Proceedings of Machine Learning Research vol 75:1–34, 2018 31st Annual Conference on Learning Theory

Optimal Single Sample Tests for Structured versus Unstructured Network Data

Guy Bresler GUY@MIT.EDU
Dheeraj Nagaraj DHEERAJ@MIT.EDU
Department of Electrical Engineering and Computer Science, MIT

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the problem of testing, using only a single sample, between mean field distributions (like Curie-Weiss, Erdős-Rényi) and structured Gibbs distributions (like Ising model on sparse graphs and Exponential Random Graphs). Our goal is to test without knowing the parameter values of the underlying models: only the *structure* of dependencies is known. We develop a new approach that applies to both the Ising and Exponential Random Graph settings based on a general and natural statistical test. The test can distinguish the hypotheses with high probability above a certain threshold in the (inverse) temperature parameter, and is optimal in that below the threshold no test can distinguish the hypotheses.

The thresholds do not correspond to the presence of long-range order in the models. By aggregating information at a global scale, our test works even at very high temperatures. The proofs are based on distributional approximation and sharp concentration of quadratic forms, when restricted to Hamming spheres. The restriction to Hamming spheres is necessary, since otherwise any scalar statistic is useless without explicit knowledge of the temperature parameter. At the same time, this restriction changes the behavior of the functions under consideration, making it hard to directly apply standard methods (i.e., Stein's method) for concentration of weakly dependent variables. Instead, we carry out an additional tensorization argument using a Markov chain that respects the symmetry of the Hamming sphere.

1. Introduction

Hypothesis testing for network data has received a lot of attention in recent years. There are two basic types of network data: first, the network or graph itself; and second, observations from the nodes in a network, where the network describes interactions between the nodes. A recent example of the first type is studied in the paper of Bubeck et al. (2016), which gives an optimal single-sample test to distinguish between geometric random graphs and Erdős-Rényi random graphs by counting the number triangles in the graph. Similarly, Gao and Lafferty (2017) use distributional approximation for a specific statistic to distinguish between an Erdős-Rényi random graph and sample from the Stochastic Block model. Another paper in this direction is that of Ghoshdastidar et al. (2017), who consider the problem of deciding whether two given graphs are samples from the same graph model or from two different models. Their method is based on existence of a statistic that concentrates at different values for the two different graph models. The problem of testing if

a known graph (with atleast $\Omega(\log n)$ vertices) is planted in an Erdős-Rényi random graph with known edge parameter was studied by Javadi and Montanari (2015). They give sharp single sample thresholds for the problem and the corresponding statistical test which can achieve this threshold. As will be seen below, our result on testing graph models differs in the fact that we consider the appearance of much smaller subgraphs and the subgraphs are not 'planted' explicitly.

As far as data from nodes in a network, Martín del Campo et al. (2017) considers the problem of tractably finding goodness-of-fit for Ising models. Daskalakis et al. (2016) developed methods for testing whether samples are coming from a given known Ising model. They assume full knowledge of all the model parameters, and use a test based on the empirical correlations between nodes, requiring polynomially many independent samples. In contrast, we focus on testing using a single sample and use the special structure present in the mean-field case to give a sharp threshold above which single sample testing is possible, using a general framework applicable to other models. Daskalakis et al. (2017) and Gheissari et al. (2017) show concentration for polynomials of Ising models at high temperature, and improve the sample complexities obtained in Daskalakis et al. (2016) for testing whether samples are from the product distribution (i.e., coordinates are independent) or from an Ising model π guaranteed to have KL-divergence at least ϵ from the product distribution. Analogously, Canonne et al. (2017) consider the problem of determining whether observed samples from a distribution P agree with a known fully-specified Bayesian network Q, using multiple samples; and also the problem of testing whether two unknown Bayes nets are identical or not, using multiple samples. Finally, this latter paper considers also *structure testing*, i.e., testing if samples are from a Bayes net with a certain structure. Our objectives differ from these papers in that: 1) our test is based on a single sample; and 2) there are no assumptions of separation in KL-divergence or total variation on the distributions generating the sample. Instead, the guarantees are in terms of the natural model parameter. Mukherjee (2013) considers the problem of consistent parameter estimation of the two star (wedge graph) ERGM considered in this paper. Their method assumes that the strength of the 'wedge interaction' $\beta_2 \in (0, \infty)$ is fixed. Whereas, in our work, the sharp threshold for distinguishing this graph from Erdős-Rényi graphs is shown to be $\beta_2 = \Theta(\frac{1}{\sqrt{n}})$, which goes to 0 with n. It is unclear how their parameter estimation methods can be used in this case to obtain the sharp threshold behavior.

1.1. Results

In this paper we prove an abstract result, Theorem 9, that provides a framework for establishing near-optimal hypothesis tests between data from a network with a given dependency structure (like Ising model, Exponential Random Graph Model) and unstructured data (like Curie-Weiss, Erdős-Rényi). We do not assume knowledge of model parameters, which makes the problem more challenging, but also more applicable to many settings where there is no way to learn them accurately based on one sample.

As the first of two applications developed in this paper, we consider the problem of testing whether the network data comes from an Ising model over a known d-regular graph with unknown inverse temperature β (also with possibly nonzero external field) or alternatively from a permutation invariant distribution (which includes the Curie-Weiss

model at unknown temperature). We motivate this problem by discussing an adversarial data scenario in Section 1.2.

Theorem 1 (Informal version of Theorem 11) *We can distinguish Ising models on d-regular graphs from the Curie-Weiss model (complete graph) with high probability with one sample if the inverse temperature β of the Ising model satisfies $\beta\sqrt{nd} \to \infty$. Conversely, if $\beta\sqrt{nd} \to 0$, then there is no statistical test that can distinguish them with high probability, even using a constant number of i.i.d. samples.*

Remark 2 *We interpret the result above as follows: whenever $\beta\sqrt{nd} \to \infty$, an adversary cannot come up with a Curie-Weiss sample at some temperature such that it can be confused for a sample from the d regular Ising model. Conversely, whenever $\beta\sqrt{nd} \to 0$, the adversary can choose a Curie-Weiss model at a specific temperature depending only on β such that the total variation distance between these distributions converges to 0. The problem is formulated in the minimax sense.*

The result works for every d-regular graph. Basak and Mukherjee (2017) showed that certain properties of Ising models are well-approximated by the Curie-Weiss (mean-field) model, including the limit of the log-partition function. It was shown in Bresler and Nagaraj (2017) that pairwise correlations, and more generally kth-order moments, of the Curie-Weiss model can be well approximated on average by expander graphs, yet the result above holds even when the underlying graph is an expander. The test also works deep inside the high temperature regime ($\beta \leq \Theta(\frac{1}{d})$), when there is no global order, by aggregating small dependencies from the entire network.

Our results also apply to certain random graph distributions, and in Section 7 we apply our framework to compare $G(n, p_n)$ (the Erdős-Rényi model) and exponential random graphs. Let $\mathrm{ERGM}(\beta_1, \beta_2)$ be the exponential random graph with respect to the single edge E and the V-graph ($\bigvee$) with inverse temperature parameters $\beta = (\beta_1, \beta_2) \in \mathbb{R}^2$. The parameter β_1 controls edge density, while β_2 encourages presence of V-subgraphs.

Theorem 3 (Informal version of Theorem 15) *We can distinguish $G(n, p)$ and $\mathrm{ERGM}(\beta)$ with high probability with one sample if $\sqrt{n}\beta_2 \to \infty$. Conversely, if $\sqrt{n}\beta_2 \to 0$, then there is no test which can distinguish them with high probability using a constant number of i.i.d. samples.*

Remark 4 *Specifically, we can distinguish between these models even with the same edge density, as long as $\beta_2\sqrt{n} \to \infty$. Whenever $\beta_2\sqrt{n} \to 0$, we can choose p and β_1 such that the total variation distance between these distributions converges to 0.*

In Bhamidi et al. (2011) it is shown that in the high-temperature regime $\beta_2 \leq \Theta(1)$, any finite collection of k edges converges in distribution to independence. (In $G(n, p)$ all edges are independent.) Our test aggregates global information to distinguish between them and works when the dependence parameter β_2 is much smaller than the high-temperature threshold. Bhamidi et al. (2011) and Eldan and Gross (2017) consider existence of unique solutions to a certain fixed point equation to define the high temperature regime in ERGMs. We use an entirely different method to identify the phases in our setup—where we choose parameters of degree 2 polynomials of binomial random variables to minimize the variance—to choose β_1 as a function of β_2 and p such that $\mathrm{ERGM}(\beta_1, \beta_2)$ converges in total variation distance to $G(n, p)$ whenever $\beta_2\sqrt{n} \to 0$. This is illustrated in Appendix F.

Outline. The next subsection motivates our results with an adversarial data detection scenario. Section 2 introduces notation and defines the Ising and exponential random graph models, formulates the statistical problem, and gives intuition for the test we use in our applications. In Section 3 we state our abstract hypothesis testing result, which is based on distributional approximation. In Section 4 we apply our framework to prove Theorem 11 for the Ising model. In Section 5 we prove the required distributional approximation for quadratic forms using Stein's method and in Section 6 we prove sharp concentration inequalities for quadratic forms over the Hamming sphere using a novel method.

1.2. Motivating example: detecting fraudulent data

Suppose that we have collected responses to a survey from a set of people, indicating a binary preference for something (iPhone or Android, Democrat or Republican, etc.). Moreover, we have access to the network structure G (e.g., induced Facebook subgraph) and the data is modeled by a family of probability distributions $\{Q_{G,\lambda} : \lambda \in \Lambda\}$ (e.g., Ising models on G) for some parameter set Λ. An adversary may attempt to counterfeit the data generated by the network using instead a distribution P, possibly biased (e.g., to fix an election). We assume that the adversary may know the graph, but does not know the labeling of the nodes. The adversary seeks to minimize the probability of the tampering being detected, which amounts to minimizing $\mathbb{E}_\pi \inf_{\lambda\in\Lambda} d_{\mathsf{TV}}(P, Q_{\pi G,\lambda})$, where π is a uniformly random permutation encoding the adversary's prior over the node labels.

The analysis of the quantity $\mathbb{E}_\pi \inf_{\lambda\in\Lambda} d_{\mathsf{TV}}(P, Q_{\pi G,\lambda})$ is fairly involved and requires convexity of the class of distributions. Our framework is able to handle testing against a convex combination of distributions, but for this manuscript we instead relax this objective to $\inf_{\lambda\in\Lambda} \mathbb{E}_\pi d_{\mathsf{TV}}(P, Q_{G,\lambda})$.

For any permutation π, let the distribution πP be defined by $\pi P(x) = P(\pi(x))$. For arbitrary $\lambda \in \Lambda$,

$$\begin{aligned}\mathbb{E}_\pi d_{\mathsf{TV}}(P, Q_{\pi G,\lambda}) = \mathbb{E}_\pi d_{\mathsf{TV}}(\pi^{-1}P, Q_{G,\lambda}) &= \frac{1}{n!}\sum_\pi d_{\mathsf{TV}}(\pi^{-1}P, Q_{G,\lambda}) \\ &\geq d_{\mathsf{TV}}\left(\sum_\pi \tfrac{1}{n!}\pi^{-1}P, Q_{G,\lambda}\right). \qquad (1)\end{aligned}$$

In the first step, we have used the fact that $\pi Q_{G,\lambda} \stackrel{d}{=} Q_{\pi G,\lambda}$ (due to relabeling of vertices). In the third step we have used Jensen's inequality for the convex function d_{TV}. Clearly, the distribution $\hat{P} := \frac{1}{n!}\sum_\pi \pi^{-1}P$ is permutation invariant.

If there is a unique optimal distribution P_0 for the adversary, we conclude that it must be a permutation invariant distribution. Some of the key features of the problem above are: There is only one sample available, the underlying network structure and model is known and the adversary, who is agnostic to the network structure, comes up with permutation invariant data to mimic the data from the network. The considerations above justify the setup in Section 4, where the true network data is taken to be from an Ising model.

2. Notation and Definitions

$\mathbb{E}_p f$ denotes the expectation with respect to the probability measure p. For any two probability measures μ and ν over $\mathbb{R}$, we denote the Kolmogorov-Smirnoff distance as $d_{\mathsf{KS}}(\mu,\nu) := \sup_{x_0\in\mathbb{R}} |\mu(\{x : x \le x_0\}) - \nu(\{x : x \le x_0\})|$. Let $\mathsf{Lip}_1(\mathbb{R})$ be the class of all 1-Lipschitz real-valued functions over $\mathbb{R}$. For μ and ν probability measures over $\mathbb{R}$, the Wasserstein distance is defined as: $d_{\mathsf{W}}(\mu,\nu) = \sup_{f\in\mathsf{Lip}_1(\mathbb{R})} \mathbb{E}_\mu f - \mathbb{E}_\nu f$. For any random variable X, let $\mathcal{L}(X)$ be the probability law of X. Let $\Phi(x)$ denote the standard normal cumulative distribution function.

2.1. Ising Model

The *interaction matrix* J is a real-valued symmetric $n \times n$ matrix with zeros on the diagonal and the *external field* is a real number h. Define the Hamiltonian $\mathcal{H}_{J,h} : \{-1,1\}^n \to \mathbb{R}$ by $\mathcal{H}_{J,h}(x) = \frac{1}{2}x^\intercal Jx + h\left(\sum_i x_i\right)$. Construct the graph $G_J = ([n], E_J)$ with $(i,j) \in E_J$ iff $J_{ij} \neq 0$. An Ising model over graph G_J with interaction matrix J and external field h is the probability measure π over $\{-1,1\}^n$ such that $\pi(x) \propto \exp\left(H_J(x)\right)$.

For any simple graph $G = ([n], E)$ there is an associated symmetric $n \times n$ adjacency matrix $\mathcal{A}(G) := (\mathcal{A}_{ij})$, where $\mathcal{A}_{ij} = 1$ if $(i,j) \in E$ and $\mathcal{A}_{ij} = 0$ otherwise.

