann. Latent distance estimation for random geometric graphs. *Advances in Neural Information Processing Systems*, 32, 2019.

Vivek Bagaria, Jian Ding, David Tse, Yihong Wu, and Jiaming Xu. Hidden hamiltonian cycle recovery via linear programming. *Operations research*, 68(1):53–70, 2020.

Afonso S Bandeira, Ahmed El Alaoui, Samuel B Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The Franz-Parisi criterion and computational trade-offs in high dimensional statistics. *arXiv preprint arXiv:2205.09727*, 2022.

Albert-László Barabási. The science of networks. *Cambridge MA: Perseus*, 2012.

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

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. 2019.

Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields*, 178(3-4):1215–1289, 2020.

Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

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.

Jian Ding, Yihong Wu, Jiaming Xu, and Dana Yang. Consistent recovery threshold of hidden nearest neighbor graphs. In *Conference on Learning Theory*, pages 1540–1553. PMLR, 2020a.

Jingqiu Ding, Samuel Hopkins, and David Steurer. Estimating rank-one spikes from heavy-tailed noise via self-avoiding walks. *Advances in Neural Information Processing Systems*, 33:5576–5586, 2020b.

Quentin Duchemin and Yohann De Castro. Random geometric graph: Some recent developments and perspectives. *arXiv preprint arXiv:2203.15351*, 2022.

Ronen Eldan, Dan Mikulincer, and Hester Pieters. Community detection and percolation of information in a geometric setting. *Combinatorics, Probability and Computing*, 31(6):1048–1069, 2022.

Christophe Giraud, Yann Issartel, and Nicolas Verzelen. Localization in 1d non-parametric latent space models from pairwise affinities. *arXiv preprint arXiv:2108.03098*, 2021.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Computational lower bounds for community detection on random graphs. In *Conference on Learning Theory*, pages 899–928. PMLR, 2015.

Samuel Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

Jeannette Janssen and Aaron Smith. Reconstruction of line-embeddings of graphons. *Electronic Journal of Statistics*, 16(1):331–407, 2022.

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions, volume 2*, volume 289. John wiley & sons, 1995.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950): 289–293, 2009.

Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. Testing thresholds for high-dimensional sparse random geometric graphs. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 672–677, 2022.

Suqi Liu and Miklós Z Rácz. Phase transition in noisy high-dimensional random geometric graphs. *arXiv preprint arXiv:2103.15249*, 2021.

Suqi Liu and Miklos Z Racz. A probabilistic view of latent space graphs and phase transitions. *arXiv preprint arXiv:2110.15886*, 2021.

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

Amine Natik and Aaron Smith. Consistency of spectral seriation. *arXiv preprint arXiv:2112.04408*, 2021.

Mathew Penrose. *Random geometric graphs*, volume 5. OUP Oxford, 2003.

William S Robinson. A method for chronologically ordering archaeological deposits. *American antiquity*, 16(4):293–301, 1951.

Tselil Schramm and Alexander S Wein. Computational barriers to estimation from low-degree polynomials. *The Annals of Statistics*, 50(3):1833–1858, 2022.

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

Daniel L Sussman, Minh Tang, and Carey E Priebe. Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):48–57, 2013.

Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American journal of sociology*, 111(2):447–504, 2005.

Duncan J Watts. The "new" science of networks. *Annual review of sociology*, pages 243–270, 2004.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393 (6684):440–442, 1998.

Appendix A. Additional proofs

A.1. Detection lower bound

In this subsection, we establish Proposition 20 which leads to Theorem 8. Recall the models $\mathcal{P}$ and $\mathcal{Q}$ in Definitions 1 and 2 respectively. For $\alpha \subseteq \binom{[n]}{2}$, define

$$\eta(z;\alpha) := |\{(i,j) \in \alpha : \mathfrak{d}(z_i, z_j) \le \tau/2\}|.$$

Lemma 16 *For $\alpha \subseteq \binom{[n]}{2}$, let $v := |V(\alpha)|$ and suppose that the graph α is connected. Then we have*

$$\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha)}}\Big] \le v^{2v}\tau^{v-|\alpha|-1}.$$

Proof Suppose that $V(\alpha) = \{i_1, \dots, i_v\}$. For any realization of z, there is a unique partition $B_1 \sqcup \cdots \sqcup B_m$ of $\{i_1, \dots, i_v\}$ such that the following two conditions hold:

1. For any distinct $j, j' \in [m]$ and any $\ell \in B_j$ and $\ell' \in B_{j'}$, we have $\mathfrak{d}(z_\ell, z_{\ell'}) > \tau/2$;
2. For any $j \in [m]$, B_j cannot be partitioned into two sub-blocks satisfying Condition 1.

In other words, we partition $z_{i_1}, \dots, z_{i_v}$ into blocks so that the distance between two consecutive points in the same block is at most $\tau/2$.

Now fix a partition $\{i_1, \dots, i_v\} = B_1 \sqcup \cdots \sqcup B_m$. We claim that

$$\mathbb{P}_z\{\text{Conditions 1 and 2 are satisfied for } B_1, \dots, B_m\} \le (v\tau)^{v-m}. \tag{19}$$

To prove (19), it suffices to use Condition 2. Fix $\ell_j \in B_j$ for $j \in [m]$. By Condition 2, for any $j \in [m]$ and $\ell \in B_j$, we have $\mathfrak{d}(z_\ell, z_{\ell_j}) \le |B_j|\,\tau/2 \le v\tau/2$. For any realization of $z_{\ell_1}, \dots, z_{\ell_m}$, it holds that

$$\mathbb{P}_z\{\mathfrak{d}(z_\ell, z_{\ell_j}) \le v\tau/2 \text{ for all } \ell \in B_j \text{ and all } j \in [m] \mid z_{\ell_1}, \dots, z_{\ell_m}\} \le \prod_{j=1}^{m} (v\tau)^{|B_j|-1} = (v\tau)^{v-m}.$$

Then (19) follows.

Since the graph α is connected, there are at least $m-1$ edges between vertices $\ell \in B_j$ and $\ell' \in B_{j'}$ for distinct $j, j' \in [m]$. If Condition 1 is satisfied, then we have

$$\eta(z;\alpha) = |\{(\ell, \ell') \in \alpha : \mathfrak{d}(z_\ell, z_{\ell'}) \le \tau/2\}| \le |\alpha| - (m-1). \tag{20}$$

Combining (19) and (20), we obtain

$$\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha)}}\Big] \le \sum_{B_1 \sqcup \cdots \sqcup B_m = \{i_1, \dots, i_v\}} (v\tau)^{v-m} \cdot \frac{1}{\tau^{|\alpha|-(m-1)}} \le \sum_{B_1 \sqcup \cdots \sqcup B_m = \{i_1, \dots, i_v\}} v^v \tau^{v-|\alpha|-1}.$$

Finally, bound the number of partitions by v^v. ∎

Lemma 17 *For $\alpha \subseteq \binom{[n]}{2}$, let $v := |V(\alpha)|$, and let m be the number of connected components of the graph α. Then we have*

$$\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha)}}\Big] \le v^{2v}\tau^{v-|\alpha|-m}.$$

Proof Let $\alpha^{(1)} \sqcup \cdots \sqcup \alpha^{(m)}$ denote the partition of α into connected components, and note that

$$|\alpha| = \sum_{i=1}^{m} |\alpha^{(i)}|, \qquad v = |V(\alpha)| = \sum_{i=1}^{m} |V(\alpha^{(i)})|, \qquad \eta(z;\alpha) = \sum_{i=1}^{m} \eta(z;\alpha^{(i)}).$$

Lemma 16 shows that for each connected component $\alpha^{(i)}$, we have

$$\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha^{(i)})}}\Big] \le |V(\alpha^{(i)})|^{2|V(\alpha^{(i)})|}\,\tau^{|V(\alpha^{(i)})|-|\alpha^{(i)}|-1} \le v^{2|V(\alpha^{(i)})|}\,\tau^{|V(\alpha^{(i)})|-|\alpha^{(i)}|-1}.$$

Crucially, the random variables $\eta(z;\alpha^{(1)}),\dots,\eta(z;\alpha^{(m)})$ are independent because the connected components have mutually disjoint vertex sets and thus involve independent collections of latent variables z_j. We conclude that

$$\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha)}}\Big] = \prod_{i=1}^{m}\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha^{(i)})}}\Big] \le \prod_{i=1}^{m} v^{2|V(\alpha^{(i)})|}\,\tau^{|V(\alpha^{(i)})|-|\alpha^{(i)}|-1} = v^{2v}\tau^{v-|\alpha|-m}.$$

■

Lemma 18 *For $\alpha \subseteq \binom{[n]}{2}$, let $v := |V(\alpha)|$, and let m be the number of connected components of the graph α. Recall that $\mu = \frac{p-r}{\sqrt{r(1-r)}}$. Then we have*

$$|\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)]| \le \Big(\frac{\mu}{1-\tau}\Big)^{|\alpha|} v^{2v}\tau^{v-m}.$$

Proof We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)] &= \frac{1}{(r(1-r))^{|\alpha|/2}}\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\alpha}(A_{ij}-r)\Big] \\
&= \frac{1}{(r(1-r))^{|\alpha|/2}}\mathbb{E}_z\Big[\prod_{(i,j)\in\alpha}\mathbb{E}[A_{ij}-r \mid z]\Big] \\
&= \frac{1}{(r(1-r))^{|\alpha|/2}}\mathbb{E}_z\Big[(p-r)^{\eta(z;\alpha)}(q-r)^{|\alpha|-\eta(z;\alpha)}\Big].
\end{aligned}$$

Recall that $r = \tau p + (1-\tau)q$ so that $\frac{p-r}{q-r} = \frac{\tau-1}{\tau}$, and $\mu = \frac{p-r}{\sqrt{r(1-r)}}$. It follows that

$$\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)] = \Big(\frac{\mu\tau}{\tau-1}\Big)^{|\alpha|}\mathbb{E}_z\Big[\Big(\frac{\tau-1}{\tau}\Big)^{\eta(z;\alpha)}\Big].$$

By Lemma 17, we obtain

$$|\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)]| \le \Big(\frac{\mu\tau}{1-\tau}\Big)^{|\alpha|}\mathbb{E}_z\Big[\frac{1}{\tau^{\eta(z;\alpha)}}\Big] \le \Big(\frac{\mu\tau}{1-\tau}\Big)^{|\alpha|} v^{2v}\tau^{v-|\alpha|-m},$$

finishing the proof. ■

Lemma 19 *If the graph α has a dangling edge, i.e., an edge (i,j) where i is connected only to j in α, then $\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)] = 0$.*

Proof Let $z_{-i} = (z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)$. Since i is connected only to j in α, conditional on z_{-i}, the edges $\{A_{i'j'} : (i', j') \in \alpha \setminus \{(i, j)\}\}$ are mutually independent; further, they are independent from z_i and A_{ij}. It follows that

$$\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)] = \mathbb{E}_{z_{-i}}\Big[\prod_{(i',j')\in\alpha} \mathbb{E}_{\mathcal{P}}[\bar{A}_{i'j'} \mid z_{-i}]\Big] = 0,$$

because

$$\mathbb{E}_{\mathcal{P}}[\bar{A}_{ij} \mid z_{-i}] = \frac{\mathbb{E}_{\mathcal{P}}[A_{ij} - r \mid z_j]}{\sqrt{r(1-r)}} = \frac{\tau(p-r) + (1-\tau)(q-r)}{\sqrt{r(1-r)}} = 0$$

by the definition $r := \tau p + (1-\tau)q$. ∎

Proposition 20 *Recall (5). We have* $\mathsf{Adv}^2_{\le D} \le 2$ *in either of the following situations:*

- $2n^3\tau^4(2D)^{13}R^{6\sqrt{D}} \le 1/2$, *where* $R := \max\{2D\frac{\mu}{1-\tau}, 1\}$;
- $L(2D\frac{\mu}{1-\tau})^2 \le 1/2$ *and* $n^3\tau^4(2D)^{19}(\frac{\mu}{1-\tau})^6 \le 1/2$, *where* $L := \max\{n\tau^2(2D)^4, 1\}$.

Proof To ease the notation, we consider D such that $\sqrt{D}/3$ is an integer; the proof can be easily adapted to the general case by using floors $\lfloor\cdot\rfloor$ or ceilings $\lceil\cdot\rceil$.

We start with (5) which states

$$\mathsf{Adv}^2_{\le D} = \sum_{\alpha\subseteq\binom{[n]}{2} : |\alpha|\le D} (\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)])^2.$$

Recall that $\phi_\varnothing \equiv 1$. Let $\mathsf{c}(\alpha)$ denote the number of connected components of α. If any connected component of α has less than three edges, then it must contains a dangling edge; if the number of vertices of α exceeds the number of edges, then α also has a dangling edge. Therefore, Lemma 19 shows that $\mathbb{E}_{\mathcal{P}}[\phi_\alpha(A)] = 0$ if $|\alpha| < 3\,\mathsf{c}(\alpha)$ or $|V(\alpha)| > |\alpha|$. Then, by Lemma 18, we obtain

$$\mathsf{Adv}^2_{\le D} \le 1 + \sum_{m=1}^{D/3}\sum_{\ell=3m}^{D}\sum_{v=3m}^{D} \sum_{\substack{\alpha\subseteq\binom{[n]}{2} : \mathsf{c}(\alpha)=m,\\ |\alpha|=\ell,\, |V(\alpha)|=v}} \Big(\frac{\mu}{1-\tau}\Big)^{2\ell} v^{4v}\tau^{2v-2m}\cdot\mathbb{1}\{v\le\ell\le v^2/2\}.$$

There are at most $\binom{n}{v}\Big[\binom{\binom{v}{2}}{\ell} \wedge 2^{\binom{v}{2}}\Big]$ graphs α with $|V(\alpha)| = v$ and $|\alpha| = \ell$. By the inequalities $\binom{n}{v} \le n^v$, $\binom{\binom{v}{2}}{\ell} \le v^{2\ell} \le (2D)^{2\ell}$, and $2^{\binom{v}{2}} \le 2^{v^2}$, we have

$$\mathsf{Adv}^2_{\le D} \le 1 + \sum_{m=1}^{D/3}\sum_{\ell=3m}^{D}\sum_{v=3m}^{D} \Big[(2D)^{2\ell} \wedge 2^{v^2}\Big]\cdot\Big(\frac{\mu}{1-\tau}\Big)^{2\ell}(n\tau^2(2D)^4)^v\tau^{-2m}\cdot\mathbb{1}\{v\le\ell\le v^2/2\}. \tag{21}$$

Let us consider two cases.

Case 1: We bound (21) by splitting it into the following terms according the value of v:

$$\mathsf{Adv}^2_{\le D} \le 1 + \sum_{m=1}^{D/3} \sum_{v=3m}^{\sqrt{D}} \sum_{\ell=3m}^{D} 2^{v^2} \Big(\frac{\mu}{1-\tau}\Big)^{2\ell} (n\tau^2(2D)^4)^v \tau^{-2m} \cdot \mathbb{1}\{\ell \le v^2/2\} \tag{22a}$$

$$+ \sum_{m=1}^{D/3} \sum_{v=\sqrt{D}\vee 3m}^{D} \sum_{\ell=3m}^{D} (2D)^{2\ell} \Big(\frac{\mu}{1-\tau}\Big)^{2\ell} (n\tau^2(2D)^4)^v \tau^{-2m}. \tag{22b}$$

We then bound (22a) and (22b) respectively. Recall that $R := \max\{2D\frac{\mu}{1-\tau}, 1\}$. By the assumption $2n^3\tau^4(2D)^{13}R^{6\sqrt{D}} \le 1/2$, we have

$$n\tau^2(2D)^4 R^{\sqrt{D}} \le 1/2, \qquad n^3\tau^4(2D)^{12}R^{3\sqrt{D}} \le 1/2.$$

For $v \le \sqrt{D}$, we have $2\ell \le v^2 \le \sqrt{D}\, v$, so the sum in (22a) is bounded by

$$\begin{aligned}\sum_{m=1}^{D/3} \sum_{v=3m}^{\sqrt{D}} D\, R^{\sqrt{D}\, v} (n\tau^2(2D)^4)^v \tau^{-2m} &\le 2D \sum_{m=1}^{D/3} (n\tau^2(2D)^4 R^{\sqrt{D}})^{3m} \tau^{-2m} \\ &\le 4D\, n^3\tau^4(2D)^{12}R^{3\sqrt{D}} \le 1/2\end{aligned}$$

by the assumption $2n^3\tau^4(2D)^{13}R^{6\sqrt{D}} \le 1/2$. Next, using

$$n\tau^2(2D)^4 \le 1/2, \qquad n^3\tau^4(2D)^{12} \le 1/2,$$

we see that the sum in (22b) is bounded by

$$\begin{aligned}&\sum_{m=1}^{D/3} DR^{2D} (n\tau^2(2D)^4)^{\sqrt{D}\vee 3m} \tau^{-2m} \\ &\le \sum_{m=1}^{\sqrt{D}/3} DR^{2D} (n\tau^2(2D)^4)^{\sqrt{D}} \tau^{-2\sqrt{D}/3} + \sum_{m=\sqrt{D}/3}^{D/3} DR^{2D} (n^3\tau^4(2D)^{12})^m \\ &\le D^{3/2} R^{2D} (n^3\tau^4(2D)^{12})^{\sqrt{D}/3} + DR^{2D} (n^3\tau^4(2D)^{12})^{\sqrt{D}/3} \\ &\le \Big(n^3\tau^4(2D)^{12}(2D)^{9/(2\sqrt{D})} R^{6\sqrt{D}}\Big)^{\sqrt{D}/3} \le 1/2\end{aligned}$$

by the assumption $2n^3\tau^4(2D)^{13}R^{6\sqrt{D}} \le 1/2$. Combining the two terms yields $\mathsf{Adv}^2_{\le D} \le 2$.

Case 2: It follows from (21) that

$$\begin{aligned}\mathsf{Adv}^2_{\le D} &\le 1 + \sum_{m=1}^{D/3} \sum_{\ell=3m}^{D} \sum_{v=3m}^{\ell} n^v (2D)^{2\ell} \Big(\frac{\mu}{1-\tau}\Big)^{2\ell} (2D)^{4v} \tau^{2v-2m} \\ &= 1 + \sum_{m=1}^{D/3} \tau^{-2m} \sum_{\ell=3m}^{D} \left(\Big(2D\frac{\mu}{1-\tau}\Big)^2\right)^{\ell} \sum_{v=3m}^{\ell} (n\tau^2(2D)^4)^v.\end{aligned}$$

Recall that $L := \max\{n\tau^2(2D)^4, 1\}$, $L(2D\frac{\mu}{1-\tau})^2 \le 1/2$, and $n^3\tau^4(2D)^{19}(\frac{\mu}{1-\tau})^6 \le 1/2$. We have

$$\begin{aligned}
\mathsf{Adv}^2_{\le D} &\le 1 + \sum_{m=1}^{D/3} \tau^{-2m} \sum_{\ell=3m}^{D} \left(\left(2D\frac{\mu}{1-\tau}\right)^2\right)^{\ell} (\ell - 3m + 1)(n\tau^2(2D)^4)^{3m} L^{\ell-3m} \\
&\le 1 + D\sum_{m=1}^{D/3} (L^{-3}n^3\tau^4(2D)^{12})^m \sum_{\ell=3m}^{D} \left(L\left(2D\frac{\mu}{1-\tau}\right)^2\right)^{\ell} \\
&\le 1 + 2D\sum_{m=1}^{D/3} (L^{-3}n^3\tau^4(2D)^{12})^m \left(L\left(2D\frac{\mu}{1-\tau}\right)^2\right)^{3m} \\
&= 1 + 2D\sum_{m=1}^{D/3} \left(n^3\tau^4(2D)^{18}\left(\frac{\mu}{1-\tau}\right)^6\right)^m \\
&\le 1 + 4Dn^3\tau^4(2D)^{18}\left(\frac{\mu}{1-\tau}\right)^6 \le 2,
\end{aligned}$$

finishing the proof. ■

We are ready to prove Theorem 8.

Proof [Proof of Theorem 8] To prove that no polynomial of degree at most D strongly separates $\mathcal{P}$ and $\mathcal{Q}$, recall the discussion after Theorem 8: it suffices to show that the advantage in (5) is $O(1)$. Note that it is bounded by 2 in Proposition 20, which we now apply. It suffices to verify that (4) implies the assumptions of Proposition 20. To this end, we consider two cases:

- If $\mu = \tilde{\Theta}(1)$, $n^3\tau^4 \le n^{-\delta}$, and $D = o\left((\frac{\log n}{\log\log n})^2\right)$, then we can check the first set of conditions in Proposition 20: $R = \max\{2D\frac{\mu}{1-\tau}, 1\} = \tilde{O}(1)$ and $2n^3\tau^4(2D)^{13}R^{6\sqrt{D}} \le n^3\tau^4 \cdot n^{o(1)} \le 1/2$.

- Next, suppose that $\mu \le (\log n)^{-100}$, $n^3\tau^4\mu^6 \le n^{-\delta}$, and $D \le (\log n)^{10}$. We can check the second set of conditions in Proposition 20 by further considering two subcases:

  - If $\tau \le n^{-1/2}$, then $L = \max\{n\tau^2(2D)^4, 1\} \le (2D)^4$, $L\left(2D\frac{\mu}{1-\tau}\right)^2 \le \mu^2(4D)^6 \le 1/2$, and $n^3\tau^4(2D)^{19}\left(\frac{\mu}{1-\tau}\right)^6 \le n^{-\delta}(4D)^{19} \le 1/2$.

  - If $\tau > n^{-1/2}$, then $L = n\tau^2(2D)^4$ and $L\left(2D\frac{\mu}{1-\tau}\right)^2 \le n\tau^2\mu^2(4D)^6 \le 1/2$ because $n\tau^2\mu^2 = (n^3\tau^6\mu^6)^{1/3} \le (n^3\tau^4\mu^6)^{1/3} \le n^{-\delta/3}$. Finally, $n^3\tau^4(2D)^{19}\left(\frac{\mu}{1-\tau}\right)^6 \le 1/2$ as in the previous subcase.

Combining all the cases completes the proof. ■

A.2. Detection upper bound

While it suffices to focus on the signed triangle count for the upper bound, we consider cliques with $v \geq 3$ vertices, because some intermediate results hold for a general $v \geq 3$ and may be interesting in their own right. Define

$$S_v(A) := \sum_{H \in \binom{[n]}{v}} \prod_{(i,j) \in \binom{H}{2}} \bar{A}_{ij}.$$

Recall the models $\mathcal{P}$ and $\mathcal{Q}$ in Definitions 1 and 2 respectively.

Proposition 21 *We have*

$$\mathbb{E}_{\mathcal{Q}}[S_v(A)] = 0, \qquad \mathrm{Var}_{\mathcal{Q}}(S_v(A)) = \binom{n}{v}.$$

Proof It is clear that $\mathbb{E}_{\mathcal{Q}}[S_v(A)] = 0$. Moreover, the variance of $S_v(A)$ under $\mathcal{Q}$ is equal to

$$\mathbb{E}_{\mathcal{Q}}[S_v(A)^2] = \sum_{H,H' \in \binom{[n]}{v}} \mathbb{E}_{\mathcal{Q}}\Big[\prod_{(i,j) \in \binom{H}{2}} \prod_{(i',j') \in \binom{H'}{2}} \bar{A}_{ij} \bar{A}_{i'j'} \Big] = \sum_{H \in \binom{[n]}{v}} \mathbb{E}_{\mathcal{Q}}\Big[\prod_{(i,j) \in \binom{H}{2}} \bar{A}_{ij}^2 \Big] = \binom{n}{v}.$$

■

For $H \in \binom{[n]}{v}$, define

$$\zeta(z; H) := \left| \left\{ (i,j) \in \binom{H}{2} : \mathfrak{d}(z_i, z_j) \leq \tau/2 \right\} \right|. \tag{23}$$

Lemma 22 *Suppose that* $3 \leq v \leq n$ *and* $0 < \tau \leq \frac{1}{2^{v+1}(1+2v^{v-1})}$. *If* $|H| = v$, *then*

$$(-1)^{\binom{v}{2}} \cdot \mathbb{E}_z\Big[\Big(\frac{\tau - 1}{\tau} \Big)^{\zeta(z;H)} \Big] \geq \frac{1}{2^{v+1} \tau^{\binom{v}{2} - v + 1}}.$$

Proof Without loss of generality, we assume that $H = [v]$. It holds that

$$\begin{aligned}
&(-1)^{\binom{v}{2}} \cdot \mathbb{E}_z\Big[\Big(\frac{\tau - 1}{\tau} \Big)^{\zeta(z;H)} \Big] \\
&= (-1)^{\binom{v}{2}} \cdot \sum_{m=0}^{\binom{v}{2}} \mathbb{P}_z\{\zeta(z;H) = m\} \cdot \Big(\frac{\tau - 1}{\tau} \Big)^m \\
&\geq \mathbb{P}_z\left\{ \zeta(z;H) = \binom{v}{2} \right\} \cdot \Big(\frac{1-\tau}{\tau} \Big)^{\binom{v}{2}} - \sum_{m=0}^{\binom{v}{2}-1} \mathbb{P}_z\{\zeta(z;H) = m\} \cdot \Big(\frac{1-\tau}{\tau} \Big)^m.
\end{aligned} \tag{24}$$

We now further bound this quantity from below.

First, conditional on any realization of z_1, it holds that

$$\mathbb{P}_z\{\mathfrak{d}(z_1, z_i) \leq \tau/4 \text{ for all } i \in H \mid z_1\} = (\tau/2)^{v-1}.$$

If $\mathfrak{d}(z_1, z_i) \le \tau/4$ for all $i \in H$, then $\mathfrak{d}(z_i, z_j) \le \tau/2$ for all $i, j \in H$ so that $\zeta(z; H) = \binom{v}{2}$. Consequently,

$$\mathbb{P}_z\left\{\zeta(z;H) = \binom{v}{2}\right\} \cdot \left(\frac{1-\tau}{\tau}\right)^{\binom{v}{2}} \ge \left(\frac{\tau}{2}\right)^{v-1} \cdot \frac{1-\tau\binom{v}{2}}{\tau^{\binom{v}{2}}} \ge \frac{1}{2^v \tau^{\binom{v}{2}-v+1}}, \tag{25}$$

where the last step holds because $\tau \le \frac{1}{v(v-1)}$ by assumption.

Second, we have

$$\sum_{m=0}^{\binom{v}{2}-v} \mathbb{P}_z\{\zeta(z;H) = m\} \cdot \left(\frac{1-\tau}{\tau}\right)^m \le \left(\sum_{m=0}^{\binom{v}{2}-v} \mathbb{P}_z\{\zeta(z;H) = m\}\right) \cdot \frac{1}{\tau^{\binom{v}{2}-v}} \le \frac{1}{\tau^{\binom{v}{2}-v}}. \tag{26}$$

Third, for $\binom{v}{2} - v + 2 \le m \le \binom{v}{2} - 1$ and $\zeta(z; H) = m$, the graph on H with the edge set

$$\left\{(i,j) \in \binom{H}{2} : \mathfrak{d}(z_i, z_j) \le \tau/2\right\} \tag{27}$$

must be connected. As a result, $\mathfrak{d}(z_1, z_i) \le v\tau/2$ for all $i \in H$. Conditional on any realization of z_1, it holds that

$$\mathbb{P}_z\{\mathfrak{d}(z_1, z_i) \le v\tau/2 \text{ for all } i \in H \mid z_1\} \le (v\tau)^{v-1}.$$

Then we obtain

$$\begin{aligned}\sum_{m=\binom{v}{2}-v+2}^{\binom{v}{2}-1} \mathbb{P}_z\{\zeta(z;H) = m\} \cdot \left(\frac{1-\tau}{\tau}\right)^m &\le \left(\sum_{m=\binom{v}{2}-v+2}^{\binom{v}{2}-1} \mathbb{P}_z\{\zeta(z;H) = m\}\right) \cdot \frac{1}{\tau^{\binom{v}{2}-1}} \\ &\le (v\tau)^{v-1} \cdot \frac{1}{\tau^{\binom{v}{2}-1}} = \frac{v^{v-1}}{\tau^{\binom{v}{2}-v}}. \end{aligned} \tag{28}$$

Fourth, for $\zeta(z; H) = \binom{v}{2} - v + 1$, the graph on H with the edge set (27) is either connected or has only one isolated vertex z_{i^*}. Let $j^* = j^*(i^*)$ be any vertex in H not equal to i^*. Then $\mathfrak{d}(z_{j^*}, z_i) \le v\tau/2$ for all $i \in H \setminus \{i^*\}$. Conditional on any realization of z_{j^*}, it holds that

$$\mathbb{P}_z\{\mathfrak{d}(z_{j^*}, z_i) \le v\tau/2 \text{ for all } i \in H \setminus \{i^*\} \mid z_{j^*}\} \le (v\tau)^{v-2}.$$

Then we obtain

$$\mathbb{P}_z\left\{\zeta(z;H) = \binom{v}{2} - v + 1\right\} \cdot \left(\frac{1-\tau}{\tau}\right)^{\binom{v}{2}-v+1} \le v(v\tau)^{v-2} \cdot \frac{1}{\tau^{\binom{v}{2}-v+1}} \le \frac{v^{v-1}}{\tau^{\binom{v}{2}-v}}, \tag{29}$$

where we used the assumption $v \ge 3$ in the last step.