Let K_n be the complete graph on n nodes. The *Curie-Weiss model* at inverse temperature $\beta^{\mathrm{CW}} > 0$ and external field h^{CW} is the Ising model with interaction matrix $\frac{\beta^{\mathrm{CW}}}{n}\mathcal{A}(K_n)$, which corresponds to the distribution $p(x) \propto e^{\frac{\beta^{\mathrm{CW}}}{2}nm^2 + nh^{\mathrm{CW}}m}$. Here $m = m(x) = \frac{1}{n}\sum_{i=1}^n x_i$ is called the magnetization. The Curie-Weiss model is an unstructured/mean field model. It is permutation invariant and assigns the same probability to states with the same magnetization $m(x)$.

We will compare the above model to the Ising model on a d-regular graph $G_d = ([n], E_d)$. For a given inverse temperature β^{dreg}, we consider the Ising model with interaction matrix $\beta^{\mathrm{dreg}}\mathcal{A}(G_d)$ and external field h^{dreg}. We shall call this 'd-regular Ising model'.

Remark 5 *The Curie-Weiss model exhibits non-trivial behavior when $\beta^{\mathrm{CW}} = \Theta(1)$. It undergoes a phase transition when $\beta^{\mathrm{CW}} = 1$, below which pairwise correlations are $O(\frac{1}{n})$ and when $\beta^{\mathrm{CW}} > 1$ they are $\Theta(1)$. The pairwise correlations tend to 1 as $\beta^{\mathrm{CW}} \to \infty$. As considered in Section 4, $\beta^{\mathrm{CW}} \le \beta_{\mathsf{max}}$ for fixed β_{max} is a natural choice of Curie-Weiss models. The non-trivial regime for the d-regular Ising model is $\beta_n^{\mathrm{dreg}} = \Theta(\frac{1}{d})$. As shown in Section 4, our test works as long as $\beta_n^{\mathrm{dreg}} \gg \frac{1}{\sqrt{nd}}$, which includes the regime of interest.*

2.2. Exponential Random Graph Model

The Erdős-Rényi random graph model $G(n,p)$ for $p \in [0,1]$ is the distribution of simple graphs on n vertices such that each edge is included independently with probability p.

Consider fixed finite simple graphs $\{H_i\}_{i=1}^K$, such that H_1 is the graph with two vertices and a single edge. Let $\beta \in \mathbb{R} \times (\mathbb{R}^+)^{K-1}$. Given a graph G over n vertices, define $N_i(G)$ to be the number of edge preserving isomorphisms from H_i into G (i.e, no. of subgraphs of G (not necessarily induced), which are isomorphic to H_i). In particular $N_1(G)$ is twice the number of edges in G. Let v_i be the number of vertices in H_i. In the following definition

of Exponential Random Graph Model (ERGM), to allow non-trivial behavior we follow the convention in Bhamidi et al. (2011) that $N_i(G)$ are of the same order of magnitude.

We construct the Hamiltonian $\mathcal{H}_\beta(G) = \sum_{i=1}^{K} \beta_i \frac{N_i(G)}{n^{v_i-2}}$. The exponential random graph model $\mathrm{ERGM}(\beta)$ is the probability distribution $\nu(\cdot)$ over the set of simple graphs on n vertices such that $\nu(G) = \frac{e^{\mathcal{H}_\beta(G)}}{Z(\beta)}$, where $Z(\beta)$ is the normalizing factor. Note that when $\beta_i = 0$ for $i \geq 2$, $\mathrm{ERGM}(\beta)$ is the same as $G(n, \frac{e^{2\beta_1}}{1+e^{2\beta_1}})$. Roughly speaking, ERGM is like $G(n,p)$, but it favors the occurrence of certain subgraphs. Therefore, $G(n,p)$ is the mean field model and ERGM is the structured model. In this paper, we take $K = 2$ and fix H_2 to be the wedge graph ($\wedge$) and denote the resulting model by $\mathrm{ERGM}(\beta_1, \beta_2)$.

2.3. Problem Formulation

We formulate our problem as a minimax hypothesis testing problem:

$$H_0 : \text{Data is from some mean field model } P_\gamma \text{ with unknown } \gamma \in \Gamma$$
$$H_1 : \text{Data is from a structured model } Q_\lambda \text{ with unknown parameters } \lambda \in \Lambda.$$

A statistical test $\mathcal{T}$ is a decision function $\mathcal{D}_\mathcal{T} : \Omega \to \{H_0, H_1\}$. Let $p_1(\gamma, \mathcal{T}) := \mathbb{P}(\mathcal{D}_\mathcal{T}(\hat{X}) = H_1 | \hat{X} \sim P_\gamma)$ and $p_2(\lambda, \mathcal{T}) := \mathbb{P}(\mathcal{D}_\mathcal{T}(\hat{X}) = H_0 | \hat{X} \sim Q_\lambda)$. We take the risk of the test $\mathcal{T}$ to be worst case Bayesian probability of error:

$$R(\mathcal{T}) = \sup_{\gamma \in \Gamma} \sup_{\lambda \in \Lambda} \max(p_1(\gamma, \mathcal{T}), p_2(\lambda, \mathcal{T})) \,.$$

We describe the philosophy behind theorems 11 and 15 below. We keep Γ fixed and consider two different regimes for Λ:

1. Case 1: Λ is such that the interaction parameter β is large enough for every Q_λ.

   We explicitly construct a test $\mathcal{T}$ to for specific P_γ and Q_λ such that $p_1(\gamma, \mathcal{T}) \to 0$ and $p_2(\lambda, \mathcal{T}) \to 0$. We then extend this to the composite case by proving that the test does not actually require the knowledge of the parameters γ and λ and hence $R(\mathcal{T}) \to 0$.

2. Case 2: Λ is such that the interaction parameter β is small for every Q_λ.

   We show that in this case, for every sequence of tests $\{\mathcal{T}_n\}$, $\liminf_{n\to\infty} R(\mathcal{T}_n) \geq \frac{1}{2}$, which is the same as random labeling w.p $\frac{1}{2}$. To prove this, we show that for every $\lambda \in \Lambda$ in this set, we can find $\gamma \in \Gamma$ such that $d_{\mathsf{TV}}(P_\gamma, Q_\lambda) \to 0$. Since $\max(p_1(\gamma, \mathcal{T}), p_2(\lambda, \mathcal{T})) \geq \frac{1}{2}(1 - d_{\mathsf{TV}}(P_\gamma, Q_\lambda))$, we conclude the result.

2.4. Intuition behind the Comparison Result

Let $q(\cdot)$ be the Ising model with interaction matrix $\beta^{\mathrm{dreg}} B$ (β^{dreg} unknown) and the Curie-Weiss model $p(\cdot)$ (at an unknown temperature β^{CW}). The measure $q(\cdot)$ assigns higher probability to states with higher value of $x^\intercal B x$, so a natural idea for distinguishing between p and q would be to check if $x^\intercal B x$ has a large value. However, the inverse temperature parameters are unknown, which implies that we can have the same expected value for the statistic under both hypotheses (for some choice of temperature parameters).

Instead, we exploit the symmetry in the Curie-Weiss model. Let $\Omega_n = \{-1, 1\}^n$ and recall the magnetization $m(x) = \frac{\sum_{i=1}^n x_i}{n}$. Let $A_{m_0} = \{x \in \Omega_n : m(x) = m_0\}$ for $m_0 \in \{-1, -1 + \frac{2}{n}, \ldots, 1\} =: M_n$. We can partition Ω_n as $\Omega_n = \cup_{m_0 \in M_n} A_{m_0}$.

By symmetry, the Curie-Weiss model $p(\cdot)$ gives the uniform distribution over each set A_{m_0} irrespective of the inverse temperature and external field, which mitigates our initial problem. The distribution $q(\cdot)$, given the magnetization m_0, assigns most of the probability to states x with large values of $x^\intercal Bx$. We first prove a central limit theorem for $g(x) := x^\intercal Bx$ when x is drawn uniformly from the set A_{m_0}. Then we show that the event

$$\frac{g(x) - \mathbb{E}_p[g(x)|x \in A_{m_0}]}{\sqrt{\mathrm{var}_p(g(x)|x \in A_{m_0})}} \geq T$$

has a small probability under $p(\cdot|x \in A_{m_0})$ for large values of T, but has a large probability under $q(\cdot)$ because it favors larger values of $g(x)$. This gives us a distinguishing statistic for large enough inverse temperature β^{dreg}.

Similarly, $\mathsf{ERGM}(\beta_1, \beta_2)$ favors the appearance of 'V' subgraphs ( ) when compared to $G(n, p)$, but the expected number of 'V' subgraphs can be made equal by increasing p. We overcome this issue by exploiting the symmetry in $G(n, p)$: it assigns the same probability to all graphs with the same number of edges. So conditioned on the number of edges in the sample graph, we check if the number of 'V' subgraphs are disproportionately large.

3. Abstract Result

We consider a sequence of probability spaces $(\Omega_n, \mathcal{F}_n, p_n)$, $n \in \mathbb{N}$. Consider a $\mathcal{F}_n$ measurable, real valued function g_n such that $\mathbb{E}_{p_n}[e^{\beta_n g_n}] < \infty$ for all $\beta_n \in \mathbb{R}$ and define measure q_n using Radon-Nikodym derivative as:

$$\frac{dq_n}{dp_n} = \frac{e^{\beta g_n}}{\mathbb{E}_{p_n}[e^{\beta_n g_n}]}$$

We try to compare the distributions p_n and q_n in the total variation sense. We shall use the notation defined in the following discussion of the abstract result even when dealing with specific examples. Consider the following conditions:

C1 For some finite index set M_n such that $|M_n| = M(n) \in \mathbb{N}$, we can partition $\Omega = \cup_{m \in M_n} A_m$ with disjoint sets A_m such that $p_n(A_m) > 0 \; \forall \; m \in M_n$.

C2 For a set $S_n \subset M_n$, $p_n(\cup_{m \in S_n} A_m) \geq 1 - \alpha_n$ for some sequence $\alpha_n \to 0$.

C3 Let $p^{(m)}$ be the probability measure over A_m defined by $p^{(m)}(A) := \frac{p_n(A)}{p_n(A_m)} \; \forall \; A \subset A_m$ and $A \in \mathcal{F}_n$. Let $X_m \sim p^{(m)}$ and $X \sim p_n$. Let $e_m(g_n) := \mathbb{E}[g_n(X_m)]$ and $\sigma_m^2(g_n) := \mathrm{var}\,[g_n(X_m)]$. For all $m, m' \in S_n$,

$$0 < c \leq \frac{\sigma_m^2(g_n)}{\sigma_{m'}^2(g_n)} \leq C$$

for some constants c, C independent of n. We let σ_n be any sequence such that $c\sigma_m(g_n) \leq \sigma_n \leq C\sigma_m(g_n)$ for some absolute constants c and C for every $m \in S_n$. Although g_n can depend on β_n, $g_n(x) - e_m(g_n)$ does not depend on β_n for $x \in A_m$.

C4 There is a sequence $\tau_n \to 0$ such that

$$\sup_{m \in S_n} d_{\mathsf{KS}} \left(\mathcal{L} \left(\frac{g(X_m) - e_m(g_n)}{\sigma_m} \right), \mathcal{N}(0,1) \right) < \tau_n \,.$$

C5 Let $X \sim p_n$. **C3** holds, $\mathrm{var}(g_n(X)) = O(\sigma_n^2)$ and

$$\log \mathbb{E}\left[e^{\beta(g_n(X) - \mathbb{E} g_n(X))} \right] \leq \frac{C\beta^2\sigma_n^2}{1 - |\beta| D\sigma_n}$$

for all $|\beta| < \frac{1}{D\sigma_n}$ for absolute constants C, D independent of n.

Remark 6 *Condition **C4** can be relaxed to convergence to a fixed distribution with a strictly positive tail. We use standard normal distribution because it is sufficient for the examples in this paper.*

Remark 7 *We note that the function g_n can have β_n as a parameter but, condition **C3** requires that $g_n(x) - e_m(x)$ does not depend on β_n whenever $x \in A_m$. Therefore, the conditional variances do not depend on the value of β_n. A trivial example is: $g_n(x) = l(x) + \beta_n m(x)$. Other examples satisfying these conditions are given in Sections 4 and 7.*

We consider the following simple hypothesis testing problem for data $X \in \Omega_n$, and then extend this test to the composite case in Sections 4 and 7.

$$\begin{aligned} H_0 &: X \sim p_n \\ H_1 &: X \sim q_n \quad \text{for some} \quad \beta_n \,. \end{aligned}$$

The value of β_n may be unknown.

We define the following test:

Definition 8 (Canonical Test) *Let $T \geq 0$ and $\kappa : \Omega \to \mathbb{R}$. Define the function m so that $m(x) = m_0$ iff $x \in A_{m_0}$. Given a sample X, we define the decision function $\mathcal{D}^{\mathsf{can}}(X) \in \{H_0, H_1\}$:*

1. *if $m(X) \notin S_n$ then $\mathcal{D}^{\mathsf{can}}(X) = H_1$*
2. *if $m(X) \in S_n$ and $\frac{\kappa(X) - e_{m(X)}(\kappa)}{\sigma_{m(X)}(\kappa)} \geq T$ then $\mathcal{D}^{\mathsf{can}}(X) = H_1$*
3. *otherwise $\mathcal{D}^{\mathsf{can}}(X) = H_0$*

The statistical test with decision function $\mathcal{D}^{\mathsf{can}}$ is the canonical statistical test $\mathcal{T}^{\mathsf{can}}(T, \kappa)$.

We note that the canonical test depends only on the function κ, the set S_n and the conditional measures $p^{(m)}$. A natural choice of κ is: $\kappa = g_n$. We show the following result for this choice of κ. Our metric of comparison will the following 'probability of error' for any test $\mathcal{T}$ with decision function $\mathcal{D}$:

$$p_{\mathsf{error}} = \max(\mathbb{P}[\mathcal{D}(X) = H_0 | X \sim H_1], \mathbb{P}[\mathcal{D}(X) = H_1 | X \sim H_0]) \,.$$

Theorem 9 *Assume w.l.o.g that $\beta_n > 0$. We have the following results.*

1. *Suppose that conditions **C1**,**C2**,**C3**, and **C4**, hold. Then*

$$\lim_{n\to\infty} d_{\mathsf{TV}}(p_n, q_n) = 1 \quad \textit{if } \beta_n\sigma_n \to \infty \tag{2}$$

If $\beta_n\sigma_n \geq L_n$ for a known sequence $L_n \to \infty$ (β_n being possibly unknown), then the canonical test $\mathcal{T}^{\mathsf{can}}(T_n, g_n)$ can distinguish between p_n and q_n with high probability from a single sample for $T_n \to \infty$ depending only on L_n and τ_n. The probability of type 1 and type 2 errors can be bounded above by a function of α_n, T_n and L_n tending to 0.