Finally, combining (24), (25), (26), (28), and (29) yields that

$$\left|\mathbb{E}_z\left[\left(\frac{\tau-1}{\tau}\right)^{\zeta(z;H)}\right]\right| \ge \frac{1}{2^v \tau^{\binom{v}{2}-v+1}} - \frac{1}{\tau^{\binom{v}{2}-v}} - \frac{v^{v-1}}{\tau^{\binom{v}{2}-v}} - \frac{v^{v-1}}{\tau^{\binom{v}{2}-v}} \ge \frac{1}{2^{v+1}\tau^{\binom{v}{2}-v+1}},$$

since $\frac{1}{2^v\tau} \ge 2(1 + 2v^{v-1})$ by assumption. ∎

Proposition 23 *Suppose that* $3 \le v \le n$ *and* $0 < \tau \le \frac{1}{2^{v+1}(1+2v^{v-1})}$. *It holds that*

$$|\mathbb{E}_{\mathcal{P}}[S_v(A)]| \ge \binom{n}{v}\Big(\frac{\mu}{1-\tau}\Big)^{\binom{v}{2}}\frac{\tau^{v-1}}{2^{v+1}}.$$

Proof We have

$$\begin{aligned}\mathbb{E}_{\mathcal{P}}[S_v(A)] &= \sum_{H\in\binom{[n]}{v}}\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\bar{A}_{ij}\Big]\\ &= \sum_{H\in\binom{[n]}{v}}\mathbb{E}_{z}\Big[\prod_{(i,j)\in\binom{H}{2}}\mathbb{E}_{\mathcal{P}}[\bar{A}_{ij}\mid z]\Big]\\ &= \frac{1}{(r(1-r))^{\binom{v}{2}/2}}\sum_{H\in\binom{[n]}{v}}\mathbb{E}_{z}\Big[(p-r)^{\zeta(z;H)}(q-r)^{\binom{v}{2}-\zeta(z;H)}\Big].\end{aligned}$$

Recall that $r=\tau p+(1-\tau)q$ so that $\frac{p-r}{q-r}=\frac{\tau-1}{\tau}$, and $\mu=\frac{p-r}{\sqrt{r(1-r)}}$. It follows that

$$\mathbb{E}_{\mathcal{P}}[S_v(A)] = \Big(\frac{\mu\tau}{\tau-1}\Big)^{\binom{v}{2}}\sum_{H\in\binom{[n]}{v}}\mathbb{E}_{z}\Big[\Big(\frac{\tau-1}{\tau}\Big)^{\zeta(z;H)}\Big].$$

Then, by Lemma 22, we obtain

$$|\mathbb{E}_{\mathcal{P}}[S_v(A)]| \ge \Big(\frac{\mu\tau}{1-\tau}\Big)^{\binom{v}{2}}\binom{n}{v}\frac{1}{2^{v+1}\tau^{\binom{v}{2}-v+1}},$$

completing the proof. ∎

Proposition 24 *There is an absolute constant* $C>0$ *such that*

$$\mathrm{Var}_{\mathcal{P}}(S_3(A)) \le \frac{C}{r^3(1-r)^3}\Big(n^4(\tau p+q+r^2)(r-q)^2(p-q)^2+n^3(q^3+r^6+\tau pq^2+\tau pr^4+\tau^2p^3)\Big).$$

Proof For brevity, let $\sigma:=\sqrt{r(1-r)}$ in this proof. The variance of $S_3(A)$ under $\mathcal{P}$ is

$$\begin{aligned}&\mathrm{Var}_{\mathcal{P}}(S_3(A)) = \mathbb{E}_{\mathcal{P}}[S_3(A)^2]-\mathbb{E}_{\mathcal{P}}[S_3(A)]^2\\ &= \sum_{H,H'\in\binom{[n]}{3}}\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big]-\sum_{H,H'\in\binom{[n]}{3}}\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\bar{A}_{ij}\Big]\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{i'j'}\Big].\end{aligned}$$

For fixed $H,H'\subseteq[n]$ with $|H|=|H'|=3$, consider the following cases:

- $|H\cap H'|=0$: We have that $\{z_i : i\in H\}$ and $\{z_{i'} : i'\in H'\}$ are independent, and consequently, $\{\bar{A}_{ij}:(i,j)\in\binom{H}{3}\}$ and $\{\bar{A}_{i'j'}:(i',j')\in\binom{H'}{3}\}$ are independent. Therefore,

$$\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big]-\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\bar{A}_{ij}\Big]\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{i'j'}\Big]=0.$$

- $|H \cap H'|= 1$: Suppose $H \cap H' = \{i^*\}$. Conditional on z_{i^*}, we have that $\{z_i : i \in H \setminus \{i^*\}\}$ and $\{z_{i'} : i' \in H' \setminus \{i^*\}\}$ are independent. Moreover, $\{\bar{A}_{ij} : (i,j) \in \binom{H}{3}\}$ and $\{\bar{A}_{i'j'} : (i',j') \in \binom{H'}{3}\}$ are conditionally independent, and their distributions are not changed by the conditioning on z_{i^*}. Therefore, we still have

$$\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big] - \mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\bar{A}_{ij}\Big]\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{i'j'}\Big] = 0.$$

- $|H \cap H'|= 2$: Without loss of generality, suppose that $H = \{1,2,3\}$ and $H' = \{1,2,4\}$. Then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big] \\
&= \mathbb{E}_z\Big[\mathbb{E}_{\mathcal{P}}[\bar{A}_{12}^2 \mid z]\cdot\mathbb{E}_{\mathcal{P}}[\bar{A}_{13} \mid z]\cdot\mathbb{E}_{\mathcal{P}}[\bar{A}_{23} \mid z]\cdot\mathbb{E}_{\mathcal{P}}[\bar{A}_{14} \mid z]\cdot\mathbb{E}_{\mathcal{P}}[\bar{A}_{24} \mid z]\Big] \\
&= \frac{1}{\sigma^6}\,\mathbb{E}_z\Big[(p(1-r)^2+(1-p)r^2)^{\mathbb{1}\{\mathfrak{d}(z_1,z_2)\le\tau/2\}}(q(1-r)^2+(1-q)r^2)^{\mathbb{1}\{\mathfrak{d}(z_1,z_2)>\tau/2\}} \\
&\qquad\qquad \cdot (p-r)^{\tilde{\zeta}(z)}(q-r)^{4-\tilde{\zeta}(z)}\Big],
\end{aligned}$$

  where $\tilde{\zeta}(z) := |\{(i,j) \in \{(1,3),(2,3),(1,4),(2,4) : \mathfrak{d}(z_i,z_j) \le \tau/2\}|$. We have the following:

  - It is obvious that

$$\mathbb{P}_z\{\mathfrak{d}(z_1,z_2) \le \tau/2,\, \tilde{\zeta}(z) = 0\} \le \tau, \qquad \mathbb{P}_z\{\mathfrak{d}(z_1,z_2) > \tau/2,\, \tilde{\zeta}(z) = 0\} \le 1.$$

  - Condition on any realization of (z_1,z_2). If $1 \le \tilde{\zeta}(z) \le 2$, then one of the following four events must occur: (1) $\mathfrak{d}(z_1,z_3) \le \tau/2$, (2) $\mathfrak{d}(z_2,z_3) \le \tau/2$, (3) $\mathfrak{d}(z_1,z_4) \le \tau/2$, or (4) $\mathfrak{d}(z_2,z_4) \le \tau/2$; this holds with conditional probability at most 4τ. Therefore,

$$\mathbb{P}_z\{1 \le \tilde{\zeta}(z) \le 2 \mid z_1,z_2\} \le 4\tau,$$

    so we obtain

$$\mathbb{P}_z\{\mathfrak{d}(z_1,z_2) \le \tau/2,\, 1 \le \tilde{\zeta}(z) \le 2\} \le 4\tau^2, \qquad \mathbb{P}_z\{\mathfrak{d}(z_1,z_2) > \tau/2,\, 1 \le \tilde{\zeta}(z) \le 2\} \le 4\tau.$$

  - Condition on any realization of (z_1,z_2). If $3 \le \tilde{\zeta}(z) \le 4$, then one of the following four events must occur: (1) $\mathfrak{d}(z_1,z_3) \le \tau/2$ and $\mathfrak{d}(z_1,z_4) \le \tau/2$, (2) $\mathfrak{d}(z_1,z_3) \le \tau/2$ and $\mathfrak{d}(z_2,z_4) \le \tau/2$, (3) $\mathfrak{d}(z_2,z_3) \le \tau/2$ and $\mathfrak{d}(z_1,z_4) \le \tau/2$, (4) $\mathfrak{d}(z_2,z_3) \le \tau/2$ and $\mathfrak{d}(z_2,z_4) \le \tau/2$; this holds with conditional probability at most $4\tau^2$. Therefore,

$$\mathbb{P}_z\{3 \le \tilde{\zeta}(z) \le 4 \mid z_1,z_2\} \le 4\tau^2,$$

    so we obtain

$$\mathbb{P}_z\{\mathfrak{d}(z_1,z_2) \le \tau/2,\, 3 \le \tilde{\zeta}(z) \le 4\} \le 4\tau^3, \qquad \mathbb{P}_z\{\mathfrak{d}(z_1,z_2) > \tau/2,\, 3 \le \tilde{\zeta}(z) \le 4\} \le 4\tau^2.$$

Combining the above bounds, we see that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big]\\
&=\frac{1}{\sigma^6}\Big[\sum_{\ell=0}^{4}\mathbb{P}_z\{\mathfrak{d}(z_1,z_2)\le\tau/2,\,\tilde{\zeta}(z)=\ell\}\cdot(p(1-r)^2+(1-p)r^2)(p-r)^\ell(q-r)^{4-\ell}\\
&\qquad+\sum_{\ell=0}^{4}\mathbb{P}_z\{\mathfrak{d}(z_1,z_2)>\tau/2,\,\tilde{\zeta}(z)=\ell\}\cdot(q(1-r)^2+(1-q)r^2)(p-r)^\ell(q-r)^{4-\ell}\Big]\\
&\le\frac{1}{\sigma^6}\Big[(p(1-r)^2+(1-p)r^2)\Big(\tau(r-q)^4+4\tau^2(p-r)^2(r-q)^2+4\tau^3(p-r)^4\Big)\\
&\qquad+(q(1-r)^2+(1-q)r^2)\Big((r-q)^4+4\tau(p-r)^2(r-q)^2+4\tau^2(p-r)^4\Big)\Big],
\end{aligned}$$

where we omitted negative terms where ℓ is odd. Recall that $\tau(p-r)=(1-\tau)(r-q)\le r-q$. Also, the condition $0<q<r<p<1$ implies that

$$p(1-r)^2+(1-p)r^2\le p+r^2\le 2p,\qquad q(1-r)^2+(1-q)r^2\le q+r^2. \tag{30}$$

It then follows that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big]\\
&\le\frac{1}{\sigma^6}\Big(\tau(p(1-r)^2+(1-p)r^2)+(q(1-r)^2+(1-q)r^2)\Big)\Big((r-q)^2+2\tau(p-r)^2\Big)^2\\
&\le\frac{1}{\sigma^6}(2\tau p+q+r^2)\Big((r-q)^2+2(r-q)(p-r)\Big)^2\\
&\le\frac{4}{\sigma^6}(2\tau p+q+r^2)(r-q)^2(p-q)^2.
\end{aligned}$$

- $|H\cap H'|=3$: Without loss of generality, suppose that $H=H'=\{1,2,3\}$. Then

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big]=\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\bar{A}_{ij}^2\Big]=\mathbb{E}_z\Big[\prod_{(i,j)\in\binom{H}{2}}\mathbb{E}_{\mathcal{P}}[\bar{A}_{ij}^2\mid z]\Big]\\
&=\frac{1}{\sigma^6}\,\mathbb{E}_z\Big[(p(1-r)^2+(1-p)r^2)^{\zeta(z;H)}(q(1-r)^2+(1-q)r^2)^{3-\zeta(z;H)}\Big].
\end{aligned}$$

We have $\mathbb{P}_z\{\mathfrak{d}(z_1,z_2)\le\tau/2\}=\tau$, so by symmetry, $\mathbb{P}_z\{\zeta(z;H)=1\}\le 3\tau$. Moreover, let us condition on any realization of z_1. If $\zeta(z;H)\ge 2$, then $\mathfrak{d}(z_1,z_2)\le\tau$ and $\mathfrak{d}(z_1,z_3)\le\tau$, which occurs with conditional probability at most $(2\tau)^2$. Therefore, $\mathbb{P}_z\{\zeta(z;H)\ge 2\}\le$

$(2\tau)^2$. We then obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\Big[\prod_{(i,j)\in\binom{H}{2}}\prod_{(i',j')\in\binom{H'}{2}}\bar{A}_{ij}\bar{A}_{i'j'}\Big]\\
&=\frac{1}{\sigma^6}\sum_{\ell=0}^{3}\mathbb{P}_z\{\zeta(z;H)=\ell\}\cdot(p(1-r)^2+(1-p)r^2)^{\ell}(q(1-r)^2+(1-q)r^2)^{3-\ell}\\
&\le\frac{1}{\sigma^6}\Big[(q(1-r)^2+(1-q)r^2)^3+3\tau(p(1-r)^2+(1-p)r^2)(q(1-r)^2+(1-q)r^2)^2\\
&\quad+(2\tau)^2\Big((p(1-r)^2+(1-p)r^2)^2(q(1-r)^2+(1-q)r^2)+(p(1-r)^2+(1-p)r^2)^3\Big)\Big]\\
&\le\frac{1}{\sigma^6}\Big[(q+r^2)^3+6\tau p(q+r^2)^2+64\tau^2p^3\Big],
\end{aligned}$$

where the last step follows from (30).

In summary, we have

$$\begin{aligned}
&\mathrm{Var}_{\mathcal{P}}(S_3(A))\\
&\le\sum_{\substack{H,H'\in\binom{[n]}{3}:\\|H\cap H'|=2}}\frac{4}{\sigma^6}(2\tau p+q+r^2)(r-q)^2(p-q)^2+\sum_{H\in\binom{[n]}{3}}\frac{1}{\sigma^6}\Big[(q+r^2)^3+6\tau p(q+r^2)^2+64\tau^2p^3\Big]\\
&\le\frac{C}{\sigma^6}\Big(n^4(\tau p+q+r^2)(r-q)^2(p-q)^2+n^3(q^3+r^6+\tau pq^2+\tau pr^4+\tau^2p^3)\Big)
\end{aligned}$$

for an absolute constant $C>0$. ■

We are ready to prove Theorem 9.

Proof [Proof of Theorem 9] In view of the assumption $p\ge Cq$ for a constant $C>1$ and the definition $r=\tau p+(1-\tau)q$ where $\tau\le 1/2$, we have $\mu=\frac{p-r}{\sqrt{r(1-r)}}=\Theta(p/r^{1/2})$. By Propositions 21 and 23, we obtain

$$|\mathbb{E}_{\mathcal{P}}[S_3(A)]-\mathbb{E}_{\mathcal{Q}}[S_3(A)]|\ge\binom{n}{3}\Big(\frac{\mu}{1-\tau}\Big)^3\frac{\tau^2}{16}=\Omega(n^3\tau^2p^3/r^{3/2})$$

and

$$\mathrm{Var}_{\mathcal{Q}}(S_3(A))\le n^3.$$

Moreover, the definition $r=\tau p+(1-\tau)q$ where $\tau\le 1/2$ implies $r-q=\frac{\tau}{1-\tau}(p-r)\le 2\tau p$. Hence, the bound in Proposition 24 simplifies to

$$\mathrm{Var}_{\mathcal{P}}(S_3(A))=O\Big(\frac{1}{r^3}(n^4\tau^3p^5+n^4\tau^2p^4r+n^3r^3+n^3\tau pr^2+n^3\tau^2p^3)\Big).$$

Consequently, for $S_3(A)$ to strongly separate $\mathcal{P}$ and $\mathcal{Q}$, it suffices to have

$$\sqrt{n^3+n^4\tau^3p^5/r^3+n^4\tau^2p^4/r^2+n^3\tau p/r+n^3\tau^2p^3/r^3}=o(n^3\tau^2p^3/r^{3/2}),$$

which (by dividing the square of the RHS by each term on the LHS) is equivalent to

$$\min\left\{n^3\tau^4p^6/r^3, n^2\tau p, n^2\tau^2p^2/r, n^3\tau^3p^5/r^2, n^3\tau^2p^3\right\} = \omega(1).$$

The first and the last quantity on the LHS are assumed to be $\omega(1)$ in (7), and the middle three quantities are $\omega(1)$ because $(n^2\tau p)^3 \geq (n^3\tau^2p^3)^2$, $(n^2\tau^2p^2/r)^3 \geq (n^3\tau^4p^6/r^3)\cdot(n^3\tau^2p^3)$, and $(n^3\tau^3p^5/r^2)^3 \geq (n^3\tau^4p^6/r^3)^2\cdot(n^3\tau^2p^3)$. ■

A.3. Recovery lower bound

We first prove Proposition 11.

Proof [Proof of Proposition 11] Recall Definition 10. Define $Z, Y \in \mathbb{R}^N$ by $Z_i := \frac{B_i - q}{\sqrt{q(1-q)}}$ and $Y_i := \frac{A_i - q}{\sqrt{q(1-q)}}$ for $i \in [N]$. Since any polynomial of in $(A_i)_{i\in[N]}$ is also a polynomial in $(Y_i)_{i\in[N]}$ of the same degree (and vice versa), we have

$$\mathsf{Corr}_{\leq D} = \sup_{\substack{f\in\mathbb{R}[Y]_{\leq D},\\ \mathbb{E}[f(Y)^2]\neq 0}} \frac{\mathbb{E}[f(Y)\cdot\chi]}{\sqrt{\mathbb{E}[f(Y)^2]}}.$$

For $f \in \mathbb{R}[Y]_{\leq D}$, we can write

$$f(Y) = \sum_{\alpha\subseteq[N]\,:\,|\alpha|\leq D} \hat{f}_\alpha\, Y^\alpha,$$

where $Y^\alpha := \prod_{i\in\alpha} Y_i$ and $\hat{f}_\alpha$ denotes the coefficient of f in the basis $\{Y^\alpha : \alpha \subseteq [N]\}$.

Recall that $\lambda = \frac{p-q}{\sqrt{q(1-q)}}$. Then we have $\mathbb{E}[Y_i \mid i \in W] = \lambda$. It holds that

$$\mathbb{E}[f(Y)\,\chi] = \sum_{\alpha\subseteq[N]\,:\,|\alpha|\leq D} \hat{f}_\alpha\, \mathbb{E}[Y^\alpha\,\chi] =: \langle \hat{f}, v\rangle,$$

where

$$v_\alpha := \mathbb{E}[Y^\alpha\,\chi] = \mathbb{P}\{\alpha\cup\{1\}\subseteq W\}\cdot\lambda^{|\alpha|}. \tag{31}$$

Moreover, by Jensen's inequality,

$$\mathbb{E}[f(Y)^2] \geq \mathbb{E}\Big[(\mathbb{E}[f(Y)\mid Z])^2\Big] =: \mathbb{E}[g(Z)^2],$$

where

$$g(Z) := \mathbb{E}[f(Y)\mid Z] = \sum_{\alpha\subseteq[N]\,:\,|\alpha|\leq D} \hat{f}_\alpha\, \mathbb{E}[Y^\alpha \mid Z].$$

Moreover, we have

$$\mathbb{E}[Y^\alpha \mid Z] = \sum_{\beta\subseteq\alpha} \mathbb{P}\{\alpha\setminus W = \beta\}\cdot Z^\beta\lambda^{|\alpha|-|\beta|}.$$

Together with the definitions of $g(Z)$ and $P_{\alpha\beta}$, this implies that

$$g(Z) = \sum_{\beta\subseteq[N]\,:\,|\beta|\leq D} Z^\beta \sum_{\alpha\subseteq[N]\,:\,\alpha\supseteq\beta,\,|\alpha|\leq D} \hat{f}_\alpha\,\lambda^{|\alpha|-|\beta|}\,P_{\alpha\beta} = \sum_{\beta\subseteq[N]\,:\,|\beta|\leq D} \hat{g}_\beta Z^\beta,$$

where

$$\hat{g}_\beta = \sum_{\alpha \subseteq [N] \,:\, \alpha \supseteq \beta,\, |\alpha| \le D} \hat{f}_\alpha \, \lambda^{|\alpha|-|\beta|} \, P_{\alpha\beta}.$$

Therefore, we have $\hat{g} = M\hat{f}$ where the matrix M is indexed by $\beta, \alpha \subseteq [N]$ with $|\beta|, |\alpha| \le D$, and M is defined by

$$M_{\beta\alpha} := \mathbb{1}_{\beta \subseteq \alpha} \, \lambda^{|\alpha|-|\beta|} \, P_{\alpha\beta}. \tag{32}$$

Since the basis $\{Z^\beta : \beta \subseteq [N]\}$ is orthonormal, we have $\mathbb{E}[f(Y)^2] \ge \mathbb{E}[g(Z)^2] = \|\hat{g}\|^2$. Note that M is invertible because it is upper triangular with nonzero diagonal entries. This means $\mathbb{E}[f(Y)\,\chi] = \langle \hat{f}, v \rangle = v^\top M^{-1} \hat{g}$. Therefore,

$$\mathsf{Corr}_{\le D} = \sup_{f \in \mathbb{R}[Y]_{\le D},\, \mathbb{E}[f(Y)^2] \ne 0} \frac{\mathbb{E}[f(Y)\,\chi]}{\sqrt{\mathbb{E}[f(Y)^2]}} \le \sup_{\hat{g} \ne 0} \frac{v^\top M^{-1} \hat{g}}{\|\hat{g}\|} = \|v^\top M^{-1}\| =: \|w\|, \tag{33}$$

where w is defined by $w^\top M = v^\top$. Moreover, we can solve for w recursively as

$$w_\alpha = \frac{1}{M_{\alpha\alpha}} \Big(v_\alpha - \sum_{\beta \subsetneq \alpha} w_\beta M_{\beta\alpha} \Big). \tag{34}$$

Let us define $\rho_\alpha := w_\alpha \lambda^{-|\alpha|}$. Then by (31), (32), and (34),

$$\rho_\varnothing = w_\varnothing = v_\varnothing = \mathbb{P}\{1 \in W\},$$

and

$$\rho_\alpha = \frac{1}{M_{\alpha\alpha}} \left(\lambda^{-|\alpha|} v_\alpha - \sum_{\beta \subsetneq \alpha} \lambda^{-(|\alpha|-|\beta|)} \rho_\beta M_{\beta\alpha} \right) = \frac{1}{P_{\alpha\alpha}} \left(\mathbb{P}\{\alpha \cup \{1\} \subseteq W\} - \sum_{\beta \subsetneq \alpha} \rho_\beta \, P_{\alpha\beta} \right).$$

The conclusion then follows from (33) together with the definition of ρ_α. ∎

We assume model $\mathcal{P}$ in the rest of this section.

Lemma 25 *For any $\alpha \subseteq \binom{[n]}{2}$, we have*

$$P_{\alpha\alpha} \ge 1 - \tau \, |\alpha| \, (2|\alpha| - 1) \ge 1 - 2\tau |\alpha|^2.$$

Proof Recall that $P_{\alpha\alpha} = \mathbb{P}\{\alpha \setminus W = \alpha\} = \mathbb{P}\{\alpha \cap W = \varnothing\}$. By (10), we have

$$\begin{aligned} \mathbb{P}\{\alpha \cap W = \varnothing\} &= \mathbb{P}\{\mathfrak{d}(z_i, z_j) > \tau/2 \text{ for all } (i,j) \in \alpha\} \\ &\ge \mathbb{P}\{\mathfrak{d}(z_i, z_j) > \tau/2 \text{ for all } i, j \in V(\alpha),\, i \ne j\} \\ &\ge \prod_{m=1}^{|V(\alpha)|-1} (1 - m\tau) \ge 1 - \tau \binom{|V(\alpha)|}{2}. \end{aligned}$$

Since $|V(\alpha)| \le 2\,|\alpha|$, the desired bound follows. ∎

Lemma 26 *For $\alpha \ne \varnothing$, suppose that $P_{\beta\beta} > 0$ for all $\beta \subseteq \alpha$. We have $\rho_\alpha = 0$ in either of the following situations:*

- $1 \notin V(\alpha)$ *or* $2 \notin V(\alpha)$*;*
- α *is a disconnected graph.*

Proof To facilitate the proof, we consider two cases that are split in a different way from the two cases in the statement of the lemma:

1. either $1 \notin V(\alpha)$, $2 \notin V(\alpha)$, or vertices 1 and 2 are in different connected components of α;
2. α is disconnected, but vertices 1 and 2 are in the same connected component of α.

Note that $|\alpha| \geq 1$ in Case 1 and $|\alpha| \geq 2$ in Case 2. For both cases, we prove $\rho_\alpha = 0$ by induction on $|\alpha|$. Each proof will establish the base case and the induction step simultaneously.

Case 1: Let G be the union of the graph α and the (potentially isolated) vertices 1 and 2. Let G_1 denote the connected component of G that contains 1, and let G_2 be the complement of G_1 in G. Let $E(G_i)$ denote the (potentially empty) edge set of G_i for $i = 1, 2$. By (10) and the fact that G_1 is not connected to G_2 , we see that

$$\begin{aligned}\mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} &= \mathbb{P}\{E(G_1) \subseteq W\} \cdot \mathbb{P}\{(1,2) \in W\} \cdot \mathbb{P}\{E(G_2) \subseteq W\} \\ &= \tau\, \mathbb{P}\{E(G_1) \subseteq W\} \cdot \mathbb{P}\{E(G_2) \subseteq W\} \\ &= \tau\, \mathbb{P}\{\alpha \subseteq W\} = \tau\, P_{\alpha\varnothing}.\end{aligned}$$

In the base case $|\alpha| = 1$, there is no nonempty $\beta \subsetneq \alpha$; in the case $|\alpha| > 1$, if $\varnothing \subsetneq \beta \subsetneq \alpha$, then $\rho_\beta = 0$ by the induction hypothesis. Combining these facts with (12) gives

$$\rho_\alpha\, P_{\alpha\alpha} = \mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} - \sum_{\beta \subsetneq \alpha} \rho_\beta\, P_{\alpha\beta} = \tau\, P_{\alpha\varnothing} - \rho_\varnothing\, P_{\alpha\varnothing} = 0.$$

Since $P_{\alpha\alpha} > 0$ by assumption, we conclude that $\rho_\alpha = 0$.

Case 2: Consider a subgraph $\beta \subsetneq \alpha$. If $1 \notin V(\alpha)$, $2 \notin V(\alpha)$, or vertices 1 and 2 are in different connected components of β, then $\rho_\beta = 0$ by Case 1 above. If β is disconnected while 1 and 2 are in the same connected component of β, then $\rho_\beta = 0$ by the induction hypothesis (and there is simply no such β in the base case $|\alpha| = 2$). Therefore, if $\rho_\beta \neq 0$, then β must be a connected graph containing both vertices 1 and 2.

Let γ be the connected component of α that contains vertices 1 and 2. We obtain from (12) that

$$\begin{aligned}\rho_\alpha\, P_{\alpha\alpha} &= \mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} - \sum_{\beta \subseteq \gamma} \rho_\beta\, P_{\alpha\beta} \\ &= \mathbb{P}\{\alpha \setminus \gamma \subseteq W\} \cdot \mathbb{P}\{\gamma \cup \{(1,2)\} \subseteq W\} - \rho_\gamma\, P_{\alpha\gamma} - \sum_{\beta \subsetneq \gamma} \rho_\beta\, P_{\alpha\beta}.\end{aligned}$$

Furthermore, using (12) again yields

$$\rho_\gamma = \frac{\mathbb{P}\{\gamma \cup \{(1,2)\} \subseteq W\}}{P_{\gamma\gamma}} - \sum_{\beta \subsetneq \gamma} \rho_\beta\, \frac{P_{\gamma\beta}}{P_{\gamma\gamma}}.$$

The above two equations together imply

$$\rho_\alpha P_{\alpha\alpha} = \mathbb{P}\{\gamma \cup \{(1,2)\} \subseteq W\} \left(\mathbb{P}\{\alpha \setminus \gamma \subseteq W\} - \frac{P_{\alpha\gamma}}{P_{\gamma\gamma}} \right) - \sum_{\beta \subsetneq \gamma} \rho_\beta \left(P_{\alpha\beta} - P_{\alpha\gamma} \frac{P_{\gamma\beta}}{P_{\gamma\gamma}} \right). \quad (35)$$

In view of the assumption $P_{\alpha\alpha} > 0$, it remains to show that the two terms in the brackets are zero.

First, by definition,

$$P_{\alpha\gamma} = \mathbb{P}\{\alpha \setminus W = \gamma\} = \mathbb{P}\{\alpha \setminus \gamma \subseteq W, \gamma \cap W = \varnothing\}.$$

Since γ is disconnected from $\alpha \setminus \gamma$ by construction, we obtain

$$P_{\alpha\gamma} = \mathbb{P}\{\alpha \setminus \gamma \subseteq W\} \cdot \mathbb{P}\{\gamma \cap W = \varnothing\} = \mathbb{P}\{\alpha \setminus \gamma \subseteq W\} \cdot P_{\gamma\gamma}.$$

Hence the first term in (35) is zero.

Second, since $\beta \subseteq \gamma \subseteq \alpha$, we have $\alpha \setminus W = \beta$ if and only if $\gamma \setminus W = \beta$ and $\alpha \setminus \gamma \subseteq W$. Thus

$$\frac{P_{\alpha\beta}}{P_{\gamma\beta}} = \frac{\mathbb{P}\{\gamma \setminus W = \beta,\, \alpha \setminus \gamma \subseteq W\}}{\mathbb{P}\{\gamma \setminus W = \beta\}} = \mathbb{P}\{\alpha \setminus \gamma \subseteq W\},$$

where the last equality holds because γ is disconnected from $\alpha \setminus \gamma$ and thus $\alpha \setminus \gamma \subseteq W$ is independent of $\gamma \setminus W = \beta$. Note that this ratio does not depend on β, so we can set $\beta = \gamma$ and obtain

$$\frac{P_{\alpha\gamma}}{P_{\gamma\gamma}} = \mathbb{P}\{\alpha \setminus \gamma \subseteq W\} = \frac{P_{\alpha\beta}}{P_{\gamma\beta}}.$$

Consequently, $P_{\alpha\beta} - P_{\alpha\gamma} \frac{P_{\gamma\beta}}{P_{\gamma\gamma}} = 0$ and so the second term in (35) is also zero. ■

Lemma 27 *Fix $\alpha \subseteq \binom{[n]}{2}$ such that $1, 2 \in V(\alpha)$ and α is a connected graph. We have*

$$\mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} \le P_{\alpha\varnothing} \le (\tau\,|V(\alpha)|)^{|V(\alpha)|-1}. \quad (36)$$

Moreover, fix $\beta \subseteq \binom{[n]}{2}$ such that $\varnothing \subsetneq \beta \subsetneq \alpha$ and $1, 2 \in V(\beta)$. We have

$$P_{\alpha\beta} \le (\tau\,|V(\alpha)|)^{|V(\alpha)|-|V(\beta)|}. \quad (37)$$

Proof By the definition of W, it holds that

$$\mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} = \mathbb{P}\{\mathfrak{d}(z_i, z_j) \le \tau/2 \text{ for all } (i,j) \in \alpha \cup \{(1,2)\}\},$$
$$P_{\alpha\varnothing} = \mathbb{P}\{\alpha \setminus W = \varnothing\} = \mathbb{P}\{\mathfrak{d}(z_i, z_j) \le \tau/2 \text{ for all } (i,j) \in \alpha\},$$

so the first inequality in (36) is obvious. Next, suppose $\mathfrak{d}(z_i, z_j) \le \tau/2$ for all $(i,j) \in \alpha$. Fix $\ell \in V(\alpha)$. Since α is a connected graph, there is a path from ℓ to any $i \in V(\alpha)$ that has length at most $|V(\alpha)|$. As a result, $\mathfrak{d}(z_i, z_\ell) \le |V(\alpha)| \cdot \tau/2$ for all $i \in V(\alpha)$. Conditional on any realization of z_ℓ, the probability that $\mathfrak{d}(z_i, z_\ell) \le |V(\alpha)| \cdot \tau/2$ for all $i \in V(\alpha)$ is at most $(\tau\,|V(\alpha)|)^{|V(\alpha)|-1}$. Hence (36) follows.