2. *Suppose that condition **C5** holds. Then*

$$\lim_{n\to\infty} d_{\mathsf{TV}}(p_n, q_n) = 0 \quad \textit{if } \beta_n\sigma_n \to 0 \tag{3}$$

We defer the proof to Appendix A. The idea behind the first part of the proof is described in Section 2.4. To understand the proof of the second part of the theorem, we take Ω to be a finite space. Then, $q(x) = p(x)\frac{e^{\beta_n g(x)}}{\mathbb{E}_p e^{\beta_n g}}$. Condition **C5** along with Jensen's inequality implies that whenever $\beta_n\sigma_n \to 0$,

$$e^{\beta_n \mathbb{E}_p g} \leq \mathbb{E}_p e^{\beta_n g} \leq e^{\beta_n \mathbb{E}_p g} e^{\frac{C\beta_n^2\sigma_n^2}{1-D|\beta_n|\sigma_n}} = (1+o(1))e^{\beta_n \mathbb{E}_p g}$$

Therefore, $q(x) = (1-o(1))p(x)e^{\beta_n(g(x)-\mathbb{E}_p(g))}$. We use Chebyshev inequality to show that $\beta_n(g(x) - \mathbb{E}_p(g))$ is small most of the time i.e, $q(x) = (1 \pm o(1))p(x)$ with high probability. This proves that the total variation distance converges to zero.

4. Testing Ising Model Structure

We intend to test between the following hypotheses for data $\hat{X} \in \{-1, 1\}^n$:

H_0 : The data is generated by a Curie-Weiss model at an unknown inverse temperature $0 \leq \beta^{\mathrm{CW}}(n) \leq \beta_{\mathsf{max}}$ and external field $|h^{\mathrm{CW}}| \leq h_{\mathsf{max}} < \infty$

H_1 : The data is generated by an Ising model on a known d-regular graph at an unknown inverse temperature $0 \leq \beta_n^{\mathrm{dreg}} < \infty$ and arbitrary external field $h^{\mathrm{dreg}} \in \mathbb{R}$ such that $(\beta_n^{\mathrm{dreg}}, h^{\mathrm{dreg}}) \in \Lambda_{\mathsf{Ising}}$

We will apply Theorem 9 to prove Theorem 11. We use the notation from the conditions of Theorem 9. Let $x \in \Omega := \{-1, 1\}^n$. We take p_n to be Curie-Weiss model at inverse temperature $\beta^{\mathrm{CW}} \leq \beta_{\mathsf{max}}$ and external field h^{CW} such that $|h^{\mathrm{CW}}| \leq h_{\mathsf{max}} < \infty$ i.e, $p_n(x) \propto e^{\frac{n}{2}\beta^{\mathrm{CW}}m^2 + nh^{\mathrm{CW}}m(x)}$, where $m := m(x) = \frac{1}{n}\sum_i x_i$.

Let G be any known d-regular graph over n vertices with adjacency matrix A and $d = o(n)$. We take q to be the Ising model with interaction matrix $\beta_n^{\mathrm{dreg}}A$ and external field h^{dreg} such that $\beta_n^{\mathrm{dreg}} > 0$ and $h^{\mathrm{dreg}} \in \mathbb{R}$. That is, $q_n(x) \propto e^{\frac{\beta_n^{\mathrm{dreg}}}{2}x^\intercal Ax + nh^{\mathrm{dreg}}m(x)}$.

We take $g_n(x) = \frac{1}{2}x^\intercal Ax - \frac{n}{2}\frac{\beta^{\mathrm{CW}}}{\beta_n^{\mathrm{dreg}}}m^2 + \frac{nd}{2(n-1)\beta_n^{\mathrm{dreg}}} + \frac{n(h^{\mathrm{dreg}}-h^{\mathrm{CW}})}{\beta_n^{\mathrm{dreg}}}m(x)$. Therefore,

$$q_n(x) = \frac{p_n(x)e^{\beta_n^{\mathrm{dreg}}g_n(x)}}{\mathbb{E}_{p_n}[e^{\beta_n^{\mathrm{dreg}}g(x)}]}$$

Let M_n and A_{m_0} for $m_0 \in M_n$ be as defined in Section 2.4. Clearly, $A_{m_0} = \{x : |\{i : x_i = 1\}| = \frac{1+m_0}{2}n\}$. Magnetization concentration of Curie-Weiss model is well studied (c.f. Ellis (2007)). The magnetization for the Curie Weiss model concentrates at the roots of the equation $m^* = \tanh(\beta^{\mathrm{CW}} m^* + h^{\mathrm{CW}})$. Since $\beta^{\mathrm{CW}} \le \beta_{\mathsf{max}} < \infty$ and $|h^{\mathrm{CW}}| < h_{\mathsf{max}} < \infty$ one has for some $\epsilon > 0$ and constants $B, C(\beta_{\mathsf{max}}, h_{\mathsf{max}}) > 0$ depending only on β_{max} and $h_{\max}$,

$$p_n\left(m(x) \in [-1+\epsilon, 1-\epsilon]\right) \ge 1 - Be^{-C(\beta_{\mathsf{max}}, h_{\mathsf{max}})n} =: 1 - \alpha_n\,.$$

Therefore, we let $S_n = M_n \cap [-1+\epsilon, 1-\epsilon]$.

Remark 10 *Consider the canonical test for H_0 and H_1 given in Definition 8. Let a sample $\hat{X}$ with magnetization $\hat{m} = m(\hat{X})$ be given. By definition of S_n, we can determine whether $m(\hat{X}) \in S_n$ without using $(\beta_n^{\mathrm{dreg}}, h^{\mathrm{dreg}})$ and $(\beta^{\mathrm{CW}}, h^{\mathrm{CW}})$. Clearly, $p^{(\hat{m})}$ is the uniform measure over $A_{\hat{m}}$ irrespective of the value of β^{CW} and h^{CW}. Let $X_{\hat{m}} \sim p^{(\hat{m})}$. A calculation shows that*

$$\frac{1}{2}\hat{X}^\intercal A\hat{X} - \mathbb{E}\left[\frac{1}{2}X_{\hat{m}}^\intercal A X_{\hat{m}}\right] = g(\hat{X}) - e_{\hat{m}}(g)$$

Therefore, $\sigma_m^2 := \mathrm{var}(g(X_m)) = \mathrm{var}\left(\frac{1}{2}X_m^\intercal A X_m\right)$. We observe that neither of the quantities above depend on the values of the unknown parameters and hence the same is true for whether or not $\frac{g(\hat{X}) - e_{\hat{m}}(g)}{\sigma_{\hat{m}}(g)} \ge T$. Letting $\kappa_{\mathsf{Ising}}(\hat{X}) := \hat{X}^\intercal A\hat{X}$, we have that $\mathcal{T}^{\mathrm{can}}(T_n, g_n) = \mathcal{T}^{\mathrm{can}}(T_n, \kappa_{\mathsf{Ising}})$.

By Theorem 12, $\sigma_m = \Theta(\sqrt{nd})$ uniformly for all $m \in S_n$ and

$$\sup_{m \in S_n} d_{\mathsf{KS}}\left(\mathcal{L}\left(\tfrac{g(X_m) - e_m(g)}{\sigma_m}\right), \mathcal{N}(0,1)\right) < C(\epsilon)\sqrt[4]{\frac{d}{n}} =: \tau_n$$

Theorem 11 *Let $d = o(n)$ and L_n be any positive sequence diverging to infinity.*

1. *If $\Lambda_{\mathsf{Ising}} = \{(\beta_n^{\mathrm{dreg}}, h^{\mathrm{dreg}}) : \beta_n^{\mathrm{dreg}} \ge \frac{L_n}{\sqrt{nd}}, |h^{\mathrm{dreg}}| \le h_{\max}\}$, the canonical test $\mathcal{T}^{\mathrm{can}}(T_n, \kappa_{\mathsf{Ising}})$, which depends only on β_{max}, h_{max} and L_n can distinguish H_0 and H_1 with high probability for some choice of $T_n(\beta_{\mathsf{max}}, h_{\mathsf{max}}, L_n) \to \infty$.*
2. *If $\Lambda_{\mathsf{Ising}} = \{(\beta_n^{\mathrm{dreg}}, h^{\mathrm{dreg}}) : \beta_n^{\mathrm{dreg}} = \frac{1}{L_n\sqrt{nd}}, |h^{\mathrm{dreg}}| \le h_{\max}\}$, there is no statistical test which can distinguish H_0 and H_1 with high probability using constant number of i.i.d. samples.*

We defer the proof to Appendix D. The idea is to use Remark 10 to conclude $\mathcal{T}^{\mathrm{can}}(T_n, g_n) = \mathcal{T}^{\mathrm{can}}(T_n, \kappa_{\mathsf{Ising}})(\hat{X})$ and then use Theorem 9 to conclude the result.

We note from the proof that above the threshold, the distribution p_n need not necessarily be the Curie-Weiss model. It can be any family of permutation invariant probability distribution such that $p_n(m(x) \in [\delta, 1-\delta]) \to 1$ for some $\delta > 0$ and our proof for the success of our statistical test goes through.

5. A Central Limit Theorem for Quadratic Forms over Hamming Sphere

In order to apply Theorem 9 to problems of interest, we would like to prove a central limit theorem with Berry-Esseen type bounds for quadratic forms over Hamming Spheres. Consider $\mathcal{S} = \{(x_1, ..., x_n) \in \{-1, 1\}^n : |\{i : x_i = 1\}| = sn\}$. That is, $\mathcal{S}$ is the Hamming sphere of radius sn for a fixed $s \in (0, 1)$. Let $X \sim \text{unif}(\mathcal{S})$. Given an adjacency matrix A, we intend to prove a central limit theorem for the quadratic form $\mathcal{A}(X) = \frac{1}{2}X^\intercal AX$. The problem of limiting distributions has been well studied for quadratic forms of i.i.d random variables (see Hall (1984), Rotar et al. (1979), de Jong (1987), Götze and Tikhomirov (2002)). These methods use the independence of the entries of the random vector, which does not hold here. We use Stein's method to prove the following result:

Theorem 12 *Let $d = o(n)$ and A be the adjacency matrix of a d regular graph. Let $0 < \delta < s < 1 - \delta < 1$ and $\sigma_s^2 := \text{var}(\mathcal{A}(X))$ and $L = \frac{\mathcal{A}(X) - \mathbb{E}\mathcal{A}(X)}{\sigma_s}$. Then,*

1. $\sigma_s^2 = 8nds^2(1-s)^2(1 + O(\frac{d}{n}))$
2. $d_{\mathsf{KS}}(\mathcal{L}(L), \mathcal{N}(0, 1)) \leq C\sqrt[4]{\frac{d}{n}}$

C depends only on δ and the bound $O\left(\frac{d}{n}\right)$ holds uniformly for all $s \in (\delta, 1 - \delta)$.

A pair of random variables (T, T') is called exchangeable if $(T, T') \stackrel{d}{=} (T', T)$.

Definition 13 *We call a real valued exchangeable pair (T, T') an a-Stein pair with respect to the sigma algebra $\mathcal{F}$ if T is $\mathcal{F}$ measurable and $\mathbb{E}(T'|\mathcal{F}) = (1 - a)T + a\mathbb{E}(T)$*

We prove Theorem 12 using the following CLT (Theorem 3.7 in Ross et al. (2011)).

Theorem 14 *Let (W, W') be an a-Stein pair with respect to the sigma algebra $\mathcal{F}$ such that W has 0 mean and unit variance. Let N have the standard normal distribution. Then,*

$$d_{\mathsf{W}}(W, N) \leq \frac{\sqrt{\text{var}\left(\mathbb{E}\left[(W' - W)^2|\mathcal{F}\right]\right)}}{\sqrt{2\pi}a} + \frac{\mathbb{E}\left(|W - W'|^3\right)}{3a}$$

Consider the set $S(x) = \{i \in [n] : x_i = 1\}$. Let χ_S be the n dimensional column vector such that $\chi_S(i) = 1$ if $i \in S$ and $\chi_S(i) = 0$ if $i \in S^c$. We shall henceforth use $S(x)$, $\chi_S(x)$ and x interchangeably. Define $d(A, B)$ to be the number of edges of G with one vertex in A and the other in B. When $A = \{j\}$, we denote $d(A, B)$ be d_{jB}. We can easily show that $\frac{1}{2}x^\intercal Ax = \frac{nd}{2} - 2d(S(X), S(X)^c)$. It is sufficient to prove the CLT for $d(S(X), S(X)^c)$ when $X \sim \text{unif}(\mathcal{S})$. We continue with the proofs of the theorems in Appendix B.

6. Concentration of Quadratic Forms over Hamming Sphere

Let S be the uniform random set of constant size and $T(S) = d(S, S^c)$ be the size of the edge-cut, just like in Section 5. Here, we relax the constraint on the size of S so that $0 \leq |S| \leq n$. To lower bound the total variation distance, we need Condition **C5**. To prove this condition, for the examples considered in this paper, we need sub-exponential bounds of the form:

$$\log \mathbb{E} \exp\left(\beta T - \beta \mathbb{E} T\right) \leq \frac{C\beta^2 nd}{1 - D\sqrt{nd}|\beta|} \tag{4}$$

In the case of centered independent random variables, i.e, when $y_i = \mathrm{Ber}(s) - s$, Hanson-Wright inequality for quadratic forms gives a sub-exponential concentration inequality like (4), but it is not clear how to extend this to case when the space is restricted to the Hamming sphere (and therefore has weak dependencies).