Next, we have

$$P_{\alpha\beta} = \mathbb{P}\{\alpha \setminus W = \beta\} \le \mathbb{P}\{\mathfrak{d}(z_i, z_j) \le \tau/2 \text{ for all } (i,j) \in \alpha \setminus \beta\}.$$

Suppose $\mathfrak{d}(z_i, z_j) \le \tau/2$ for all $(i,j) \in \alpha \setminus \beta$. Fix any vertex $s \in V(\alpha) \setminus V(\beta)$. We claim that there is a path from s to a vertex $t_s \in V(\beta)$ which has length at most $|V(\alpha)|$ and lies entirely in $\alpha \setminus \beta$. Given the claim, it follows that $\mathfrak{d}(z_s, z_{t_s}) \le |V(\alpha)| \cdot \tau/2$. Now, conditional on any realization of $\{z_i : i \in V(\beta)\}$, the probability that $\mathfrak{d}(z_s, z_{t_s}) \le |V(\alpha)| \cdot \tau/2$ for all $s \in V(\alpha) \setminus V(\beta)$ is at most $(\tau\,|V(\alpha)|)^{|V(\alpha)|-|V(\beta)|}$. Hence (37) follows.

It remains to prove the claim. Pick any $r \in V(\beta)$ and take a path from s to r in the graph $\alpha \cup \{(1,2)\}$. Since $1, 2 \in V(\beta)$, the first edge in the path (that is, the edge adjacent to $s \in V(\alpha) \setminus V(\beta)$) is neither $(1,2)$ nor belongs to β, and so it must belong to $\alpha \setminus \beta$. Following the path, we can find the first vertex t_s that is in $V(\beta)$. For the same reason, all edges between s and t_s must belong to $\alpha \setminus \beta$, proving the claim. ■

Lemma 28 *Fix $\alpha \subseteq \binom{[n]}{2}$ such that $1, 2 \in V(\alpha)$ and α is a connected graph. Suppose that $\tau|\alpha|^4 \le 0.1$. If α^* consists of the single edge $(1,2)$, then $|\rho_{\alpha^*}| \le \tau$. More generally,*

$$|\rho_\alpha| \le (1 + \tau|\alpha|^4)\,(|\alpha|+1)^{|\alpha|}\,(\tau\,|V(\alpha)|)^{|V(\alpha)|-1}.$$

Proof We prove the result by induction on $|\alpha|$. First, consider the base case $|\alpha| = 1$. We must have $(1,2) \in \alpha$ since $1, 2 \in V(\alpha)$. By (12) and Lemma 25, we have

$$|\rho_{\alpha^*}| \le \frac{1}{1-\tau}\left(\mathbb{P}\{\alpha^* \subseteq W\} - \rho_\varnothing\, P_{\alpha^*\varnothing}\right) = \frac{\tau - \tau^2}{1-\tau} = \tau,$$

since $\mathbb{P}\{\alpha^* \subseteq W\} = P_{\alpha^*\varnothing} = \mathbb{P}\{\mathfrak{d}(1,2) \le \tau/2\} = \tau$.

Next, fix $\alpha \subseteq \binom{[n]}{2}$ with $|\alpha| \ge 2$. Assume $|\rho_\beta| \le (1 + \tau|\beta|^4)\,(|\beta|+1)^{|\beta|}\,(\tau\,|V(\beta)|)^{|V(\beta)|-1}$ for all $\beta \subsetneq \alpha$ as the induction hypothesis. Applying (12) and Lemma 25 again, we obtain

$$|\rho_\alpha| \le \frac{1}{1 - 2\tau|\alpha|^2}\left(\mathbb{P}\{\alpha \cup \{(1,2)\} \subseteq W\} + |\rho_\varnothing|\, P_{\alpha\varnothing} + \sum_{\beta \subsetneq \alpha} |\rho_\beta|\, P_{\alpha\beta}\right),$$

where $\rho_\beta = 0$ if either 1 or 2 is not in $V(\beta)$ by Lemma 26. We then apply (36) and (37) for β such that $1, 2 \in V(\beta)$ to obtain

$$|\rho_\alpha| \le \frac{1}{1 - 2\tau|\alpha|^2}\left((1+\tau)\,(\tau\,|V(\alpha)|)^{|V(\alpha)|-1} + \sum_{\beta \subsetneq \alpha} |\rho_\beta|\,(\tau\,|V(\alpha)|)^{|V(\alpha)|-|V(\beta)|}\right).$$

Then, by the induction hypothesis $|\rho_\beta| \le (1+\tau|\beta|^4)\,(|\beta|+1)^{|\beta|}\,(\tau\,|V(\beta)|)^{|V(\beta)|-1}$ together with the assumption $\tau|\alpha|^4 \le 0.1$, we see that

$$\begin{aligned}
|\rho_\alpha| &\le 2(\tau\,|V(\alpha)|)^{|V(\alpha)|-1} + \frac{1}{1 - 2\tau|\alpha|^2} \cdot \sum_{\beta \subsetneq \alpha} (1 + \tau|\beta|^4)\,(|\beta|+1)^{|\beta|} (\tau\,|V(\alpha)|)^{|V(\alpha)|-1} \\
&= (\tau\,|V(\alpha)|)^{|V(\alpha)|-1}\left(2 + \frac{1}{1 - 2\tau|\alpha|^2} \cdot \sum_{\beta\,:\,\varnothing \subsetneq \beta \subsetneq \alpha} (1 + \tau|\beta|^4)\,(|\beta|+1)^{|\beta|}\right).
\end{aligned}$$

Finally, since $\tau|\alpha|^4 \le 0.1$ and $|\beta| \le |\alpha|-1$, we have $\frac{1}{1-2\tau|\alpha|^2}(1+\tau|\beta|^4) \le (1+\tau|\alpha|^4)$, and then

$$\begin{aligned}
2 + \frac{1}{1-2\tau|\alpha|^2} \sum_{\beta\,:\,\varnothing \subsetneq \beta \subsetneq \alpha} (1+\tau|\beta|^4)\,(|\beta|+1)^{|\beta|} &\le 2 + (1+\tau|\alpha|^4) \sum_{i=1}^{|\alpha|-1} \sum_{\beta\,:\,\beta\subsetneq\alpha,\,|\beta|=i} (i+1)^i \\
&= 2 + (1+\tau|\alpha|^4) \sum_{i=1}^{|\alpha|-1} \binom{|\alpha|}{i} (i+1)^i \\
&\le (1+\tau|\alpha|^4) \sum_{i=0}^{|\alpha|} \binom{|\alpha|}{i} |\alpha|^i \\
&= (1+\tau|\alpha|^4)\,(|\alpha|+1)^{|\alpha|}.
\end{aligned}$$

Combining the above two displays finishes the induction. ■

Proposition 29 *Recall (11) and (12). Suppose that $\tau D^4 \le 0.1$. We have:*

- *If $n\tau^2(D+1)^2 Q^{4\sqrt{D}} \le 1/2$ where $Q := \max\{\lambda(D+1)^2, 1\}$, then*
$$\mathsf{Corr}^2_{\le D} \le \tau^2(1 + \lambda^2 + 4n\tau^2(D+1)^5 Q^{4\sqrt{D}}).$$
- *If $\lambda^2(D+1)^4 M \le 1/2$ where $M := \max\{n\tau^2(D+1)^2, 1\}$, then*
$$\mathsf{Corr}^2_{\le D} \le \tau^2(1 + 4\lambda^2(D+1)^7 M).$$

Proof To ease the notation, we consider D such that $\sqrt{D}/2$ is an integer; the proof can be easily adapted to the general case by using floors $\lfloor\cdot\rfloor$ or ceilings $\lceil\cdot\rceil$.

To bound $\mathsf{Corr}^2_{\le D}$, we apply (11) and (12). By Lemma 26, it suffices to consider connected graphs $\alpha \subseteq \binom{[n]}{2}$ such that $1, 2 \in V(\alpha)$, for otherwise $\rho_\alpha = 0$. In the sequel, we focus on such α but suppress the conditions for brevity. Note that there is only one such α with $|\alpha| = 1$, i.e., the graph α^* consisting of a single edge $(1,2)$. For other graphs, we have $|\alpha| \ge 2$ and $|V(\alpha)| \ge 3$. Since the graph α is connected, we have $|V(\alpha)| \le |\alpha|+1$. Applying (11), (12), Lemma 28, and the assumption $\tau D^4 \le 0.1$, we obtain

$$\begin{aligned}
\mathsf{Corr}^2_{\le D} &\le \rho_\varnothing^2 + \lambda^2 \rho_{\alpha^*}^2 + \sum_{\alpha\,:\,2 \le |\alpha| \le D} \lambda^{2|\alpha|}\,(1+\tau|\alpha|^4)^2\,(|\alpha|+1)^{2|\alpha|}\,(\tau\,|V(\alpha)|)^{2|V(\alpha)|-2} \\
&\le \tau^2 + \lambda^2\tau^2 + 2\sum_{\ell=2}^{D} \sum_{v=3}^{D+1} \sum_{\alpha\,:\,|\alpha|=\ell,\,|V(\alpha)|=v} \\
&\qquad (\lambda(D+1))^{2\ell}\,(\tau(D+1))^{2v-2}\,\mathbb{1}\{v-1 \le \ell \le v^2/2\}.
\end{aligned}$$

Since $1, 2 \in V(\alpha)$, there are at most $\binom{n}{v-2}\left[\binom{\binom{v}{2}}{\ell} \wedge 2^{\binom{v}{2}}\right]$ graphs α with $|V(\alpha)| = v$ and $|\alpha| = \ell$. By the bounds $\binom{n}{v-2} \le n^{v-2}$, $\binom{\binom{v}{2}}{\ell} \le \binom{v}{2}^\ell \le (v-1)^{2\ell} \le (D+1)^{2\ell}$, and $2^{\binom{v}{2}} \le 2^{v^2}$, it follows that

$$\begin{aligned}
\mathsf{Corr}^2_{\le D} &\le \tau^2(1+\lambda^2) + 2\tau^2(D+1)^2 \cdot \\
&\quad \sum_{\ell=2}^{D} \sum_{v=3}^{D+1} \left[(D+1)^{2\ell} \wedge 2^{v^2}\right] (\lambda(D+1))^{2\ell}\,(n\tau^2(D+1)^2)^{v-2}\,\mathbb{1}\{v-1 \le \ell \le v^2/2\}. \quad (38)
\end{aligned}$$

Let us consider two cases.

Case 1: We bound the summation in (38) by splitting it into the following terms according to the value of v:

$$\mathsf{Corr}^2_{\le D} \le \tau^2(1+\lambda^2) + 2\tau^2(D+1)^2 \cdot \Bigg[\sum_{v=3}^{\sqrt{D}} \sum_{\ell=2}^{D} 2^{v^2} (\lambda(D+1))^{2\ell} \left(n\tau^2(D+1)^2\right)^{v-2} \mathbb{1}\{\ell \le v^2/2\} \tag{39a}$$

$$+ \sum_{\ell=2}^{D} \sum_{v=\sqrt{D}+1}^{D+1} (D+1)^{2\ell} (\lambda(D+1))^{2\ell} \left(n\tau^2(D+1)^2\right)^{v-2} \Bigg]. \tag{39b}$$

Recall that $Q := \max\{\lambda(D+1)^2, 1\}$. Moreover, by assumption,

$$Q^{\sqrt{D}}\, n\tau^2(D+1)^2 \le 1/2.$$

For $v \le \sqrt{D}$, we have $2\ell \le v^2 \le \sqrt{D}\, v$, so the sum in (39a) is bounded by

$$\begin{aligned} \sum_{v=3}^{\sqrt{D}} DQ^{\sqrt{D}\, v} \left(n\tau^2(D+1)^2\right)^{v-2} &= DQ^{2\sqrt{D}} \sum_{v=3}^{\sqrt{D}} \left(Q^{\sqrt{D}}\, n\tau^2(D+1)^2\right)^{v-2} \\ &\le 2DQ^{2\sqrt{D}} Q^{\sqrt{D}}\, n\tau^2(D+1)^2. \end{aligned}$$

Next, $n\tau^2(D+1)^2 \le 1/2$ by assumption, so the sum in (39b) is bounded by

$$\begin{aligned} \sum_{\ell=2}^{D} \sum_{v=\sqrt{D}+1}^{D+1} Q^{2\ell} \left(n\tau^2(D+1)^2\right)^{v-2} &\le 2DQ^{2D} (n\tau^2(D+1)^2)^{\sqrt{D}-1} \\ &\le 2D(Q^{4\sqrt{D}} n\tau^2(D+1)^2)^{\sqrt{D}/2} \\ &\le 2DQ^{4\sqrt{D}} n\tau^2(D+1)^2, \end{aligned}$$

where the last step holds because $Q^{4\sqrt{D}} n\tau^2(D+1)^2 \le 1/2$. Plugging the above two bounds into (39a) and (39b) respectively, we complete the proof.

Case 2: Continuing from (38), we have

$$\mathsf{Corr}^2_{\le D} \le \tau^2(1+\lambda^2) + 2\tau^2(D+1)^2 \sum_{\ell=2}^{D} \sum_{v=3}^{\ell+2} (\lambda(D+1)^2)^{2\ell} \left(n\tau^2(D+1)^2\right)^{v-2}.$$

Recall that $M = \max\{n\tau^2(D+1)^2, 1\}$ and $\lambda^2(D+1)^4 M \le 1/2$. We conclude that

$$\begin{aligned} \mathsf{Corr}^2_{\le D} &\le \tau^2(1+\lambda^2) + 2\tau^2(D+1)^2 \sum_{\ell=2}^{D} \left(\lambda^2(D+1)^4\right)^{\ell} M^{\ell} \ell \\ &\le \tau^2(1+\lambda^2) + 2\tau^2(D+1)^2 \cdot 2\lambda^2(D+1)^4 MD \\ &\le \tau^2(1+4(D+1)^7\lambda^2 M), \end{aligned}$$

finishing the proof. ■

We now prove Theorem 12.

Proof [Proof of Theorem 12] It suffices to apply Proposition 29 to bound $\mathsf{Corr}^2_{\le D}$. Consider two cases:

- If $(\log n)^{-100} \le \lambda \le O(1)$, then $n\tau^2 \le n^{-\delta/2}$ by (9). We now apply the first statement of Proposition 29. Since $D = o\Big((\frac{\log n}{\log\log n})^2\Big)$, we have $Q = \max\{\lambda(D+1)^2, 1\} = \tilde{O}(1)$ and $n\tau^2(D+1)^2Q^{4\sqrt{D}} \le n\tau^2 \cdot n^{o(1)} \le 1/2$. It follows that

$$\mathsf{Corr}^2_{\le D} \le \tau^2(1+\lambda^2+4n\tau^2(D+1)^5Q^{4\sqrt{D}}) \le \tau^2(1+\lambda^2+n\tau^2\cdot n^{o(1)}) = O(\tau^2).$$

- Next, suppose that $\lambda \le (\log n)^{-100}$, $n\tau^2\lambda^2 \le n^{-\delta}$, and $D \le (\log n)^{10}$, which hold by (9). We apply the second statement of Proposition 29 in each of the following two subcases:
  - If $\tau \le n^{-1/2}$, then $M = \max\{n\tau^2(D+1)^2, 1\} \le (D+1)^2$ and $\lambda^2(D+1)^4M \le 1/2$. Therefore, $\mathsf{Corr}^2_{\le D} \le \tau^2(1+4\lambda^2(D+1)^7M) = O(\tau^2)$.
  - If $\tau > n^{-1/2}$, then $M = n\tau^2(D+1)^2$ and $\lambda^2(D+1)^4M = n\tau^2\lambda^2(D+1)^6 \le 1/2$. We again obtain $\mathsf{Corr}^2_{\le D} \le \tau^2(1+4\lambda^2(D+1)^7M) = O(\tau^2)$.

Combining the above cases, we conclude that $\mathsf{Corr}_{\le D} = O(\tau)$ if (9) holds. This completes the proof once we recall Definition 6 and that $\mathbb{E}_{\mathcal{P}}[\chi] = \tau$. ■

A.4. Recovery upper bound

For brevity, write $T = T(A)$ in the sequel. We let $i_0 := 1$ and $i_{\ell+1} := 2$, so that a length-$(\ell+1)$ self-avoiding walk in consideration is through vertices i_j for $j = 0, 1, \dots, \ell+1$. Hence we can rewrite (16) as

$$T = \sum_{\substack{3 \le i_1, \dots, i_\ell \le n \\ i_1 \ne \cdots \ne i_\ell}} \prod_{j=0}^{\ell} \tilde{A}_{i_j i_{j+1}}. \tag{40}$$

We assume $2\tau(\ell+1) \le 1$ in the rest of this section.

Lemma 30 *Let $\tilde{A}_{ij}$ be defined by (13) and λ be defined by (8). We have*

- $\mathbb{E}\Big[\tilde{A}_{ij}|z_i, z_j\Big] = \lambda \cdot \mathbb{1}\{\mathfrak{d}(z_i, z_j) \le \tau/2\}$;
- $\mathbb{E}\Big[\tilde{A}^2_{ij}|z_i, z_j\Big] \le p/q$.

Proof The first statement is obvious in view of (13) and (8). For the second statement, note that if $\mathfrak{d}(z_i, z_j) > \tau/2$, then $\mathbb{E}\Big[\tilde{A}^2_{ij}|z_i, z_j\Big] = 1$, and if $\mathfrak{d}(z_i, z_j) \le \tau/2$, then

$$\mathbb{E}\Big[\tilde{A}^2_{ij}|z_i, z_j\Big] = \frac{p(1-q)^2 + (1-p)q^2}{q(1-q)} \le \frac{p(1-q)+q^2}{q} \le \frac{p}{q}.$$

■

Proposition 31 *If* $\mathfrak{d}(z_1, z_2) > \frac{(\ell+1)\tau}{2}$, *then* $\mathbb{E}[T \mid z_1, z_2] = 0$. *If* $\mathfrak{d}(z_1, z_2) \le \frac{(\ell+1)\tau}{2}$, *then*

$$\mathbb{E}[T \mid z_1, z_2] = \binom{n-2}{\ell} \tau^\ell \lambda^{\ell+1} \int_{\frac{\ell}{2} + \frac{\mathfrak{d}(z_1,z_2)}{\tau} - \frac{1}{2}}^{\frac{\ell}{2} + \frac{\mathfrak{d}(z_1,z_2)}{\tau} + \frac{1}{2}} f_\ell(t) \mathrm{d}t, \tag{41}$$

where $f_\ell(x)$ *is the probability density function of the Irwin-Hall distribution with parameter* ℓ, *i.e.,*

$$f_\ell(x) = \frac{1}{(\ell-1)!} \sum_{k=0}^{\lfloor x \rfloor} (-1)^k \binom{\ell}{k} (x-k)^{\ell-1} \quad \text{for } x \in [0, \ell], \tag{42}$$

and $f_\ell(x) = 0$ *otherwise. Moreover,* $u \mapsto \mathbb{E}[T \mid \mathfrak{d}(z_1, z_2) = u]$ *is a decreasing function on* $[0, \frac{(\ell+1)\tau}{2}]$.

Proof Throughout this proof, we condition on z_1 and z_2, and use $\mathbb{E}$ and $\mathbb{P}$ to denote the conditional expectation and conditional probability respectively. By (40) and the independence of $(A_{i_j i_{j+1}})_{j=0}^{\ell}$ conditional on $(z_{i_j})_{j=0}^{\ell+1}$, we have

$$\mathbb{E}[T] = \sum_{\substack{3 \le i_1, \dots, i_\ell \le n \\ i_1 \ne \cdots \ne i_\ell}} \mathbb{E}\left[\prod_{j=0}^{\ell} \mathbb{E}\left[\tilde{A}_{i_j i_{j+1}} \mid (z_{i_s})_{s=0}^{\ell+1} \right] \right].$$

Applying the first statement of Lemma 30, we then obtain

$$\mathbb{E}[T] = \lambda^{\ell+1} \sum_{\substack{3 \le i_1, \cdots, i_\ell \le n \\ i_1 \ne \cdots \ne i_\ell}} \mathbb{P}\left\{ \mathfrak{d}(z_{i_s}, z_{i_{s+1}}) \le \tau/2 \text{ for all } s \in [\ell] \right\}. \tag{43}$$

If $\mathfrak{d}(z_{i_s}, z_{i_{s+1}}) \le \tau/2$ for all $s \in [\ell]$, then $\mathfrak{d}(z_1, z_2) \le (\ell+1)\tau/2$ by the triangle inequality. Therefore, we see that $\mathbb{E}[T] = 0$ if $\mathfrak{d}(z_1, z_2) > (\ell+1)\tau/2$.

Next, suppose that $\mathfrak{d}(z_1, z_2) \le (\ell+1)\tau/2$. Fix vertices $i_1, \dots, i_\ell$ and define

$$E_s := \left\{ \mathfrak{d}(z_{i_s}, z_{i_{s+1}}) \le \frac{\tau}{2} \right\}, \qquad \mathcal{E} = \bigcap_{s=0}^{\ell} E_s,$$

where we suppress the dependency on $i_1, \dots, i_\ell$ for brevity. We now compute $\mathbb{P}\{\mathcal{E}\}$, i.e., the probability in (43). Let us write

$$\mathbb{P}\{\mathcal{E}\} = \prod_{s=0}^{\ell} \mathbb{P}\left\{ E_s \,\middle|\, \bigcap_{j=0}^{s-1} E_j \right\}. \tag{44}$$

Since $(z_{i_s})_{s=1}^{\ell}$ are i.i.d. uniform random variables in $[0, 1]$, it is not hard to see that

$$\mathbb{P}\left\{ E_s \,\middle|\, \bigcap_{j=0}^{s-1} E_j \right\} = \tau \quad \text{for } 0 \le s \le \ell - 1. \tag{45}$$

It remains to compute the conditional probability $\mathbb{P}\left\{ E_\ell \,|\, \bigcap_{j=0}^{\ell-1} E_j \right\}$. For any $0 \le s \le \ell - 1$, conditional on any realization of $z_{i_0}, z_{i_1}, \dots, z_{i_s}$ and the event E_s, the random variable $z_{i_{s+1}} - z_{i_s}$

is uniform $[-\tau/2, \tau/2]$. Crucially, this distribution does not depend on $z_{i_0}, z_{i_1}, \dots, z_{i_s}$. Applying this argument for $s = 0, 1, \dots, \ell - 1$, we see that conditional on $\bigcap_{s=0}^{\ell-1} E_s$, the random variables $z_{i_1} - z_{i_0}, z_{i_2} - z_{i_1}, \dots, z_{i_\ell} - z_{i_{\ell-1}}$ are i.i.d. and uniform in $[-\tau/2, \tau/2]$. We can write

$$z_{i_\ell} = z_{i_0} + \tau I_\ell - \frac{\ell\tau}{2}, \qquad \text{where } I_\ell := \sum_{s=0}^{\ell-1} \Big(\frac{z_{i_{s+1}} - z_{i_s}}{\tau} + \frac{1}{2} \Big).$$

Since I_r is a sum of ℓ i.i.d. uniform random variables in $[0, 1]$, it has the Irwin–Hall distribution with parameter ℓ (see, e.g., Johnson et al. (1995)). Moreover, since $i_0 = 1$ and $i_{\ell+1} = 2$, the event E_ℓ occurs if and only if $\mathfrak{d}(z_2, z_1 + \tau I_\ell - \frac{\ell\tau}{2}) \le \tau/2$, i.e.,

$$\frac{z_2 - z_1}{\tau} + \frac{\ell}{2} - \frac{1}{2} \le I_\ell \le \frac{z_2 - z_1}{\tau} + \frac{\ell}{2} + \frac{1}{2}.$$

Let $f_\ell(x)$ be the PDF of the Irwin–Hall distribution with parameter ℓ. Then

$$\mathbb{P}\Big\{ E_\ell \;\Big|\; \bigcap_{j=0}^{\ell-1} E_j \Big\} = \int_{\frac{\ell}{2} + \frac{z_2 - z_1}{\tau} - \frac{1}{2}}^{\frac{\ell}{2} + \frac{z_2 - z_1}{\tau} + \frac{1}{2}} f_\ell(t) \mathrm{d}t. \tag{46}$$

Plugging (45) and (46) into (44) and then combining the result with (43), we obtain

$$\mathbb{E}[T] = \lambda^{\ell+1} \binom{n-2}{\ell} \tau^\ell \int_{\frac{\ell}{2} + \frac{z_2 - z_1}{\tau} - \frac{1}{2}}^{\frac{\ell}{2} + \frac{z_2 - z_1}{\tau} + \frac{1}{2}} f_\ell(t) \mathrm{d}t,$$

which is almost (41). It remains to show that the above quantity is an even function in $u := z_2 - z_1$ and decreasing for $u \in [0, \frac{(\ell+1)\tau}{2}]$. Its derivative as a function of u is proportional to

$$f_\ell\Big(\frac{\ell}{2} + \frac{u}{\tau} + \frac{1}{2}\Big) - f_\ell\Big(\frac{\ell}{2} + \frac{u}{\tau} - \frac{1}{2}\Big). \tag{47}$$

The PDF $f_\ell(t)$ is symmetric around $\ell/2$, increasing on $[0, \ell/2]$, decreasing on $[\ell/2, \ell]$, and zero outside $[0, \ell]$ (and the monotonicity of $f_\ell(t)$ on $[0, \ell/2]$ and $[\ell/2, \ell]$ is strict if $\ell > 1$). Hence, the difference in (47) is an odd function in u; it is positive if $u \in [-\frac{(\ell+1)\tau}{2}, 0]$ and negative if $u \in [0, \frac{(\ell+1)\tau}{2}]$. Consequently, $\mathbb{E}[T]$ is an even function in $u = z_2 - z_1$, and it is increasing on $[-\frac{(\ell+1)\tau}{2}, 0]$ and decreasing on $[0, \frac{(\ell+1)\tau}{2}]$, proving the last statement. ∎

Lemma 32 *Let $\epsilon \in (0, \tau/2)$. There is a constant $c_\ell > 0$ depending only on ℓ such that*

$$\Delta(\epsilon) := \mathbb{E}\Big[T \,|\, \mathfrak{d}(z_1, z_2) = \frac{\tau}{2}\Big] - \mathbb{E}\Big[T \,|\, \mathfrak{d}(z_1, z_2) = \frac{\tau}{2} + \epsilon\Big] \ge c_\ell \, n^\ell \epsilon \, \tau^{\ell-1} \lambda^{\ell+1}. \tag{48}$$

Proof It follows from Proposition 31 that

$$\begin{aligned}
\Delta(\epsilon) &= \left(\int_{\frac{\ell}{2}}^{\frac{\ell}{2}+1} f_\ell(t)\mathrm{d}t - \int_{\frac{\ell}{2}+\frac{\epsilon}{\tau}}^{\frac{\ell}{2}+1+\frac{\epsilon}{\tau}} f_\ell(t)\mathrm{d}t \right) \binom{n-2}{\ell} \tau^\ell \lambda^{\ell+1} \\
&= \left(\int_{\frac{\ell}{2}}^{\frac{\ell}{2}+\frac{\epsilon}{\tau}} f_\ell(t)\mathrm{d}t - \int_{\frac{\ell}{2}+1}^{\frac{\ell}{2}+1+\frac{\epsilon}{\tau}} f_\ell(t)\mathrm{d}t \right) \binom{n-2}{\ell} \tau^\ell \lambda^{\ell+1}.
\end{aligned}$$

By the mean value theorem, there exists $\xi_1 \in \left(\frac{\ell}{2}, \frac{\ell}{2} + \frac{\epsilon}{\tau}\right)$ and $\xi_2 \in \left(\frac{\ell}{2}+1, \frac{\ell}{2}+1+\frac{\epsilon}{\tau}\right)$ such that

$$\Delta(\epsilon) = \frac{\epsilon}{\tau}(f_\ell(\xi_1) - f_\ell(\xi_2))\binom{n-2}{\ell}\tau^\ell \lambda^{\ell+1}. \tag{49}$$

If $\ell = 1$, then $f_\ell(\xi_1) - f_\ell(\xi_2) = 1$ as $\epsilon/\tau \in (0, 1/2)$; if $\ell > 1$, then $f_\ell(x)$ is strictly decreasing for $x \in [\ell/2, \ell]$, so $f_\ell(\xi_1) - f_\ell(\xi_2) \geq c'_\ell$ for a constant $c'_\ell > 0$. The conclusion follows from (49). ■

Recall that we identify an edge set $\alpha \subseteq \binom{[n]}{2}$ with the graph induced by α, and $V(\alpha) \subseteq [n]$ denotes the vertex set of α. Recall (15). For $\alpha, \beta \in \mathsf{SAW}_\ell$, we consider the graph $\alpha \triangle \beta$ and introduce the following notation which will be used in the rest of this section:

$$\mathsf{e} := |\alpha \triangle \beta|, \tag{50a}$$
$$\mathsf{v} := |V(\alpha \triangle \beta)|, \tag{50b}$$
$$\mathsf{c} := \text{number of connected components of } \alpha \triangle \beta. \tag{50c}$$

Lemma 33 *For $\alpha, \beta \in \mathsf{SAW}_\ell$, let* e, v, *and* c *be defined in* (50). *Recall* (14). *We have*

$$\mathbb{E}\left[\tilde{A}^2_{\alpha\cap\beta}\tilde{A}_{\alpha\triangle\beta} \mid z_1, z_2\right] \leq \begin{cases} (p/q)^{\ell+1} & \text{if } \alpha\triangle\beta = \varnothing, \\ (p/q)^{\ell+1-\mathsf{e}/2}\,\lambda^{\mathsf{e}}(\ell\tau)^{\mathsf{v}-\mathsf{c}-1} & \text{if } \alpha\triangle\beta \neq \varnothing. \end{cases}$$

Proof For brevity, write $z = \{z_i : i \in V(\alpha \cup \beta)\}$, and let $\mathbb{E}$ and $\mathbb{P}$ be the expectation and probability conditional on z_1, z_2 in the proof. By the independence of $(\tilde{A}_{ij})_{(i,j)\in\alpha\cap\beta}$ and $(\tilde{A}_{ij})_{(i,j)\in\alpha\triangle\beta}$ conditional on z, we have

$$\mathbb{E}\Big[\tilde{A}^2_{\alpha\cap\beta}\tilde{A}_{\alpha\triangle\beta}\Big] = \mathbb{E}\Bigg[\prod_{(i,j)\in\alpha\cap\beta}\mathbb{E}\Big[\tilde{A}^2_{ij} \mid z\Big] \cdot \prod_{(i,j)\in\alpha\triangle\beta}\mathbb{E}\Big[\tilde{A}_{ij} \mid z\Big]\Bigg].$$

It then follows from Lemma 30 that

$$\mathbb{E}\Big[\tilde{A}^2_{\alpha\cap\beta}\tilde{A}_{\alpha\triangle\beta}\Big] \leq (p/q)^{|\alpha\cap\beta|}\,\lambda^{\mathsf{e}} \cdot \mathbb{P}\{\mathfrak{d}(z_i, z_j) \leq \tau/2 \text{ for all } (i,j) \in \alpha\triangle\beta\}. \tag{51}$$

If $\alpha\triangle\beta = \varnothing$, then $\mathsf{e} = 0$ and $|\alpha \cap \beta| = \ell + 1$, so the first bound of the lemma follows. For the second bound where $\alpha\triangle\beta \neq \varnothing$, note that $|\alpha \cap \beta| = \frac{1}{2}(|\alpha|+|\beta|-|\alpha\triangle\beta|) = \frac{1}{2}(2\ell + 2 - \mathsf{e})$. Hence, it remains to bound the probability in (51) by $(\ell\tau)^{\mathsf{v}-\mathsf{c}-1}$.