To deal with this, tensorization of roughly the following form is normally proved: $\log \mathbb{E} \exp\left(\beta T - \beta \mathbb{E} T\right) \leq C\beta^2 \sum_{i=1}^{n} \mathbb{E}\Delta_i^2(T)$, where $\Delta_i(f(x)) := f(x_i^+) - f(x_i^-)$ is the discrete derivative. Here we run into a second problem: since our random set S has constant size almost surely, we cannot remove a single element and the discrete derivative $\Delta_i f(x)$ cannot be defined within our space. We use the exchangeable pair used in Section 5 and Appendix B to prove a well defined tensorization similar to the one above.

Using our method, based on Burkhölder-Davis-Gundy type inequalities proved in Chatterjee (2007), we show that

$$\log \mathbb{E} \exp \gamma(T - \mathbb{E} T) \leq \frac{32nd\gamma^2(1 + o(1))}{1 - 16nd\gamma^2(1 + o(1))} \tag{5}$$

We defer the full proof to Appendix C.

7. Comparing ERGM to Erdős-Rényi Model

Here, we compare $G(n, p_n)$ to $\mathsf{ERGM}(\beta_1, \beta_2)$. Fix $\delta > 0$. Consider the following hypothesis testing problem given a single sample of a random simple graph G over n vertices:

H_0 : G is drawn from the distribution $G(n, p)$ for some $p \in (\delta, 1 - \delta)$

H_1 : G is drawn from $\mathsf{ERGM}(\beta_1, \beta_2)$ for $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}^+$ for unknown β_1 and β_2 such that $(\beta_1, \beta_2) \in \Lambda_{\mathsf{ERGM}}$

Given a sample graph X, we let $V(X)$ be the number of wedge graphs ($\vee$) in X.

Theorem 15 *Let L_n be any positive sequence diverging to infinity.*

1. *If $\Lambda_{\mathsf{ERGM}} = \{(\beta_1, \beta_2) : \beta_2 \geq L_n \frac{1}{\sqrt{n}}, \beta_1 \in \mathbb{R}\}$ then the canonical statistical test $\mathcal{T}^{\mathrm{can}}(T_n, V)$, which depends only on δ and L_n, can distinguish H_0 and H_1 with high probability for some choice of $T_n(\delta, L_n) \to \infty$.*

2. *If $\Lambda_{\mathsf{ERGM}} = \{(\beta_1, \beta_2) : 0 \leq \beta_2 = \frac{1}{L_n\sqrt{n}}, \beta_1 \in \mathbb{R}\}$, then there is no statistical test which can distinguish H_0 and H_1 with high probability using constant number of i.i.d. samples.*

We proceed in a way similar to Section 4 by proving each of the conditions (C1) - (C5). We defer the proof to Appendix E.

Acknowledgment

We thank the anonymous reviewers for their helpful suggestions. This work was supported in part by ONR N00014-17-1-2147, DARPA W911NF-16-1-0551, and NSF CCF-1565516.

References

Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the Potts model. *Probability Theory and Related Fields*, 168(3-4):557–600, 2017.

Shankar Bhamidi, Guy Bresler, and Allan Sly. Mixing time of exponential random graphs. *The Annals of Applied Probability*, pages 2146–2170, 2011.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Guy Bresler and Dheeraj M Nagaraj. Stein's method for stationary distributions of markov chains and application to Ising models. *arXiv preprint arXiv:1712.05743*, 2017.

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.

Clement L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing Bayesian networks. In *Conference on Learning Theory*, pages 370–448, 2017.

Sourav Chatterjee. Concentration inequalities with exchangeable pairs (ph. d. thesis). *arXiv preprint math/0507526*, 2005.

Sourav Chatterjee. Stein's method for concentration inequalities. *Probability theory and related fields*, 138(1):305–321, 2007.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. *arXiv preprint arXiv:1612.03147*, 2016.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems*, pages 12–22, 2017.

Peter de Jong. A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields*, 75(2):261–277, 1987.

Ronen Eldan and Renan Gross. Exponential random graphs behave like mixtures of stochastic block models. *arXiv preprint arXiv:1707.01227*, 2017.

Richard Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, 2007.

Chao Gao and John Lafferty. Testing network structure using relations between small subgraph probabilities. *arXiv preprint arXiv:1704.06742*, 2017.

Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting Ising models. *arXiv preprint arXiv:1706.00121*, 2017.

Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-sample tests for large random graphs using network statistics. *Proceedings of Machine Learning Research vol*, 65:1–24, 2017.

F Götze and A Tikhomirov. Asymptotic distribution of quadratic forms and applications. *Journal of Theoretical Probability*, 15(2):423–475, 2002.

Peter Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, 14(1):1–16, 1984.

Hamid Javadi and Andrea Montanari. A statistical model for motifs detection. *arXiv preprint arXiv:1511.05254*, 2015.

Abraham Martín del Campo, Sarah Cepeda, and Caroline Uhler. Exact goodness-of-fit testing for the ising model. *Scandinavian Journal of Statistics*, 44(2):285–306, 2017.

Sumit Mukherjee. Consistent estimation in the two star exponential random graph model. *arXiv preprint arXiv:1310.4526*, 2013.

Nathan Ross et al. Fundamentals of stein's method. *Probability Surveys*, 8:210–293, 2011.

Vladimir I Rotar et al. Limit theorems for polylinear forms. *Journal of Multivariate analysis*, 9(4):511–530, 1979.

Appendix A. Proof of Main Abstract Theorem 9

We first consider the case when $\sigma_n\beta_n \to \infty$. Given a sample from p_n or q_n, we prove that the statistical test $\mathcal{T}^{\mathsf{can}}(T_n, g_n)$ succeeds with high probability for some choice of T_n. Let $\mathcal{D}^{\mathsf{can}}$ be the decision function associated with the test $\mathcal{T}^{\mathsf{can}}(T_n, g_n)$.
Consider the type 1 error rate:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{D}^{\mathsf{can}}(X) = H_1 | X \sim H_0\right) &= p_n\left(m(X) \notin S_n\right) + \sum_{m\in S_n} p_n\left(\frac{g(X)-e_m(g)}{\sigma_m} \geq T \middle| m(X)=m\right) p_n\left(A_m\right) \\
&\leq \alpha_n + \sum_{m\in S_n}\left[1-\Phi(T) + d_{\mathsf{KS}}\left(\mathcal{L}\left(\tfrac{g(X_m)-e_m(g)}{\sigma_m}\right), \mathcal{N}(0,1)\right)\right] p\left(A_m\right) \\
&\leq \alpha_n + 1 - \Phi(T) + \tau_n \qquad (6)
\end{aligned}$$

Now consider the type 2 error rate:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{D}^{\mathsf{can}}(X) = H_0 | X \sim H_1\right) &= q_n\left(\frac{g(X)-e_{m(X)}(g)}{\sigma_{m(X)}} < T, m(X) \in S_n\right) \\
&= \sum_{m\in S_n} q_n\left(\frac{g(X)-e_m(g)}{\sigma_m} < T \middle| m(X)=m\right) q_n\left(A_m\right) \\
&= \sum_{m\in S_n} \frac{q_n\left(\frac{g(X)-e_m(g)}{\sigma_m} < T \middle| X\in A_m\right)}{q_n\left(\frac{g(X)-e_m(g)}{\sigma_m} < 2T \middle| X\in A_m\right) + q_n\left(\frac{g(X)-e_m(g)}{\sigma_m} \geq 2T \middle| X\in A_m\right)} q_n\left(A_m\right) \\
&\leq \sum_{m\in S_n} \frac{q_n\left(\frac{g(X)-e_m(g)}{\sigma_m} < T \middle| X\in A_m\right)}{q_n\left(\frac{g(X)-e_m(g)}{\sigma_m} \geq 2T \middle| X\in A_m\right)} q_n\left(A_m\right) \\
&= \sum_{m\in S_n} \frac{\int_{g<e_m+T\sigma_m} e^{\beta_n g} dp^{(m)}}{\int_{g\geq e_m+2T\sigma_m} e^{\beta_n g} dp^{(m)}} q_n\left(A_m\right) \\
&\leq \sum_{m\in S_n} \frac{e^{(\beta_n e_m + T\beta_n\sigma_m)}}{p^{(m)}(\{g\geq e_m+2T\sigma_m\}) e^{(\beta_n e_m + 2T\beta_n\sigma_m)}} q\left(A_m\right) \\
&\leq \sum_{m\in S_n} \frac{e^{-T\beta_n\sigma_m}}{1-\Phi(2T)-\tau_n} q\left(A_m\right) \\
&\leq \frac{e^{-cT\beta_n\sigma_n}}{1-\Phi(2T)-\tau_n} \qquad (7)
\end{aligned}$$

We use the fact that for positive x and y, $\max(x,y) \leq x+y$, equation (6) and (7), to conclude that for every $T>0$ such that $1-\Phi(2T) > \tau_n$ the error rate p_{error}

$$p_{\mathsf{error}} \leq \alpha_n + 1 - \Phi(T) + \tau_n + \frac{e^{-cT\beta_n\sigma_n}}{1-\Phi(2T)-\tau_n} \leq \alpha_n + 1 - \Phi(T) + \tau_n + \frac{e^{-cTL_n}}{1-\Phi(2T)-\tau_n} \quad (8)$$

For n large enough, $\tau_n + e^{-cL_n} < \frac{1}{2}$. For such n, we can pick $T = T_n > 0$ such that

$$1 - \Phi(2T_n) = \tau_n + e^{-cL_n}$$

Clearly, $T_n \to \infty$, therefore, $1-\Phi(T_n) \to 0$ and

$$\frac{e^{-cT_nL_n}}{1-\Phi(2T_n)-\tau_n} = e^{-c(T_n-1)L_n} \to 0$$

Using the equations above in equation (8), we conclude that:

$$p_{\mathsf{error}} \leq \alpha_n + 1 - \Phi(T_n) + \tau_n + e^{-c(T_n-1)L_n} \to 0$$

Therefore, the decision function $\mathcal{D}^{\mathsf{can}}(X)$ has a vanishing error rate for the choice of $T = T_n$ made above. Let $A^{\mathsf{can}} = \{x \in \Omega : \mathcal{D}^{\mathsf{can}}(x) = H_0\}$

$$\begin{aligned}
d_{\mathsf{TV}}(p_n, q_n) &= \sup_{A \in \mathcal{F}_n} p_n(A) - q_n(A) \\
&\geq p_n(A^{\mathsf{can}}) - q_n(A^{\mathsf{can}}) \\
&= 1 - p_n\left((A^{\mathsf{can}})^c\right) - q_n(A^{\mathsf{can}}) \\
&= 1 - \mathbb{P}(\mathcal{D}^{\mathsf{can}}(X) = H_1 | X \sim H_0) - \mathbb{P}(\mathcal{D}^{\mathsf{can}}(X) = H_0 | X \sim H_1) \\
&\geq 1 - 2\max(\mathbb{P}(\mathcal{D}^{\mathsf{can}}(X) = H_1 | X \sim H_0), \mathbb{P}(\mathcal{D}^{\mathsf{can}}(X) = H_0 | X \sim H_1)) \\
&= 1 - 2p_{\mathsf{error}}
\end{aligned} \tag{9}$$

Using Equation (9) we conclude that whenever $\sigma_n \beta_n \to \infty$,

$$d_{\mathsf{TV}}(p_n, q_n) \to 1$$

We now consider the case $\beta_n \sigma_n \to 0$. Consider the set $A_{g_n} = \{x \in \Omega : \frac{e^{\beta_n g_n}}{\mathbb{E}_{p_n} e^{\beta_n g_n}} < 1\}$. It can be easily shown that $A_{g_n} \in \mathcal{F}_\setminus$ and $d_{\mathsf{TV}}(p_n, q_n) = p_n(A_{g_n}) - q_n(A_{g_n})$. Let $Z_{g_n} := \mathbb{E}_{p_n} e^{\beta_n g_n}$. Since, $\beta_n \sigma_n \to 0$, the following inequalities hold when $\beta_n \sigma_n$ is small enough and any $T > 0$

$$\begin{aligned}
d_{\mathsf{TV}}(p_n, q_n) &= p_n(A_{g_n}) - q_n(A_{g_n}) \\
&= \int \mathbb{1}_{A_{g_n}} \left(1 - \frac{e^{\beta_n g_n}}{\mathbb{E}_p e^{\beta_n g_n}}\right) dp_n \\
&= \int \mathbb{1}_{A_{g_n}} \left(1 - e^{-|\beta_n g_n - \log Z_{g_n}|}\right) dp_n \\
&\leq \int \left(1 - e^{-|\beta_n g_n - \log Z_{g_n}|}\right) dp_n \\
&\leq \int \left(1 - e^{-|\beta_n g_n - \beta_n \mathbb{E}_{p_n}[g_n]|} e^{-|\log Z_{g_n} - \beta_n \mathbb{E}_{p_n}[g]|}\right) dp_n \\
&\leq \int \left(1 - e^{-\frac{A\beta_n^2\sigma_n^2}{1 - B|\beta_n|\sigma_n}} e^{-|\beta_n(g_n - \mathbb{E}_{p_n}[g_n])|}\right) dp_n \\
&\leq p_n\left(|g_n - \mathbb{E}_{p_n}[g_n]| \geq T\right) + 1 - \exp\left(-\frac{A\beta_n^2\sigma_n^2}{1 - B\beta_n\sigma_n}\right) e^{-\beta_n T} \\
&\leq \frac{\sigma_n^2}{T^2} + 1 - \exp\left(-\frac{A\beta_n^2\sigma_n^2}{1 - B\beta_n\sigma_n}\right) e^{-\beta_n T}
\end{aligned} \tag{10}$$

Where we have used the Chebyshev bound in the last step and the subexponentiality of g_n. The coefficients (A, B) are consistent with coefficients (C, D) in condition **C5**. Let $\gamma_n \to 0$ be any positive sequence such that $\frac{\beta_n \sigma_n}{\gamma_n} \to 0$. Let $T = \frac{\gamma_n}{\beta_n}$. Using this choice of T in Equation (10), we conclude that:

$$d_{\mathsf{TV}}(p_n, q_n) \leq \frac{\beta_n^2\sigma_n^2}{\gamma_n^2} + 1 - \exp\left(-\frac{A\beta_n^2\sigma_n^2}{1 - B\beta_n\sigma_n}\right) \exp\left(-\gamma_n\right) \to 0.$$

Appendix B. Proof of Central Limit Theorem

We use the quantities as defined in Section 5.