Suppose that $\mathfrak{d}(z_i, z_j) \leq \tau/2$ for all $(i,j) \in \alpha\triangle\beta$. Choose vertices $j_1, \ldots, j_{\mathsf{c}} \in V(\alpha\triangle\beta)$, one from each of the c connected components of $\alpha\triangle\beta$; in particular, if $1 \in V(\alpha\triangle\beta)$, we choose $j_1 = 1$. For every $i \in V(\alpha\triangle\beta) \setminus \{1, 2\}$, there is a path of length at most ℓ from vertex i to vertex j_{s_i} for some $s_i \in [\mathsf{c}]$ such that the path lies entirely in (the sth connected component of) $\alpha\triangle\beta$. It follows that $\mathfrak{d}(z_i, z_{j_{s_i}}) \leq \ell\tau/2$. Therefore,

$$\begin{aligned} &\mathbb{P}\{\mathfrak{d}(z_i, z_j) \leq \tau/2 \text{ for all } (i,j) \in \alpha\triangle\beta\} \\ &\leq \mathbb{P}\Big\{\mathfrak{d}(z_i, z_{j_{s_i}}) \leq \ell\tau/2 \text{ for all } i \in V(\alpha\triangle\beta) \setminus (\{j_1, \ldots, j_{\mathsf{c}}\} \cup \{1,2\})\Big\} \leq (\ell\tau)^{\mathsf{v}-\mathsf{c}-1}, \end{aligned} \tag{52}$$

since the random variables $\{z_i : i \in V(\alpha\triangle\beta) \setminus (\{j_1, \ldots, j_{\mathsf{c}}\} \cup \{1,2\})\}$ are i.i.d. uniform in $[0,1]$ conditional on any realization of $z_{j_1}, \ldots, z_{j_{\mathsf{c}}}, z_1, z_2$. Plugging the above bound into (51) completes the proof. ■

We now state a slightly improved version of the above lemma when z_1 and z_2 are far apart.

Lemma 34 *In the setting of the above lemma, if* $\mathfrak{d}(z_1, z_2) > \frac{(\ell+1)\tau}{2}$ *and* $\alpha \triangle \beta \neq \varnothing$*, we have*

$$\mathbb{E}\left[\tilde{A}^2_{\alpha\cap\beta}\tilde{A}_{\alpha\triangle\beta} \mid z_1, z_2\right] \leq (p/q)^{\ell+1-\mathsf{e}/2}\,\lambda^{\mathsf{e}}(\ell\tau)^{\mathsf{v}-\mathsf{c}}.$$

Proof The only difference from the above lemma is that we now have $(\ell\tau)^{\mathsf{v}-\mathsf{c}}$ instead of $(\ell\tau)^{\mathsf{v}-\mathsf{c}-1}$. This difference originates from (52). Recall that we suppose $\mathfrak{d}(z_i, z_j) \leq \tau/2$ for all $(i,j) \in \alpha\triangle\beta$. However, since $\mathfrak{d}(z_1, z_2) > \frac{(\ell+1)\tau}{2}$, vertices 1 and 2 cannot be in the same connected components of $\alpha\triangle\beta$. Therefore, when selecting the vertices $j_1, \ldots, j_{\mathsf{c}}$, we can choose $j_1 = 1$ and $j_2 = 2$ without loss of generality. Then (52) becomes

$$\begin{aligned}&\mathbb{P}\{\mathfrak{d}(z_i, z_j) \leq \tau/2 \text{ for all } (i,j) \in \alpha\triangle\beta\}\\ &\leq \mathbb{P}\Big\{\mathfrak{d}(z_i, z_{j_{s_i}}) \leq \ell\tau/2 \text{ for all } i \in V(\alpha\triangle\beta)\setminus\{j_1, \ldots, j_{\mathsf{c}}\}\Big\} \leq (\ell\tau)^{\mathsf{v}-\mathsf{c}},\end{aligned}$$

thereby improving the bound by a factor $\ell\tau$. ∎

Lemma 35 *For* $\alpha, \beta \in \mathsf{SAW}_\ell$*, let* e*,* v*, and* c *be defined in* (50)*. We have*

$$|V(\alpha\cup\beta)| \leq \mathsf{v} - \frac{1}{2}\mathsf{e} - \mathsf{c} + \ell + 2.$$

Proof Note that the graph $\alpha\cup\beta$ is the disjoint union of $\alpha\triangle\beta$ and $\alpha\cap\beta$. To bound the number of vertices of $\alpha\cup\beta$, we start from the graph $K = \alpha\triangle\beta$, which has v vertices, and then sequentially add vertices and edges of $\alpha\cap\beta$ to K until we eventually reach $K = \alpha\cup\beta$. Hence, $|V(\alpha\cup\beta)|$ will be bounded by v plus the total number of vertices we add in this procedure.

To be more precise, at each step, we choose an edge (i,j) of $\alpha\cap\beta$ that has not yet been added to the current K, such that $i \in V(K)$. Such an edge exists because the eventual graph $\alpha\cup\beta$ is connected. Now we add (i,j) to K, and there are two cases: $j \in V(K)$ or $j \notin V(K)$. If $j \in V(K)$, then $|V(K)|$ does not increase; if $j \notin V(K)$, then $|V(K)|$ increases by 1. Moreover, the number of connected components of K may decrease by 1 if $j \in V(K)$ (when (i,j) connects two components); the number of connected components of K will not decrease if $j \notin V(K)$. Since the number of connected components of K decreases from c to 1 in the entire procedure, the first case must occur at least $\mathsf{c}-1$ times, so $|V(K)|$ does not increase in at least $\mathsf{c}-1$ steps. Since there are $|\alpha\cap\beta|$ steps of adding an edge in total, the number of vertices added is at most $|\alpha\cap\beta|-\mathsf{c}+1$. Therefore, we obtain

$$|V(\alpha\cup\beta)| \leq \mathsf{v} + |\alpha\cap\beta| - \mathsf{c} + 1.$$

To complete the proof, it suffices to recall that $|\alpha| = |\beta| = \ell+1$ so that $2|\alpha\cap\beta| + \mathsf{e} = 2\ell+2$. ∎

Lemma 36 *For* $\alpha, \beta \in \mathsf{SAW}_\ell$*, let* v *be defined in* (50)*. Suppose* $\alpha\cap\beta \neq \varnothing$*. Then we have* $\mathsf{v} \leq 2\ell$*.*

Proof Let $(i,j) \in \alpha\cap\beta$ where $i < j$. If $i = 1$, then $1 \notin V(\alpha\triangle\beta)$ and $|V(\alpha\cup\beta)| \leq 2\ell+1$. We see that $\mathsf{v} \leq |V(\alpha\cup\beta)|-1 \leq 2\ell$. The case $j = 2$ is similar. In other cases where $i \neq 1$ and $j \neq 2$, we have $\mathsf{v} \leq |V(\alpha\cup\beta)| \leq 2\ell$. ∎

Lemma 37 *For $\alpha, \beta \in \mathsf{SAW}_\ell$, let e, v, and c be defined in (50). Suppose $\alpha \triangle \beta \neq \varnothing$. Then the graph $\alpha \triangle \beta$ does not contain any dangling edge, i.e., an edge (i,j) such that vertex j is connected to only vertex i in $\alpha \triangle \beta$. As a result, we have $\mathsf{c} + \mathsf{e} - \mathsf{v} \geq 1$.*

Proof The quantity $\mathsf{c} + \mathsf{e} - \mathsf{v}$ is known as the excess of the graph $\alpha \triangle \beta$; it is always nonnegative and is zero only if $\alpha \triangle \beta$ is a forest. Since a forest obviously contains a dangling edge, it remains to prove that $\alpha \triangle \beta$ does not contain a dangling edge.

To see this, it is convenient to view $\alpha \cup \beta$ as a multigraph, which has even degree at each vertex. Further, to obtain $\alpha \triangle \beta$ from $\alpha \cup \beta$, we delete all the double edges in $\alpha \cap \beta$, so $\alpha \triangle \beta$ also has even degree at each vertex. As a result, $\alpha \triangle \beta$ does not contain a dangling edge. ■

Proposition 38 *There is a constant $C_\ell > 0$ that depends only on ℓ such that*

$$\mathrm{Var}(T \,|\, z_1, z_2) \leq C_\ell \left[n^\ell \left(\frac{p}{q} \right)^{\ell+1} + n^{2\ell-1} \tau^{2\ell-2} \lambda^{2\ell} \frac{p}{q} + n^{\ell+\frac{1}{2}} \tau \lambda^3 \left(\frac{p}{q} \right)^{\ell-\frac{1}{2}} \right].$$

Moreover, if $\mathfrak{d}(z_1, z_2) > \frac{(\ell+1)\tau}{2}$, then

$$\mathrm{Var}(T \,|\, z_1, z_2) \leq C_\ell \left[n^\ell \left(\frac{p}{q} \right)^{\ell+1} + n^{2\ell-1} \tau^{2\ell-1} \lambda^{2\ell} \frac{p}{q} + n^{\ell+\frac{1}{2}} \tau^2 \lambda^3 \left(\frac{p}{q} \right)^{\ell-\frac{1}{2}} \right].$$

Proof Throughout this proof, we condition on z_1 and z_2, and the notations $\mathbb{E}$, $\mathbb{P}$, and Var are all with respect to the conditional probability. By (16), we have

$$\mathrm{Var}(T) = \mathbb{E}[T^2] - (\mathbb{E}T)^2 = \sum_{\alpha, \beta \in \mathsf{SAW}_\ell} \mathbb{E}[\tilde{A}_\alpha \tilde{A}_\beta] - \mathbb{E}[\tilde{A}_\alpha] \cdot \mathbb{E}[\tilde{A}_\beta]. \tag{53}$$

Note that $\ell + 2 \leq |V(\alpha \cup \beta)| \leq 2\ell + 2$.

Let us first consider the extreme case $|V(\alpha \cup \beta)| = 2\ell + 2$, where the two walks α and β have disjoint edge sets and only common vertices 1 and 2. By the independence of $(\tilde{A}_{ij})_{(i,j) \in \alpha \cup \beta}$ conditional on $(z_i)_{i \in V(\alpha \cup \beta)}$ and the first statement of Lemma 30, we have

$$\begin{aligned} \mathbb{E}[\tilde{A}_\alpha \tilde{A}_\beta] &= \mathbb{E}\Big[\prod_{(i,j) \in \alpha \cup \beta} \mathbb{E}[\tilde{A}_{ij} \mid (z_i)_{i \in V(\alpha \cup \beta)}] \Big] \\ &= \lambda^{2\ell+2} \, \mathbb{P}\{\mathfrak{d}(z_i, z_j) \leq \tau/2 \text{ for all } (i,j) \in \alpha \cup \beta\}. \end{aligned}$$

Similarly,

$$\mathbb{E}[\tilde{A}_\alpha] \cdot \mathbb{E}[\tilde{A}_\beta] = \lambda^{2\ell+2} \, \mathbb{P}\{\mathfrak{d}(z_i, z_j) \leq \tau/2 \text{ for all } (i,j) \in \alpha\} \cdot \mathbb{P}\{\mathfrak{d}(z_i, z_j) \leq \tau/2 \text{ for all } (i,j) \in \beta\}.$$

Recall that we already condition on z_1 and z_2, and the variables $\{z_i : i \in V(\alpha),\, i \neq 1, 2\}$ and $\{z_i : i \in V(\beta),\, i \neq 1, 2\}$ are independent. Hence, the above two displays are equal, i.e.,

$$\mathbb{E}[\tilde{A}_\alpha \tilde{A}_\beta] - \mathbb{E}[\tilde{A}_\alpha] \cdot \mathbb{E}[\tilde{A}_\beta] = 0.$$

In all other case where $\ell+2 \le |V(\alpha\cup\beta)| \le 2\ell+1$, we have that $\alpha\cap\beta \neq \varnothing$. The first statement of Lemma 30 implies $\mathbb{E}[\tilde{A}_\alpha], \mathbb{E}[\tilde{A}_\beta] \ge 0$. Therefore, we conclude from (53) that

$$\mathrm{Var}(T) \le \sum_{v=\ell+2}^{2\ell+1} \sum_{\substack{\alpha,\beta\in \mathsf{SAW}_\ell,\\ |V(\alpha\cup\beta)|=v}} \mathbb{E}[\tilde{A}_\alpha \tilde{A}_\beta].$$

If $|V(\alpha\cup\beta)| = v$, there are $\binom{n-2}{v-2} \le n^{v-2}$ choices of the vertices in $V(\alpha\cup\beta)\setminus\{1,2\}$. With the vertices of $\alpha\cup\beta$ fixed, the number of possible graphs $\alpha\cup\beta$ is bounded by a constant $C_1 = C_1(\ell) > 0$. Moreover, we can write $\tilde{A}_\alpha\tilde{A}_\beta = \tilde{A}^2_{\alpha\cap\beta}\tilde{A}_{\alpha\triangle\beta}$. It follows that

$$\mathrm{Var}(T) \le C_1 \sum_{v=\ell+2}^{2\ell+1} n^{v-2} \max_{\substack{\alpha,\beta\in \mathsf{SAW}_\ell,\\ |V(\alpha\cup\beta)|=v}} \mathbb{E}[\tilde{A}^2_{\alpha\cap\beta}\tilde{A}_{\alpha\triangle\beta}].$$

Let e, v, and c be defined in (50). If $\alpha\triangle\beta = \varnothing$, then $|V(\alpha\cup\beta)| = \ell+2$; if $\alpha\triangle\beta \neq \varnothing$, then $|V(\alpha\cup\beta)| \ge \ell+3$. Applying Lemmas 33 and 35 together with the above bound on $\mathrm{Var}(T)$, we obtain

$$\begin{aligned}\mathrm{Var}(T) &\le C_1\left[n^\ell (p/q)^{\ell+1} + \sum_{v=\ell+3}^{2\ell+1} \max_{\substack{\alpha,\beta\in \mathsf{SAW}_\ell,\\ |V(\alpha\cup\beta)|=v}} n^{\mathsf{v}-\mathsf{e}/2-\mathsf{c}+\ell}(p/q)^{\ell+1-\mathsf{e}/2}\,\lambda^{\mathsf{e}}(\ell\tau)^{\mathsf{v}-\mathsf{c}-1}\right]\\ &\le C_1\left[n^\ell (p/q)^{\ell+1} + \ell \max_{\substack{\alpha,\beta\in \mathsf{SAW}_\ell,\\ \ell+3\le|V(\alpha\cup\beta)|\le 2\ell+1}} n^{\mathsf{v}-\mathsf{e}/2-\mathsf{c}+\ell}(p/q)^{\ell+1-\mathsf{e}/2}\,\lambda^{\mathsf{e}}(\ell\tau)^{\mathsf{v}-\mathsf{c}-1}\right].\end{aligned}$$

It follows that, for a sufficiently large constant $C_2 = C_2(\ell) > 0$,

$$\mathrm{Var}(T) \le C_2\, n^\ell \left(\frac{p}{q}\right)^{\ell+1}\left[1 + \frac{1}{\tau}\max_{\substack{\alpha,\beta\in \mathsf{SAW}_\ell,\\ \ell+3\le|V(\alpha\cup\beta)|\le 2\ell+1}} \left(\frac{n\tau^2\lambda^2 q}{p}\right)^{\frac{1}{2}(\mathsf{v}-\mathsf{c})}\left(\frac{\lambda^2 q}{np}\right)^{\frac{1}{2}(\mathsf{c}+\mathsf{e}-\mathsf{v})}\right]. \quad (54)$$

To further control the above maximum, we consider the two factors:

- Since $\lambda = \frac{p-q}{\sqrt{q(1-q)}}$, we have $\frac{\lambda^2 q}{np} = \frac{(p-q)^2}{np(1-q)} \le 1$. In addition, $\mathsf{c}+\mathsf{e}-\mathsf{v} \ge 1$ by Lemma 37. Hence, it holds that $\left(\frac{\lambda^2 q}{np}\right)^{\frac{1}{2}(\mathsf{c}+\mathsf{e}-\mathsf{v})} \le \left(\frac{\lambda^2 q}{np}\right)^{\frac{1}{2}}$.
- By Lemma 36 and $\mathsf{c} \ge 1$, we have $\mathsf{v}-\mathsf{c} \le 2\ell-1$. By Lemma 37, $\alpha\triangle\beta$ does not contain any dangling edge, so every connected component of it has at least three vertices, and thus $\mathsf{v}-\mathsf{c} \ge 2$. It follows that $\left(\frac{n\tau^2\lambda^2 q}{p}\right)^{\frac{1}{2}(\mathsf{v}-\mathsf{c})} \le \left(\frac{n\tau^2\lambda^2 q}{p}\right)^{\ell-\frac{1}{2}} \vee \left(\frac{n\tau^2\lambda^2 q}{p}\right)$.

Combining these facts with the above bound on $\mathrm{Var}(T)$, we see that

$$\begin{aligned}\mathrm{Var}(T) &\le C_4\, n^\ell \left(\frac{p}{q}\right)^{\ell+1}\left[1 + \frac{1}{\tau}\left(\frac{n\tau^2\lambda^2 q}{p}\right)^{\ell-\frac{1}{2}}\left(\frac{\lambda^2 q}{np}\right)^{\frac{1}{2}} + \frac{1}{\tau}\left(\frac{n\tau^2\lambda^2 q}{p}\right)\left(\frac{\lambda^2 q}{np}\right)^{\frac{1}{2}}\right]\\ &\le C_4\left[n^\ell\left(\frac{p}{q}\right)^{\ell+1} + n^{2\ell-1}\tau^{2\ell-2}\lambda^{2\ell}\frac{p}{q} + n^{\ell+\frac{1}{2}}\tau\lambda^3\left(\frac{p}{q}\right)^{\ell-\frac{1}{2}}\right]\end{aligned}$$

for a constant $C_4 = C_4(\ell) > 0$.

Finally, if $\mathfrak{d}(z_1, z_2) > \frac{(\ell+1)\tau}{2}$, then the application of Lemma 33 can be replaced by Lemma 34 in the above proof, so that we gain a factor τ in the case $\alpha \triangle \beta \neq \varnothing$. As a result, we do not have the factor $1/\tau$ before the max in (54). Consequently, we gain a factor τ in the second and the third term of the eventual bound. ■

We now prove Theorems 13 and 14.

Proof [Proof of Theorem 13] For brevity, write $T = T(A)$. We need to show that $\frac{\mathbb{E}[T\cdot\chi]}{\sqrt{\mathbb{E}[T^2]}} = \omega(\tau)$. Without loss of generality, we may condition on $z_1 = 0$ throughout the proof, because the distribution of A does not change if we condition on any realization of z_1. Let $\mathbb{E}$ and $\mathbb{P}$ be the expectation and the probability with respect to the conditional distribution respectively.

Using $\chi = \mathbb{1}\{\mathfrak{d}(0, z_2) \le \tau/2\}$ and Proposition 31, we obtain

$$\mathbb{E}[T\chi] = \int_{-\tau/2}^{\tau/2} \mathbb{E}[T \,|\, z_2] \,\mathrm{d}z_2 \ge \tau\, \mathbb{E}[T \,|\, z_2 = \tau/2] = \tau \binom{n-2}{\ell} \tau^\ell \lambda^{\ell+1} C_1 \ge c_2\, n^\ell \tau^{\ell+1} \lambda^{\ell+1},$$

where $C_1 = C_1(\ell) = \int_{\frac{\ell}{2}}^{\frac{\ell}{2}+1} f_\ell(t)\mathrm{d}t$ with f_ℓ defined in (42), and $c_2 = c_2(\ell) > 0$.

Next, we have

$$\mathbb{E}[T^2] = \mathbb{E}\big[\mathbb{E}[T^2 \mid z_2]\big] = \mathbb{E}[\mathrm{Var}(T \,|\, z_2)] + \mathbb{E}\big[(\mathbb{E}[T \,|\, z_2])^2\big].$$

By Proposition 31 again,

$$\begin{aligned}
\mathbb{E}\big[(\mathbb{E}[T \,|\, z_2])^2\big] &= \int_{-\frac{(\ell+1)\tau}{2}}^{\frac{(\ell+1)\tau}{2}} (\mathbb{E}[T \mid z_2])^2 \,\mathrm{d}z_2 \\
&\le (\ell+1)\tau\, (\mathbb{E}[T \mid z_2 = 0])^2 \\
&= (\ell+1)\tau \left[\binom{n-2}{\ell} \tau^\ell \lambda^{\ell+1} C_3\right]^2 \\
&\le C_4\, n^{2\ell} \tau^{2\ell+1} \lambda^{2\ell+2},
\end{aligned}$$

where $C_3 = C_3(\ell) = \int_{\frac{\ell}{2}-\frac{1}{2}}^{\frac{\ell}{2}+\frac{1}{2}} f_\ell(t)\mathrm{d}t$, and $C_4 = C_4(\ell) > 0$. Moreover, recall that p and q are of the same order by assumption, so for any realization of z_2, Proposition 38 gives

$$\mathrm{Var}(T \mid z_2) \le C_5 \left(n^\ell + n^{2\ell-1}\tau^{2\ell-2}\lambda^{2\ell} + n^{\ell+1/2}\tau\lambda^3\right)$$

for a constant $C_5 = C_5(\ell) > 0$. Therefore, for a constant $C_6 = C_6(\ell) > 0$,

$$\sqrt{\mathbb{E}[T^2]} \le C_6 \left(n^\ell \tau^{\ell+1/2}\lambda^{\ell+1} + n^{\ell/2} + n^{\ell-1/2}\tau^{\ell-1}\lambda^\ell + n^{\ell/2+1/4}\tau^{1/2}\lambda^{3/2}\right).$$

Combining the above bounds on $\mathbb{E}[T\chi]$ and $\sqrt{\mathbb{E}[T^2]}$, we conclude that

$$\begin{aligned}
\frac{\mathbb{E}[T\cdot\chi]}{\sqrt{\mathbb{E}[T^2]}} &\ge \frac{c_2}{C_6} \cdot \frac{n^\ell \tau^{\ell+1}\lambda^{\ell+1}}{n^\ell\tau^{\ell+1/2}\lambda^{\ell+1} + n^{\ell/2} + n^{\ell-1/2}\tau^{\ell-1}\lambda^\ell + n^{\ell/2+1/4}\tau^{1/2}\lambda^{3/2}} \\
&\ge c_7 \min\left\{\tau^{1/2}, n^{\ell/2}\tau^{\ell+1}\lambda^{\ell+1}, n^{1/2}\tau^2\lambda, n^{\ell/2-1/4}\tau^{\ell+1/2}\lambda^{\ell-1/2}\right\}.
\end{aligned}$$

For this bound to be of order $\omega(\tau)$, it suffices to have

$$\tau = o(1), \qquad n\tau^2\lambda^{2+2/\ell} = \omega(1), \qquad n\tau^2\lambda^2 = \omega(1).$$

Since $Cq \le p \le C'q$ for $C' > C > 1$, we have $\lambda = \frac{p-q}{\sqrt{q(1-q)}} = \Theta(p^{1/2})$. Therefore, the above conditions all hold by the assumptions $\ell > 1/\delta$ and (17). ■

Proof [Proof of Theorem 14] Similar to the proof of Theorem 13, we write $T = T(A)$ and condition on $z_1 = 0$ throughout the proof. We start by rewriting the expectation as the sum of type I and type II errors:

$$\mathbb{E}\big[(\hat{\chi} - \chi)^2\big] = \mathbb{P}\{\hat{\chi} \neq \chi\} = \mathbb{P}\{\chi = 1, \hat{\chi} = 0\} + \mathbb{P}\{\chi = 0, \hat{\chi} = 1\}.$$

Since $\chi = \mathbb{1}\{-\tau/2 \le z_2 \le \tau/2\}$ and $\hat{\chi} = \mathbb{1}\{T < \kappa\}$, we have

$$\begin{aligned}
\mathbb{P}\{\chi = 1, \hat{\chi} = 0\} &= \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} \mathbb{P}\{T < \kappa \,|\, z_2\} \,\mathrm{d}z_2, \\
\mathbb{P}\{\chi = 0, \hat{\chi} = 1\} &= \int_{\frac{\tau}{2} \le |z_2| \le 1} \mathbb{P}\{T \ge \kappa \,|\, z_2\} \,\mathrm{d}z_2 \\
&\le 2\epsilon + \int_{\frac{\tau}{2} + \epsilon \le |z_2| \le \frac{(\ell+1)\tau}{2}} \mathbb{P}\{T \ge \kappa \,|\, z_2\} \,\mathrm{d}z_2 + \int_{\frac{(\ell+1)\tau}{2} < |z_2| \le 1} \mathbb{P}\{T \ge \kappa \,|\, z_2\} \,\mathrm{d}z_2.
\end{aligned}$$

It remains to bound the above three integrals:

- Consider $z_2 \in [-\tau/2, \tau/2]$. Let $\Delta(\epsilon)$ be defined in (48) and κ be defined in (18). By Proposition 31 and Chebyshev's inequality,

$$\mathbb{P}\{T < \kappa \,|\, z_2\} \le \mathbb{P}\left\{|T - \mathbb{E}[T \,|\, z_2]| > \frac{\Delta(\epsilon)}{2} \;\middle|\; z_2\right\} \le \frac{4\,\mathrm{Var}(T \mid z_2)}{\Delta(\epsilon)^2}.$$

  Lemma 32 and Proposition 38 together imply that

$$\begin{aligned}
\frac{\mathrm{Var}(T \mid z_2)}{\Delta(\epsilon)^2} &\le C_1 \frac{n^\ell + n^{2\ell-1}\tau^{2\ell-2}\lambda^{2\ell} + n^{\ell+\frac{1}{2}}\tau\lambda^3}{n^{2\ell}\epsilon^2\tau^{2\ell-2}\lambda^{2\ell+2}} \\
&\le C_1 \left(\frac{1}{n^\ell \epsilon^2 \tau^{2\ell-2}\lambda^{2\ell+2}} + \frac{1}{n\epsilon^2\lambda^2} + \frac{1}{n^{\ell-1/2}\epsilon^2\tau^{2\ell-3}\lambda^{2\ell-1}} \right)
\end{aligned}$$

  for a constant $C_1 = C_1(\ell) > 0$. Using the assumptions $n\tau^2\lambda^2 = \Theta(n\tau^2 p) \ge n^\delta$, $\ell > 3/\delta$, and $\epsilon = \tau n^{-\delta/4}$, we can check

$$\frac{\mathrm{Var}(T \mid z_2)}{\Delta(\epsilon)^2} \le 3C_1\, n^{-\delta/2}. \tag{55}$$

  Therefore,

$$\int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} \mathbb{P}\{T < \kappa \,|\, z_2\} \,\mathrm{d}z_2 \le 12C_1\tau n^{-\delta/2}.$$

- For $\frac{\tau}{2}+\epsilon \le |z_2| \le \frac{(\ell+1)\tau}{2}$, again, by Proposition 31, Chebyshev's inequality, and (55),

$$\mathbb{P}\{T \ge \kappa \,|\, z_2\} \le \mathbb{P}\left\{|T - \mathbb{E}[T \,|\, z_2]| > \frac{\Delta(\epsilon)}{2} \mid z_2\right\} \le \frac{4\operatorname{Var}(T \mid z_2)}{\Delta(\epsilon)^2} \le 12C_1 n^{-\delta/2}.$$

  Therefore,

$$\int_{\frac{\tau}{2}+\epsilon \le |z_2| \le \frac{(\ell+1)\tau}{2}} \mathbb{P}\{T \ge \kappa \,|\, z_2\}\, \mathrm{d}z_2 \le 12(\ell+1)C_1 \tau n^{-\delta/2}.$$

- For $\frac{(\ell+1)\tau}{2} < |z_2| \le 1$, we have $\mathbb{E}[T \mid z_2] = 0$ by Proposition 31. Combining Chebyshev's inequality, the second bound in Proposition 38, (18), and Proposition 31, we obtain

$$\begin{aligned}\mathbb{P}\{T \ge \kappa \,|\, z_2\} \le \frac{\operatorname{Var}(T \,|\, z_2)}{\kappa^2} &\le C_2 \frac{n^\ell + n^{2\ell-1}\tau^{2\ell-1}\lambda^{2\ell} + n^{\ell+\frac{1}{2}}\tau^2\lambda^3}{n^{2\ell}\tau^{2\ell}\lambda^{2\ell+2}} \\ &\le C_2\left(\frac{1}{n^\ell \tau^{2\ell}\lambda^{2\ell+2}} + \frac{1}{n\tau\lambda^2} + \frac{1}{n^{\ell-1/2}\tau^{2\ell}\lambda^{2\ell-1}}\right)\end{aligned}$$

  for $C_2 = C_2(\ell) > 0$. Using the assumptions $n\tau^2\lambda^2 = \Theta(n\tau^2 p) \ge n^\delta$ and $\ell > 3/\delta$, we can check

$$\mathbb{P}\{T \ge \kappa \,|\, z_2\} \le 3C_2\, \tau n^{-\delta/2}.$$

  Therefore,

$$\int_{\frac{(\ell+1)\tau}{2} < |z_2| \le 1} \mathbb{P}\{T \ge \kappa \,|\, z_2\}\, \mathrm{d}z_2 \le 3C_2\, \tau n^{-\delta/2}.$$

In summary, we have obtained $\mathbb{E}\big[(\hat{\chi}-\chi)^2\big] = \mathbb{P}\{\hat{\chi} \ne \chi\} \le C_3 \tau n^{-\delta/2}$ for $C_3 = C_3(\ell) > 0$. ■