For the sake of clarity, we denote the random variable $S(X)$ by just S. Clearly, $|S| = sn =: l$. Define $T(S) := d(S, S^c)$. Recall that our objective is to prove a CLT for $T(S)$. We define the following exchangable pair (S, S') : Draw K and $J \in \{1, ..., n\}$ uniformly at random and independent of each other and independent of S. Define $\chi_{S'}$ to be the vector obtained by exchanging entries at indices K and J of χ_S.

A calculation using the fact that G is d-regular shows that

$$T(S') = \begin{cases} T(S) & \text{if } \chi_S(J) = \chi_S(K) \\ T(S) + 2(d_{J,S} - d_{K,S} + d_{J,K}) & \text{if } J \in S \text{ and } K \in S^c \\ T(S) + 2(d_{K,S} - d_{J,S} + d_{J,K}) & \text{if } K \in S \text{ and } J \in S^c \end{cases} \tag{11}$$

We apply Theorem 14 to the centered and normalized version of the Stein pair $(T(S), T'(S))$ to prove Theorem 12.

Lemma 16 *$(T(S), T(S'))$ is a λ-Stein pair with respect to $\mathcal{F}(S)$, where $\lambda = 4\frac{n-1}{n^2}$. Further, $\mathbb{E}[T(S)] = \frac{l(n-l)d}{n-1}$*

Proof Clearly,

$$\mathbb{E}[T(S')|S] = T(S) + \tfrac{4}{n^2} \sum_{j \in S} \sum_{k \in S^c} d_{j,S} - d_{k,S} + d_{j,k} \tag{12}$$

$$= T(S) + \frac{4}{n^2} \sum_{j \in S} (d - d_{j,S^c})(n - l) - \sum_{k \in S^c} l d_{k,S} + \sum_{j \in S} k \in S^c d_{j,k} \tag{13}$$

$$= \left(1 - 4\frac{n-1}{n^2}\right) T(S) + 4\frac{l(n-l)d}{n^2} \tag{14}$$

Using the fact that $\mathbb{E}T(S) = \mathbb{E}T(S')$, we conclude the result. ■

We shall henceforth shorten $T(S')$ to T' and define $\lambda := 4\frac{n-1}{n^2}$. We list some elementary results about various moments.

Lemma 17 *For a d-regular graph, when $n - d - 2 > l > d + 2$, if $l = \theta(n)$*

1. $\mathbb{E} \sum_{j \in S} \sum_{k \in S^c} d_{j,S}^2 = l(n-l) \left(d^2 \frac{(l-1)(l-2)}{(n-1)(n-2)} + d \frac{(l-1)(n-l)}{(n-1)(n-2)} \right)$
2. $\mathbb{E} \sum_{j \in S} \sum_{k \in S^c} d_{k,S}^2 = l(n-l) \left(d^2 \frac{(l)(l-1)}{(n-1)(n-2)} + d \frac{(l)(n-l-1)}{(n-1)(n-2)} \right)$
3. $\mathbb{E} \sum_{j \in S} \sum_{k \in S^c} d_{j,S} d_{k,S} = d^2 l(n-l) \frac{(l)(l-1)}{(n-1)^2} - \mathrm{var}(T)$
4. $\mathbb{E} \sum_{j \in S} \sum_{k \in S^c} d_{k,S} d_{j,k} = O(nd^2)$
5. $\mathbb{E} \sum_{j \in S} \sum_{k \in S^c} d_{j,S} d_{j,k} = O(nd^2)$

6. $\mathbb{E}\sum_{j\in S}\sum_{k\in S^c} d_{j,k}^2 = O(nd)$

Proof

1. $$\mathbb{E}\sum_{j\in S}\sum_{k\in S^c} d_{j,S}^2 = \frac{l(n-l)}{n}\sum_{j=1}^{n}\mathbb{E}(d_{j,S}^2|j\in S)$$
   Denoting the neighborhood of j by $N(j)$,
   $$\mathbb{E}(d_{j,S}^2|j\in S) = \sum_{a,b\in N(j)}\mathbb{P}(a\in S, b\in S|j\in S)$$
   A simple computation of the probability gives the result.
2. proof similar to the previous part.
3. $$\mathbb{E}\sum_{j\in S}\sum_{k\in S^c} d_{j,S}d_{k,S} = \mathbb{E}\left[(ld - T(S))\,T(S)\right]$$
   Using the fact that $\mathbb{E}\left[T(S)\right] = d\frac{l(n-l)}{n-1}$ we arrive at the result.
4. The result follows from the fact that
   $$\sum_{j\in S}\sum_{k\in S^c} d_{j,k}d_{k,S} = \sum_{k\in S^c} d_{k,S}^2$$
5. $$\sum_{j\in S}\sum_{k\in S^c} d_{j,k}d_{j,S} = \sum_{j\in S} d_{j,S}d_{j,S^c} = O(nd^2)$$
6. We note that since G is a d-regular graph, $d_{j,k}^2 = d_{j,k}$. Therefore,
   $$\mathbb{E}\sum_{j\in S}\sum_{k\in S^c} d_{j,k}^2 = \mathbb{E}d(S, S^c)$$

■

Lemma 18

$$\mathrm{var}(T) = \frac{1}{2\lambda}\mathbb{E}\left[(T-T')^2\right]\ . \tag{15}$$

If $l = \theta(n)$, then,

$$\mathrm{var}(T) = 2dn\frac{l^2(n-l)^2}{n^4} + O(d^2)$$

Denoting $l = sn$ and $s \in (0,1)$,

$$\sigma^2 := \mathrm{var}(T) = 2ds^2(1-s)^2(1+O(\frac{d}{n}))$$

. The $O(\frac{d}{n})$ holds uniformly for all $s \in [\delta, 1-\delta]$ when $0 < \delta < 1-\delta < 1$

Proof Equation 15 follows from the fact that (T, T') forms a λ-Stein pair.

$$\text{var}(T) = \frac{n^2}{8(n-1)}\mathbb{E}\left[\mathbb{E}\left[(T-T')^2|S\right]\right] = \frac{n^2}{8(n-1)}\left(\mathbb{E}\frac{8}{n^2}\sum_{j\in S}\sum_{k\in S^c}(d_{j,s}+d_{j,k}-d_{k,S})^2\right)$$

$$= \frac{1}{n-1}\mathbb{E}\sum_{j\in S}\sum_{k\in S^c}(d_{j,s}^2+d_{k,S}^2-2d_{j,S}d_{k,S}+d_{j,k}^2+2d_{j,k}d_{j,S}-2d_{j,k}d_{k,S})$$

We use Lemma 17 to compute this expectation. ■

Lemma 19 *Let $\gamma(s) = \sum_{i=0}^{r} a_i s^i$ be any polynomial such that $0 \leq \gamma(s) \leq \alpha \; \forall \; s \in [s_1, s_2]$ such that $s_1 < s_2$, then $|a_i| \leq C\alpha$ for some constant C depending only on r, s_1 and s_2*

Proof Choose distinct $x_i \in [s_1, s_2]$ for $i \in \{0, 1, .., r\}$. Let $\boldsymbol{a} = [a_0 \; a_1 \; .. \; a_r]^\intercal$ and $\boldsymbol{b} = [\gamma(x_0) \; \gamma(x_1) \; ... \; \gamma(x_r)]^\intercal$. Consider the Vandermonde matrix with entries $V_{i,j} = x_i^j$ for $i, j \in \{0, 1, ..., r\}$. V is invertible since x_i are distinct and $V\boldsymbol{a} = \boldsymbol{b}$. Therefore, $\boldsymbol{a} = V^{-1}\boldsymbol{b}$. Therefore $\|a\|_\infty \leq \|V^{-1}\|_\infty \|b\|_\infty$. Since $\|b\|_\infty \leq \alpha$, we obtain the result by setting $C = \|V^{-1}\|_\infty$. ■

Definition 20 (function type) *Let R be a subset of vertices of a given graph G. We define the following classification of functions $f(R)$*

1. *We call f to be of **type 1** of index $r \in \mathbb{N}$ if $f(R) = (d_{j,R} - d_{k,R} + d_{j,k})^r \; \mathbb{1}_{j\in R}\mathbb{1}_{k\in R^c}$,*
2. *We call f to be of **type 2** of index $r \in \mathbb{N}$ if*

$$f(R) = (d_{j_1,R} - d_{k_1,R} + d_{j_1,k_1})^{r_1} \; (d_{j_2,R} - d_{k_2,R} + d_{j_2,k_2})^{r_2} \; \mathbb{1}_{j_1\in S}\mathbb{1}_{k_1\in R^c}\mathbb{1}_{j_2\in S}\mathbb{1}_{k_2\in S^c}$$

such that $r_1, r_2 \in \mathbb{N}$ and $r = r_1 + r_2$.

Since the coordinates of the random set S are dependent (because $|S| = l$), it is hard to bound moments of functions of S. Therefore, we draw a random set $\tilde{S}$ such that each vertex is included independently with probability $p = \frac{l}{n}$. As we shall see, S is locally similar to $\tilde{S}$ and hence we can use the known tools for bounding moments of functions of independent variables to bound the moments of $f(S)$.

Lemma 21 *Let f be a function of type 1 or type 2 with G being a d-regular graph. Then, the following are true. Let τ by the 'type' of the function f.*

1. *$f(R) = \sum_{h=0}^{r+2\tau} g_h(R) \; \forall R \subset V$*
2. *If each vertex is included in the set $\tilde{S}$ independently with probability $p = \frac{l}{n}$, then,*

$$\mathbb{E}f\left(\tilde{S}\right) = \sum_{h=0}^{r+2\tau} a_h p^h$$

for some constants $a_h \in \mathbb{Z}$.

3. *If the set S is chosen uniformly at random from all vertex subsets of size l, then* $\mathbb{E}f(S) = \sum_{h=0}^{r+2\tau} a_h \prod_{i=0}^{h-1} \frac{l-i}{n-i}$

Where $g_h(S)$ is a function of the form $\sum_{i\in I}(-1)^{\eta_i}\mathbb{1}_{S_i\subset S}$. Where $\eta_i \in \{-1,+1\}$, $S_i \subset V$, $|S_i| = h$ and I_h is any finite index set.

Proof

1. We use the following identities:

$$d_{j,S} = \sum_{i\in N(j)} \mathbb{1}_{i\in S}\,.$$

$$\mathbb{1}_{i\in S^c} = 1 - \mathbb{1}_{i\in S}\,.$$

   Expanding the power and noting that $\mathbb{1}_{S_1\subset S}\mathbb{1}_{S_2\subset S} = \mathbb{1}_{S_1\cup S_2\subset S}$, we obtain the result.

2. This follows trivially since $\mathbb{E}\mathbb{1}_{S_i\subset T} = p^{|S_i|}$ and if $g_h(T)$ is of the form above, $a_h = \sum_{i\in I}(-1)^{\eta_i}$.

3. This follows from the fact that $\mathbb{E}\mathbb{1}_{S_i\subset S} = \frac{\binom{n-|S_i|}{l-|S_i|}}{\binom{n}{l}} = \prod_{i=0}^{h-1}\frac{l-i}{n-i}$, where $h = |S_i|$. If $g_h(T)$ is of the form above, $a_h = \sum_{i\in I_h}(-1)^{\eta_i}$

■

Lemma 22 *If f is of type 1 or 2 for a d regular graph G over n vertices with a fixed index r. Let τ by the 'type' of the function.*

1. $\mathbb{E}f(\tilde{S}) = O\left(d^{\frac{r}{2}}\right)$
2. $|\mathbb{E}f(\tilde{S}) - \mathbb{E}f\,(S)\,| = O\left(\frac{d^{\frac{r}{2}}}{n}\right)$ *when* $p = \frac{l}{n}$
3. $\mathbb{E}f(S) \le Cd^{\frac{r}{2}}\left(1 + O(\frac{1}{n})\right)$

Proof

1. Let f be of type 1. Then,

$$\begin{aligned}|\mathbb{E}f(\tilde{S})| &\le \mathbb{E}|d_{j,\tilde{S}} - d_{k,\tilde{S}} + d_{j,k}|^r \\ &\le \left(1 + 2\left(\mathbb{E}|d_{j,\tilde{S}} - \mathbb{E}d_{j,\tilde{S}}|^r\right)^{\frac{1}{r}}\right)^r \end{aligned} \tag{16}$$

   Where the inequalities above follow from Minkowski's inequality and the fact that $d_{j,\tilde{S}}$ and $d_{k,\tilde{S}}$ are identically distributed.

   $d_{j,\tilde{S}}$ is a 1 Lipschitz function of $\tilde{S}$ with respect to Hamming distance. We use MacDiarmid's inequality to conclude that

$$\mathbb{P}(|d_{j,\tilde{S}} - \mathbb{E}d_{j,\tilde{S}}| > t) \le 2\exp^{-\frac{2t^2}{d}}$$

From the above, we obtain the estimate:

$$\mathbb{E}|d_{j,\tilde{S}} - \mathbb{E}d_{j,\tilde{S}}|^r \leq \int_0^\infty 2rt^{r-1}e^{-\frac{2t^2}{d}} = \left(\frac{r\Gamma(\frac{r}{2})}{4}\right) d^{\frac{r}{2}} = O(d^{\frac{r}{2}})$$

Plugging it back into equation (16), we obtain the result.

For any type 2 function g, we use Cauchy Schwarz inequality to note that:

$$\begin{aligned}|\mathbb{E}g(\tilde{S})| &\leq \mathbb{E}|d_{j_1,\tilde{S}} - d_{k_1,\tilde{S}} + d_{j_1,k_1}|^{r_1}|d_{j_2,\tilde{S}} - d_{k_2,\tilde{S}} + d_{j_2,k_2}|^{r_2} \\ &\leq \sqrt{\mathbb{E}|d_{j_1,\tilde{S}} - d_{k_1,\tilde{S}} + d_{j_1,k_1}|^{2r_1}}\sqrt{\mathbb{E}|d_{j_2,\tilde{S}} - d_{k_2,\tilde{S}} + d_{j_2,k_2}|^{2r_2}}\end{aligned}$$

And note that $\sqrt{\mathbb{E}|d_{j_i,\tilde{S}} - d_{k_i,\tilde{S}} + d_{j_i,k_i}|^{2r_i}} = O\left(d^{\frac{2r_i}{2}}\right)$ for $i = 1, 2$, as shown above, to conclude the result.

2. We use Lemma 21 to conclude that $\mathbb{E}f(\tilde{S}) = \sum_{h=0}^{r+2\tau} a_h p^h = L(p)$. Using the result in part 1, we conclude that for some absolute constant depending only on r, $L(p) \leq \alpha := Cd^{\frac{r}{2}}$ for every $p \in [0, 1]$. We then invoke Lemma 19 to show that $|a_h| \leq C_1\alpha$ for all $h \in \{0, 1, ..., r + 2\tau\}$ and that

$$|\mathbb{E}f(\tilde{S}) - \mathbb{E}f(S)| \leq \sum_{h=0}^{r+2\tau} |a_h| \left| \left(\frac{l}{n}\right)^h - \prod_{i=0}^{h-1} \frac{l-i}{n-i} \right|$$

For a fixed r, $|\left(\frac{l}{n}\right)^h - \prod_{i=0}^{h-1} \frac{l-i}{n-i}| = O(\frac{1}{n})$ for every $l \leq n$. Therefore,

$$|\mathbb{E}f(\tilde{S}) - \mathbb{E}f(S)| \leq \left(\frac{C_2}{n}\right) d^{\frac{r}{2}}$$

3. This follows from parts 1 and 2.

■

Using the fact that the co-ordinates of the vector χ_S are weakly dependent, we prove the following bound on the expectation of type 1 and type 2 functions. This gives an explicit bound on the constant $C(r)$ for every l, which will be useful when proving concentration inequalities for $d(S, S^c)$.

Lemma 23 *If f is a function of type 1 or type 2 of index r and $0 \leq l \leq n$ then*

$$\mathbb{E}|f(S)| \leq C(r)d^{\frac{r}{2}}$$

where $C(r)$ is a constant depending only on r.

Proof It is suffient to prove this result for type 1 functions since this implies the result for type 2 functions through Cauchy-Schwarz inequality. Also, it is sufficient to prove this result when r is even since an application of Jensen's inequality for the concave function $x^{\frac{r-1}{r}}$ implies the result for odd integers. Assume r is even and f is a type 1 function defined by:

$$f(S) = (d_{j,S} - d_{k,S} + d_{j,k})^r \mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}$$

Define variable $y_i(S) := \mathbb{1}_{i\in S}$. We note that,

$$\begin{aligned} f(S) &= \left(\sum_{i\in N(j)\setminus k} y_i - \sum_{i_1\in N(k)\setminus j} y_{i_1}\right)^r \mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c} \\ &\leq \mathbb{E}\left|\sum_{i\in N(j)\setminus k} y_i - \sum_{i_1\in N(k)\setminus j} y_{i_1}\right|^r \mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c} \\ &\leq \mathbb{E}\left|\sum_{i\in N(j)\setminus k} y_i - \sum_{i_1\in N(k)\setminus j} y_{i_1}\right|^r \end{aligned}$$

Define

$$g_{jk}(S) = \sum_{i\in N(j)\setminus k} y_i - \sum_{i_1\in N(k)\setminus j} y_{i_1}$$

g_{jk} is a function of $(y_i)_{i\in D_{jk}}$ where $D_{j,k} = (N(j)\setminus\{k\})\,\Delta\,(N(k)\setminus\{j\})$ and $|D_{j,k}| := h \leq 2(d-1)$. g_{jk} is 1 Lipschitz in Hamming distance.

We follow the concentration inequalities as given in Section 4.2 of Chatterjee (2005). We fix j and k such that $j \neq k$. $y_{\sim r} := (y_i : i \in D_{j,k}\setminus r)$. Let μ_i be the law of y_i. Define the dependency matrix $L = (a_{rs})$ $r, s \in D_{j,k}$ to be a matrix such that

$$d_{\mathrm{TV}}(\mu_r(.|y_{\sim r}), \mu_r(.|\hat{y}_{\sim r})) \leq \sum_{s\in D_{j,k}} a_{rs}\mathbb{1}_{y_s\neq\hat{y}_s}$$

Let $h_1 = d_H(y_{\sim r})$ and $h_2 = d_H(y_{\sim r})$. We consider two cases:

1. $l > h$

$$\begin{aligned} d_{\mathrm{TV}}(\mu_r(.|y_{\sim r}), \mu_r(.|\hat{y}_{\sim r})) &= |\mu_r(1|y_{\sim r}) - \mu_r(1|\hat{y}_{\sim r})| \\ &= \left|\frac{\binom{n-h}{l-h_1-1}}{\binom{n-h+1}{l-h_1}} - \frac{\binom{n-h}{l-h_2-1}}{\binom{n-h+1}{l-h_2}}\right| \\ &= \left|\frac{h_1-h_2}{n-h+1}\right| \\ &\leq \sum_{s\in D_{j,k}\setminus\{r\}} \frac{1}{n-h+1}\mathbb{1}_{y_s\neq\hat{y}_s} \end{aligned}$$

2. $l \leq h$
   This is similar to the previous case. It is clear that $d_{\mathsf{H}}(y_{\sim r}) \leq l$ a.s. Therefore, simple calculation shows that

$$\mu_r(1|y_{\sim r}) = \begin{cases} 0 & \text{if } h_1 = l \\ \frac{l-h_1}{n-h+1} & \text{if } h_1 < l \end{cases} \tag{17}$$

   Proceeding similar to the previous case, we conclude the result.

Therefore, we set $a_{rs} = \frac{1}{n-h+1}$ when $r \neq s$ and $a_{rr} = 0$. A is a symmetric matrix. Therefore, $\|A\|_2 \leq \|A\|_1 = \frac{h-1}{n-h+1}$. Applying theorem 4.3 from Chatterjee (2005), we have

$$\mathbb{P}(|g_{j,k} - \mathbb{E}(g_{j,k})| > t) \leq 2\exp\left(-\left(\frac{1-\frac{h-1}{n-h+1}}{h}\right)t^2\right) \tag{18}$$

Since $h \leq 2(d-1) = o(n)$, we conclude that $g_{j,k}$ is subgaussian with a variance proxy of $\frac{h}{2}(1+o(1))$. We also note that $\mathbb{E}(g_{j,k}) = 0$. We can bound the centralised moments of a sub-Gaussian random variable from Equation (18) as shown in Boucheron et al. (2013) Theorem 2.1 :

$$\mathbb{E}(g_{jk})^{2q} \leq 2(q!)[h(1+o(1))]^q \leq 2(q!)[2d(1+o(1))]^q\,, \tag{19}$$

where $q \in \mathbb{N}$ is arbitrary. Taking $r = 2q$ yields the result. ■

Let $Y(S) := \frac{T(S)-\mathbb{E}T(S)}{\sigma}$. We intend to apply Theorem 14 to the Stein pair (Y, Y') when $d = o(n)$.

We first bound the term $\frac{\mathbb{E}(|Y-Y'|^3)}{3\lambda}$ in the following lemma.

Lemma 24 *$\forall\, s \in (\delta, 1-\delta)$ such that $0 < \delta < \frac{1}{2}$, we have*

$$\frac{\mathbb{E}\left(|Y-Y'|^3\right)}{3\lambda} = O\left(\sqrt{\frac{1}{n}}\right)$$

and the bound is uniform for all $s \in (\delta, 1-\delta)$.

Proof Using Lemma 18,

$$\frac{\mathbb{E}\left(|Y-Y'|^3\right)}{3\lambda} = \frac{\mathbb{E}\left(|T-T'|^3\right)}{3\lambda\sigma^3} = \frac{C\mathbb{E}\left(|T-T'|^3\right)\left(1+O\left(\frac{d}{n}\right)\right)}{\sqrt{n}d^{\frac{3}{2}}s^3(1-s)^3} \tag{20}$$

Conditioning on S,

$$\begin{aligned}
\mathbb{E}\left(|T-T'|^3\right) &= \mathbb{E}\frac{16}{n^2}\sum_{j\in S}\sum_{k\in S^c}|d_{j,S} - d_{k,S} + d_{j,k}|^3 \\
&= \frac{16}{n^2}\sum_{j\in V}\sum_{k\in V}\mathbb{E}\left(|d_{j,S} - d_{k,S} + d_{j,k}|^3 \mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}\right) \\
&= O\left(d^{\frac{3}{2}}\right)
\end{aligned}$$

Where we get the last relation using Lemma 22. Substituting in Equation 20, we conclude the result. ■

We now bound the second term. Since $|Y - Y'| = \frac{|T-T'|}{\sigma}$. Therefore,

$$\begin{aligned}\frac{\sqrt{\text{var}(\mathbb{E}((Y'-Y)^2|S))}}{\sqrt{2\pi}\lambda} &= \frac{1}{\sqrt{2\pi}\lambda\sigma^2}\sqrt{\text{var}(\mathbb{E}((T'-T)^2|S))} \\ &= \frac{1}{\sqrt{2\pi}\lambda\sigma^2}\sqrt{\text{var}(\mathbb{E}((T'-T)^2|S))} \\ &= \frac{1}{\sqrt{2\pi}\lambda\sigma^2}\sqrt{\text{var}\left(\frac{8}{n^2}\sum_{j\in S}\sum_{k\in S^c}(d_{j,S}+d_{j,k}-d_{k,S})^2\right)} \\ &= \sqrt{\frac{2}{\pi}}\frac{1}{(n-1)\sigma^2}\sqrt{\text{var}\left(\sum_{j\in V}\sum_{k\in V}(d_{j,S}+d_{j,k}-d_{k,S})^2\mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}\right)}\end{aligned} \tag{21}$$

For $j, k \in V$, define $h_{j,k}(R) := (d_{j,R} + d_{j,k} - d_{k,R})^2\mathbb{1}_{j\in R}\mathbb{1}_{k\in R^c}$. Clearly,

$$\text{var}\left(\sum_{j\in V}\sum_{k\in V}(d_{j,S}+d_{j,k}-d_{k,S})^2\mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}\right) = \sum_{j,k,j_1,k_1\in V}\text{cov}(h_{j,k}(S), h_{j_1,k_1}(S)) \tag{22}$$

Using Lemma 22, when $p = \frac{l}{n}$,

$$\text{cov}(h_{j,k}(S), h_{j_1,k_1}(S)) = \text{cov}(h_{j,k}(\tilde{S}), h_{j_1,k_1}(\tilde{S})) + O(\frac{d^2}{n}) \tag{23}$$

Using equations 22 and 23 we conclude

$$\begin{aligned}\text{var}\left(\sum_{j\in V}\sum_{k\in V}(d_{j,S}+d_{j,k}-d_{k,S})^2\mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}\right) &= \text{var}\left(\sum_{j\in V}\sum_{k\in V}(d_{j,\tilde{S}}+d_{j,k}-d_{k,\tilde{S}})^2\mathbb{1}_{j\in \tilde{S}}\mathbb{1}_{k\in \tilde{S}^c}\right) \\ &\quad + O(n^3d^2)\end{aligned} \tag{24}$$

Lemma 25

$$\text{var}\left(\sum_{j\in V}\sum_{k\in V}(d_{j,\tilde{S}}+d_{j,k}-d_{k,\tilde{S}})^2\mathbb{1}_{j\in \tilde{S}}\mathbb{1}_{k\in \tilde{S}^c}\right) = O(n^3d^3)$$

uniformly for all $p \in [0, 1]$. Using equation 24, we conclude that $\forall s \in [a, b]$ with $0 < a < b < 1$,

$$\text{var}\left(\sum_{j\in V}\sum_{k\in V}(d_{j,S}+d_{j,k}-d_{k,S})^2\mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}\right) = O(n^3d^3)$$

uniformly.

Proof The elements of $\tilde{S}$ are drawn i.i.d with probability of inclusion p. Define $\epsilon_i = \mathbb{1}_{i \in \tilde{S}}$. Then, $\epsilon_i \sim \text{Ber}(p)$ i.i.d for $1 \leq i \leq n$. We use $\boldsymbol{\epsilon}$ and $\tilde{S}$ interchangeably.

$$F(\boldsymbol{\epsilon}) := F(\tilde{S}) = \sum_{j \in V} \sum_{k \in V} h_{j,k}(\tilde{S})$$

Let $\tilde{S}_i := \tilde{S} \setminus \{i\}$ and $\Delta^i_{j,k}\left(\tilde{S}_i\right) := h_{j,k}\left(\tilde{S}_i\right) - h_{j,k}\left(\tilde{S}_i \cup \{i\}\right)$. Since entries of the vector $\boldsymbol{\epsilon}$ are independent, we use Efron-Stein method to tensorize the variance as follows:

$$\text{var}(F(\boldsymbol{\epsilon})) \leq \sum_{i=1}^{n} \mathbb{E}\text{var}_i\left(F(\boldsymbol{\epsilon})\right)$$

Where $\text{var}_i\left(F(\boldsymbol{\epsilon})\right) = \text{var}\left(F(\boldsymbol{\epsilon})|\epsilon_{\sim i}\right)$ Now, when $\boldsymbol{\epsilon}_{\sim i}$ is fixed, $F(\boldsymbol{\epsilon})$ can take two values. Therefore,

$$\text{var}_i(F(\boldsymbol{\epsilon})) = p(1-p)\left(F(\tilde{S}_i) - F(\tilde{S}_i \cup \{i\})\right)^2 = p(1-p)\left(\sum_{j,k} \Delta^i_{j,k}(\tilde{S}_i)\right)^2 \tag{25}$$

By Cauchy-Schwarz inequality,

$$\sqrt{\mathbb{E}\left(\sum_{j,k} \Delta^i_{j,k}(\tilde{S}_i)\right)^2} \leq \sum_{j,k} \sqrt{\mathbb{E}\left(\Delta^i_{j,k}(\tilde{S}_i)\right)^2} \tag{26}$$

Clearly, $\Delta^i_{j,k}(\tilde{S}_i) \neq 0$ only if one of the following is true:

1. $j = i$ and $k \neq i$
2. $j \neq i$ and $k = i$
3. $j \in N(i)$ and $k \notin N(i) \cup \{i\}$
4. $j \notin N(i) \cup \{i\}$ and $k \in N(i)$

For case 1, considering sub cases $k \in N(i)$ and $k \notin N(i)$, we conclude:

$$\Delta^i_{i,k}(\tilde{S}_i) = -\left(d_{i,\tilde{S}_i} - d_{k,\tilde{S}_i}\right)^2 \mathbb{1}_{k \in \tilde{S}_i^c}$$

$d_{i,\tilde{S}_i} \sim \text{Bin}(p,d)$, $d_{k,\tilde{S}_i} \sim \text{Bin}(p,d-1)$ if $k \in N(i)$ and $d_{k,\tilde{S}_i} \sim \text{Bin}(p,d)$ if $k \notin N(i)$. We use the same Minkowski inequality - McDiarmid concentration argument as in Lemma 22 to conclude that when $j = i$ and $k \neq i$

$$\mathbb{E}\left(\Delta^i_{i,k}(\tilde{S}_i)\right)^2 = O(d^2) \tag{27}$$

By a similar argument for case 2, when $j \neq i$ and $k = i$,

$$\mathbb{E}\left(\Delta^i_{j,i}(\tilde{S}_i)\right)^2 = O(d^2) \tag{28}$$

We consider case 3. Let $j \in N(i)$ and $k \notin N(i) \cup \{i\}$. Then,

$$\Delta^i_{j,k}(\tilde{S}_i) = -\left(2(d_{j,\tilde{S}_i} - d_{k,\tilde{S}_i} + d_{j,k}) - 1\right) \mathbb{1}_{j \in \tilde{S}_i} \mathbb{1}_{k \in \tilde{S}_i^c}$$

Clearly, $d_{j,S_i} \sim \text{Bin}(p, d-1)$ and $d_{k,S_i} \sim \text{Bin}(p, d)$. Using similar reasoning as case 1, we conclude that when $j \in N(i)$ and $k \notin N(i) \cup \{i\}$

$$\mathbb{E}\left(\Delta^i_{j,k}(\tilde{S}_i)\right)^2 = O(d) \tag{29}$$

We can repeat a similar argument for case 4 to conclude that when $j \notin N(i) \cup \{i\}$ and $k \in N(i)$,

$$\mathbb{E}\left(\Delta^i_{j,k}(\tilde{S}_i)\right)^2 = O(d) \tag{30}$$

All the $O()$ in the bounds above are uniform for $p \in [0, 1]$. There are at most $2n$ pairs j, k which satisfy cases 1 or 2. There are at most $2nd$ pairs which satisfy cases 3 or 4. Therefore, using equations (26) (27) (28) (29) (30)

$$\sqrt{\mathbb{E}\left(\sum_{j,k} \Delta^i_{j,k}(\tilde{S}_i)\right)^2} = 2nO(d) + 2ndO(\sqrt{d}) = O(nd^{\frac{3}{2}})$$

Therefore, we conclude from equation (25) that for every $i \in V$

$$\mathbb{E}(\text{var}_i(F(\boldsymbol{\epsilon}))) = O(n^2 d^3)$$

By Efron-Stein method, we conclude that

$$\text{var}(F(\boldsymbol{\epsilon})) = O(n^3 d^3)$$

■

We bound the second term in Theorem 14

Lemma 26 *Let $s \in (\delta, 1-\delta)$ with $0 < \delta < \frac{1}{2}$.*

$$\frac{\sqrt{\text{var}(\mathbb{E}((Y' - Y)^2 | S))}}{\sqrt{2\pi}\lambda} = O\left(\sqrt{\frac{d}{n}}\right) \tag{31}$$

The bound above holds uniformly for $s \in (\delta, 1-\delta)$.

Proof Using Lemma 25 in equation (21) and using the fact that $\sigma^2 = \Theta(nd)$ for all $s \in [\delta, 1-\delta]$ uniformly, we conclude

$$\frac{\sqrt{\text{var}(\mathbb{E}((Y' - Y)^2 | S))}}{\sqrt{2\pi}\lambda} = O\left(\sqrt{\frac{d}{n}}\right) \tag{32}$$

■

Proof [Proof of Theorem 12] We use Lemmas 26 and 24 along with Theorem 14 to show that

$$d_{\mathsf{W}}(\mathcal{L}(Y), \mathcal{N}(0,1)) \leq C\sqrt{\frac{d}{n}}$$

We conclude the bound for the Kolmogorov metric using the fact that when one of the arguments has the standard normal distribution, $d_{\mathsf{KS}} \leq C\sqrt{d_{\mathsf{W}}}$ for some absolute constant C. ■

Appendix C. Proof of Concentration of Quadratic Forms

We continue here from the end of Section 6. We refer to Chatterjee (2007) for details of the exchangeable pairs method for concentration inequalities and theorem 2.3 in Boucheron et al. (2013) for properties of sub-gamma distributions.

We begin with the Stein pair (S, S') defined in Section 5 with $|S| = l$ and $0 \leq l \leq n$. Following the notation in Chatterjee (2007), we take $F(S, S') := T(S) - T(S')$. Then,

$$f(S) := \mathbb{E}\left[F(S,S')|S\right] = \lambda(T - \mathbb{E}(T))$$

and

$$\begin{aligned}
\Delta(S) &:= \frac{1}{2}\mathbb{E}\left[(f(S) - f(S'))F(S,S')|S\right] \\
&= \frac{\lambda}{2}\mathbb{E}\left[(T-T')^2|S\right] \\
&= \frac{4\lambda}{n^2}\sum_{j\in V}\sum_{k\in V}(d_{j,S} - d_{k,S} + d_{j,k})^2 \mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c} \\
&:= \frac{4\lambda}{n^2}\sum_{j\in V}\sum_{k\in V} g_{j,k}^2(S)\mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c} \qquad (33)
\end{aligned}$$

From Theorem 1.5 in Chatterjee (2007),

$$\begin{aligned}
\mathbb{E}((f(S))^{2q}) &\leq (2q-1)^q \mathbb{E}(\Delta(S)^q) \\
\implies \mathbb{E}(T - \mathbb{E}(T))^{2q} &\leq \left(\frac{2q-1}{\lambda^2}\right)^q \mathbb{E}(\Delta(S)^q) \\
&= 4^q\left(\frac{2q-1}{\lambda}\right)^q \mathbb{E}\left(\frac{1}{n^2}\sum_{j\in V}\sum_{k\in V} g_{j,k}^2(S)\mathbb{1}_{j\in S}\mathbb{1}_{k\in S^c}\right)^q \\
&\leq 4^q\left(\frac{2q-1}{\lambda}\right)^q \frac{1}{n^2}\sum_{j\in V}\sum_{k\in V}\mathbb{E}\left[g_{j,k}(S)^{2q}\right] \\
&\leq 2.4^q.\left(\frac{2q-1}{\lambda}\right)^q q![2d(1+o(1))]^q \\
&\leq 2.(2q)!.\left(\sqrt{4nd(1+o(1))}\right)^{2q} \qquad (34)
\end{aligned}$$

Where we used Jensen's inequality in the third step and Equation (19) in the fourth step.

Following the proof of Theorem 2.3 in Boucheron et al. (2013), we conclude that for every γ such that $2|\gamma|\sqrt{4nd(1+o(1))} < 1$,

$$\log \mathbb{E} \exp \gamma(T - \mathbb{E}T) \leq \frac{32nd\gamma^2(1+o(1))}{1 - 16nd\gamma^2(1+o(1))}$$

Which is the required result in Equation (5) This follows from a simple power series argument.

Appendix D. Proof of Theorem 11

We use the notation established in Section 4.

Proof [Proof of Theorem 11] Consider the first case: $\beta_n^{\mathrm{dreg}}\sqrt{nd} = \Theta(\sigma_n \beta_n^{\mathrm{dreg}}) \geq L_n \to \infty$. We first fix the parameters β_n^{dreg}, β^{CW}, h^{dreg} and h^{CW}. p_n and q_n satisfy Conditions **C1-C4** of Theorem 9 as shown in Section 4.

We invoke Theorem 9 to conclude that for some choice of T_n depending only on τ_n, L_n and S_n, the canonical test $\mathcal{T}^{\mathsf{can}}(T_n, g_n)$ can distinguish between p_n (with given parameters β^{CW}, h^{CW}, β_n^{dreg} and h^{dreg}) and q_n with a single sample with probability of error

$$p_{\mathsf{error}} \leq f(L_n, \alpha_n, \tau_n) \to 0$$

We conclude from Remark 10 that the canonical tests $\mathcal{T}(T_n, g_n)$ and $\mathcal{T}(T_n, \kappa_{\mathsf{Ising}})$ are the same. Therefore, the test $\mathcal{T}(T_n, \kappa_{\mathsf{Ising}})$ has the same success probability for the same choice of T_n. κ_{Ising} doesn't depend on the unknown parameters. The parameters S_n, τ_n and α_n depend only on β_{max} and h_{max}. Therefore, given L_n, T_n can be chosen without the knowledge of the unknown parameters. The probability of error tends to 0 uniformly for every choice of the unknown parameters. Hence, we conclude that the canonical test $\mathcal{T}(T_n, \kappa_{\mathsf{Ising}})$ succeeds with high probability for any choice of the unknown parameters.

We now consider the second case: $\beta_n^{\mathrm{dreg}}\sqrt{nd} = \Theta(\sigma_n \beta_n^{\mathrm{dreg}}) \leq \frac{1}{L_n} \to 0$. It is sufficient to prove that for a specific sequence $(\beta_n^{\mathrm{CW}}, \beta_n^{\mathrm{dreg}})$ and external fields $(h^{\mathrm{dreg}}, h^{\mathrm{CW}})$,

$$d_{\mathsf{TV}}(p_n, q_n) \to 0\,.$$

We take $\beta^{\mathrm{CW}} = \frac{nd\beta_n^{\mathrm{dreg}}}{n-1}$ and $h^{\mathrm{dreg}} = h^{\mathrm{CW}}$. A simple calculation using Lemma 16 we show that

$$e_m(g) = \mathbb{E}g(X_m) = 0\,.$$

Using Equation (5), we conclude

$$g(X_m) - e_m(g) = -2\left(T\left(S_m\right) - \mathbb{E}\left[T\left(S_m\right)\right]\right)$$

Where $S_m = S(X_m)$. Clearly, for n large enough, we have $4\beta_n^{\text{dreg}}\sqrt{4nd(1+o(1))} < 1$. We shall denote $T_m \stackrel{d}{:=} T(S_m)$

$$\begin{aligned}
\mathbb{E}_p\left[e^{\beta_n^{\text{dreg}} g}\right] &= \sum_{m_0 \in M_n} p\left(m(x) = m_0\right) \mathbb{E}\left[e^{\beta_n^{\text{dreg}} g(X_m)}\right] \\
&= \sum_{m_0 \in M_n} p\left(m(x) = m_0\right) \mathbb{E}\left[e^{-2\beta_n^{\text{dreg}}(T_m - \mathbb{E}[T_m])}\right] \\
&\leq e^{\frac{128nd\left(\beta_n^{\text{dreg}}\right)^2(1+o(1))}{1-64nd\left(\beta_n^{\text{dreg}}\right)^2(1+o(1))}} \\
&\leq e^{\frac{128nd\left(\beta_n^{\text{dreg}}\right)^2(1+o(1))}{1-8\beta_n^{\text{dreg}}\sqrt{nd(1+o(1))}}} \qquad (35)
\end{aligned}$$

Where we have used Equation (5) in the second to last step. To bound the variance $\text{var}_p g$, we note that $\mathbb{E}_p(g) = 0$ and $\mathbb{E}g(X_m) = 0$. Therefore,

$$\text{var}_p g = \mathbb{E}_p g^2 = \sum_{m_0 \in M_n} p\left(m(x) = m_0\right) \mathbb{E}g^2(X_{m_0}) = \sum_{m_0 \in M_n} p\left(m(x) = m_0\right) \sigma_{m_0}^2 \qquad (36)$$

Clearly, $\beta^{\text{CW}} \to 0$. Therefore, for n large enough, $\beta^{\text{CW}} < 1$. We refer to the large deviations result for Curie-Weiss model given in Ellis (2007) to conclude that $p(|m(x)| > m_{\mathsf{max}}(h_{\max})) \leq C_1 e^{-nC}$ for some positive constants C_1 and C. Clearly,

$$\sigma_{m_0} \leq Cn^2 d^2$$

for every m_0. Invoking Theorem 12, we conlude that whenever $m_0 \leq m_{\mathsf{max}}(h_{\max})$, $\sigma_{m_0} \leq D\sqrt{nd}$ for some constant D. Plugging these results in Equation (36), we conclude that

$$\text{var}_p[g] \leq Dnd + C_1 n^2 d^2 e^{-nC} = O(nd) \qquad (37)$$

Therefore, using Equations (37) and (35) we conclude that for this particular choice of β^{CW} and h^{CW}, p and g satisfy Condition **C5**. Therefore, invoking the second part of Theorem 9, we conclude that

$$d_{\mathsf{TV}}(p_n, q_n) \to 0\,.$$

which proves our result. ■

Appendix E. Proof of Theorem 15

Consider Erdős-Rényi model over n vertices. We let $N = \binom{n}{2}$, the maximum number of edges. Consider $\mathcal{X} := \{0,1\}^N$. We index elements of $x \in \mathcal{X}$ by tuples $e = (i,j)$ such that $i, j \in [n]$ and $i < j$. We can represent any simple graph $G = (V, E)$ over n vertices by an element $x(G) \in \mathcal{X}$ such that $x(G)_e = 1$ iff $e \in E$. Henceforth we use 'eth component of x' and 'edge e' interchangeably.

Consider an $N \times N$ symmetric matrix H such that $H_{e,f} \in \{0,1\}$ and $H_{e,f} = 1$ iff e and f have a common vertex. Clearly, H is the adjacency matrix of a $d = 2(n-2)$ regular graph

over N vertices. We partition $\mathcal{X}$ into Hamming spheres. For $m \in \{0, 1, \ldots, N\} =: M_n$, define $A_m \subset \mathcal{X}$ to be the set of simple graphs over n vertices with exactly m edges. Here we define the function $E(x) := m(x)$ to be the number of edges in the graph associated with $x \in \mathcal{X}$. Clearly, the number of V graphs ($\bigvee$) which are subgraphs of the graph represented by x is

$$V(x) = \frac{1}{2} \sum_{e,f} x_e x_f H_{ef} = \frac{1}{2} x^\intercal H x$$

Let $\mathbb{1}$ be the all one vector. A simple calculation using the fact that H is a regular matrix shows that:

$$\begin{aligned}(2x - \mathbb{1})^\intercal H (2x - \mathbb{1}) &= 4x^\intercal H x - 4\mathbb{1}^\intercal H x + \mathbb{1}^\intercal H \mathbb{1} \\ &= 8V(x) - 8(n-2)E(x) + 2(n-2)N \end{aligned} \tag{38}$$

Clearly, $2x - \mathbb{1} \in \{-1, 1\}^N$ and $|\{e : 2x_e - 1 = 1\}| = E(x)$.

Let μ be the probability measure associated with $G(n, p_n)$ ($p(\cdot)$ in the notation of Theorem 9) such that $\delta < p_n < 1 - \delta$ for some constant $\delta > 0$. We shall drop the subscript of p_n for the sake of clarity. Since E is a binomial random variable, we can easily show using McDiarmid's inequality that:

$$\mu\left(E(x) \in \left[\frac{\delta}{2}N, \left(1 - \frac{\delta}{2}\right)N\right]\right) \geq 1 - 2e^{-c(\delta)N} =: 1 - \alpha_n$$

For some constant $c(\delta) > 0$. Therefore, we let

$$S_n = M_n \cap \left[\frac{\delta}{2}N, \left(1 - \frac{\delta}{2}\right)N\right]$$

We let $g(x) = \left(n\left(\frac{\beta_1 - \frac{1}{2}\log\frac{p}{1-p}}{\beta_2}\right)E(x) + V(x)\right)$.

Remark 27 *Let $\mu^{(m)}$ be the conditional distribution $\mu(\cdot|E(x) = m)$ ($p^{(m)}$ in Section 3). Similar to Remark 10 about Ising models, we note that $\mu^{(m)}$ is the uniform distribution over the graphs with fixed number of edges m irrespective of the value of p (i.e, uniform distribution over the set A_m). Proceeding as in Remark 10, let $\hat{X}$ be the given sample and $\hat{m} = m(\hat{X})$. We can decide whether $\hat{m} \in S_n$ without the knowledge of the unknown parameters and since*

$$g(\hat{X}) - e_{\hat{m}}(g) = V(\hat{X}) - e_{\hat{m}} V(x)$$

and

$$\mathrm{var}_{\hat{m}}(g) = \mathrm{var}_{\hat{m}}(V)$$

we can decide whether $\frac{g(\hat{X}) - e_{\hat{m}}(g)}{\sigma_{\hat{m}}(g)} \geq T$ without the knowledge of the unknown parameters. We conclude that $\mathcal{T}(T_n, g_n) = \mathcal{T}(T_n, V)$

Let $X_m \sim \mu^{(m)}$. Therefore, whenever $x \in A_m$ for $m \in \left[\frac{\delta}{2}N, \left(1-\frac{\delta}{2}\right)N\right]$, $2X_m - 1$ satisfies the hypothesis for Theorem 12. Using Equation (38) and the fact that $E(X_m) = m$ is a constant a.s. we conclude that:

$$\operatorname{var}\left(g(X_m)\right) =: \sigma_m^2 = \Theta(n^3)$$

and

$$d_{\mathsf{KS}}\left(\mathcal{L}\left(\frac{g(X_m)-\mathbb{E}g(X_m)}{\sigma_m}\right), \mathcal{N}(0,1)\right) \leq C\sqrt[4]{\frac{1}{n}} =: \tau_n$$

Where we have used the fact that degree $d = \Theta(n)$ and number of rows/columns is $N = \Theta(n^2)$. All the $\Theta(\cdot)$ and bounds hold uniformly for all $m \in S_n$. Let $\beta_n := \frac{2\beta_2}{n}$. We take $\nu(x) = \mu(x)\frac{e^{\beta_n g(x)}}{\mathbb{E}_\mu e^{\beta_n g}} =: \mathsf{ERGM}(\beta_1, \beta_2)$.

To prove Theorem 15, we will need the following Lemma, where we get very small variance of a quadratic function by picking the right coefficient.

Lemma 28 *Let $p \in [0,1]$ be arbitrary. Then there exists an absolute constant c such that whenever $\frac{2\beta_2}{n} =: \beta_n < \frac{c(1-o(1))}{n^3}$ for some absolute constant c then for some choice of β_1 as a function of p and β_2, the following hold:*

1. $\operatorname{var}_\mu(g) = O(n^3) = O(N^{\frac{3}{2}})$
2. $\log \mathbb{E}_\mu\left[e^{\beta_n(g-\mathbb{E}_\mu g)}\right] \leq \frac{Cn^3\beta_n^2(1+o(1))}{1-D\beta_n\sqrt{n^3(1+o(1))}} + \frac{C_1\beta_n^2 n^2(1+o(1))}{1-|\beta_n|D_1 n(1+o(1))}$

Where C, C_1, D and D_1 are absolute constants

We defer the proof of this Lemma to Appendix F.

We proceed with the proof of Theorem 15.

Proof [Proof of Theorem 15] Since $\beta_n := \frac{2\beta_2}{n}$, it is sufficient to consider the regimes: $\beta_n n^{3/2} \to \infty$ and $\beta_n n^{3/2} \to 0$. By Remark 27, $\mathcal{T}(T_n, g_n) = \mathcal{T}(T_n, V)$, which doesn't depend on parameters β_1, p or β_2. The proof of the first part is similar to that in Theorem 11 and it follows from the discussion above and Theorem 9.

We now assume $\beta_2 \leq \frac{1}{L_n}\frac{1}{\sqrt{n}}$ and fix $p \in [\delta, 1-\delta]$. To prove the second part it is sufficient to show one distribution in H_0 is near to one distribution in H_1 in the total variation sense. We will β_1 as a function of p and β_2 such that $d_{\mathsf{TV}}(\mu_n, \nu_n) \to 0$.

Using the notation of Theorem 9, we have $\sigma_n = \Theta(n^{3/2})$. By Lemma 28, we conclude that Condition **C5** for Theorem 9 holds for some choice of β_1 and hence $d_{\mathsf{TV}}(\mu_n, \nu_n) \to 0$ ■

Appendix F. Proof of Super Concentration

Lemma 29 *If $E \sim \operatorname{Bin}(N, p)$, then,*

$$\mathbb{E}(E - Np)^{2q} \leq q! C^q N^q$$

For some absolute constant C independent of N

Proof By McDiarmid's theorem,

$$\mathbb{P}\left(|E - Np| > t\right) \leq 2e^{-\frac{2t^2}{N}}$$

We use the equivalence of moment inequalities and sub-gaussian concentrations (refer Theorem 2.1 in Boucheron et al. (2013)) to conclude the result. ■

Let X be a random vector taking values in $\{0,1\}^N$ such that its coordinates are i.i.d. $\mathrm{Ber}(p)$. Consider the function $h(X) = E^2(X) - (2pN + 1 - 2p)E(X)$. As we shall see, this choice of coefficients is special since it corresponds to a very small variance.

Obtain the random variable X' as follows: Choose n random index $I \sim \mathrm{unif}([N])$. $X_i = X_i'$ whenever $i \neq I$ and $X_I' \sim \mathrm{Ber}(p)$ independent of X. Clearly, (X, X') is an exchangeable pair.

Lemma 30

1. *$(h(X), h(X'))$ is an η-Stein pair with respect to $\mathcal{F}(X)$ where $\eta = \frac{2}{N}$.*
2. $\mathbb{E}h(X) = -p^2N(N-1)$

Proof We shorten $E(X)$ to E. Let $a := -(2pN + 1 - 2p)$

$$\begin{aligned}
\mathbb{E}\left[h(X') - h(X)|X\right] &= p\left(1 - \tfrac{E}{N}\right)\left((E+1)^2 + a(E+1) - E^2 - aE\right) \\
&\quad + \tfrac{E}{N}(1-p)\left((E-1)^2 + a(E-1) - E^2 - aE\right) \\
&= -\tfrac{2}{N}E^2 + \tfrac{E}{N}\left[2pN - 2p + 1 - a\right] + p(1+a) \\
&= -\tfrac{2}{N}h(X) - 2p^2(N-1)
\end{aligned}$$

Using the definition of a Stein pair and the fact that $\mathbb{E}h(X) = \mathbb{E}h(X')$, we conclude the result. ■

We proceed in the same way as Section 6.

$$F(X, X') := h(X) - h(X')$$
$$f(X) := \mathbb{E}\left[F(X, X')|X\right] = \eta(h(X) - \mathbb{E}h(X))$$

$$\begin{aligned}
\Delta(X) &:= \frac{1}{2}\mathbb{E}\left[(f(X) - f(X'))F(X, X')|X\right] \\
&= \frac{\eta}{2}\mathbb{E}\left[\left(h(X) - h(X')\right)^2|X\right] \\
&= \frac{\eta}{2}\left[p\left(1 - \tfrac{E}{N}\right)(2E - 2pN + 2p)^2 + \tfrac{E}{N}(1-p)(2E - 2pN + 2p - 2)^2\right] \\
&\leq 2\eta\left[p(E - pN + p)^2 + (1-p)(E - pN + p - 1)^2\right] \qquad (39)
\end{aligned}$$

From Theorem 1.5 in Chatterjee (2007), for every $q \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}((f(X))^{2q}) &\leq (2q-1)^q \mathbb{E}(\Delta(X)^q) \\
\implies \mathbb{E}(h(X) - \mathbb{E}h(X))^{2q} &\leq \left(\frac{2q-1}{\eta^2}\right)^q \mathbb{E}(\Delta(X)^q) \\
&= 2^q \left(\frac{2q-1}{\eta}\right)^q \mathbb{E}\left[p(E-pN+p)^2 + (1-p)(E-pN+p-1)^2\right]^q \\
&\leq 2^q \left(\frac{2q-1}{\eta}\right)^q \mathbb{E}\left[p(E-pN+p)^{2q} + (1-p)(E-pN+p-1)^{2q}\right] \\
&= 8^q \left(\frac{2q-1}{\eta}\right)^q \mathbb{E}\left[p\left(\tfrac{E-pN}{2} + \tfrac{p}{2}\right)^{2q} + (1-p)\left(\tfrac{E-pN}{2} + \tfrac{p-1}{2}\right)^{2q}\right] \\
&\leq \frac{8^q}{2}\left(\frac{2q-1}{\eta}\right)^q \left(\mathbb{E}\left[(E-pN)^{2q}\right] + p^{2q+1} + (1-p)^{2q+1}\right) \\
&\leq C^{2q}(2q)!N^{2q} \qquad (40)
\end{aligned}$$

Where we have used Equation (39) in the second step, Jensen's inequality for the convex function $\phi(x) = |x|^q$ in the third step, Jensen's inequality again for the function $\phi(x) = |x|^{2q}$ and Lemma 29 in the final step.

Following the proof of Theorem 2.3 in Boucheron et al. (2013), we conclude that for every γ such that $|\gamma|CN < 1$,

$$\mathbb{E}e^{\gamma[h(X)-\mathbb{E}h(X)]} \leq 2\frac{C^2\gamma^2N^2}{1-|\gamma|CN} \qquad (41)$$

Where C is an absolute constant.

We use Equation (5) along with Equation (38) to conclude that for every $m \in \{0, \ldots, N\}$ and some absolute constants C and D,

$$\log \mathbb{E}e^{\beta_n g(X_m)} \leq \beta_n \mathbb{E}g(X_m) + \frac{Cn^3\beta_n^2(1+o(1))}{1 - D\beta_n\sqrt{n^3(1+o(1))}} \qquad (42)$$

Proof [Proof of Lemma 28] The bound on variance follows from the bound on MGF shown in the second part of the theorem after an application of Theorem 2.3 in Boucheron et al. (2013). Therefore, it is sufficient to show the bound on the MGF.

$$\begin{aligned}
\mathbb{E}g(X_m) &= \mathbb{E}[g(X)|E(X) = m] \\
&= \mathbb{E}V(X_m) + n\left(\frac{\beta_1 - \frac{1}{2}\log\frac{p}{1-p}}{\beta_2}\right)m \\
&= \frac{n-2}{N-1}m^2 + n\left(\frac{\beta_1 - \frac{1}{2}\log\frac{p}{1-p}}{\beta_2}\right)m - \frac{n-2}{N-1}m
\end{aligned}$$

Therefore, we can choose β_1 a function of p and β_2 such that :

$$\mathbb{E}[g(X)|E(X)] = \frac{n-2}{N-1}h(X) = \frac{n-2}{N-1}\left(E^2 - (2pN+1-2p)E\right)$$

Therefore,

$$\begin{aligned}\mathbb{E}_\mu\left[e^{\beta_n(g-\mathbb{E}_\mu g)}\right] &= \mathbb{E}_\mu\left[\mathbb{E}\left[e^{\beta_n(g(X_m)-\mathbb{E}_\mu g)}\middle| E(X)=m\right]\right] \\ &\leq e^{\frac{Cn^3\beta_n^2(1+o(1))}{1-D\beta_n\sqrt{n^3(1+o(1))}}}\mathbb{E}_\mu\left[e^{\beta_n\frac{n-2}{N-1}(h(X)-\mathbb{E}_\mu h)}\right] \\ &\leq e^{\frac{Cn^3\beta_n^2(1+o(1))}{1-D\beta_n\sqrt{n^3(1+o(1))}}}e^{2\frac{C^2\beta_n^2n^2(1+o(1))}{1-|\beta_n|Cn(1+o(1))}}\end{aligned}$$

Here, we have used Equation (42) and the fact that for this particular choice of β_1, $\mathbb{E}_\mu h = \frac{n-2}{N-1}\mathbb{E}_\mu g$. In the third step we have used Equation (41). ■