Electron. J. Probab. **26** (2021), article no. 139, 1–36.
ISSN: 1083-6489 https://doi.org/10.1214/21-EJP708

Higher-order fluctuations in dense random graph models*

Gursharn Kaur† Adrian Röllin‡

Abstract

Our main results are quantitative bounds in the multivariate normal approximation of *centred* subgraph counts in random graphs generated by a general graphon and independent vertex labels. We are interested in these statistics because they are key to understanding fluctuations of regular subgraph counts — a cornerstone of dense graph limit theory. We also identify the resulting limiting Gaussian stochastic measures by means of the theory of generalised U-statistics and Gaussian Hilbert spaces, which we think is a suitable framework to describe and understand higher-order fluctuations in dense random graph models. With this article, we believe we answer the question "What is the central limit theorem of dense graph limit theory?". We complement the theory with some statistical applications to illustrate the use of centred subgraph counts in network modelling.

Keywords: graphon; centered subgraph counts; central limit theorem; Gaussian Hilbert spaces.
MSC2020 subject classifications: Primary 60F05, Secondary 05C80.
Submitted to EJP on September 14, 2020, final version accepted on September 21, 2021.
Supersedes arXiv:2006.15805.

1 Introduction

Since the seminal paper of Lovász and Szegedy (2006) on dense graph limit theory, a considerable amount of literature devoted to this topic has been published. A book-length treatment was given by Lovász (2012), and the theory has been extended to related models, such as sparse graphs by Bollobás and Riordan (2009), Borgs, Chayes, Cohn, and Zhao (2014a,b), Caron and Fox (2017), Borgs, Chayes, Cohn, and Holden (2017) and others, multi-graphs by Ráth (2012) and Ráth and Szakács (2012), graphon-valued stochastic processes by Athreya, den Hollander, and Röllin (2019), and permutations by Hoppen, Kohayakawa, Moreira, and Sampaio (2011) to name a few.

Much of the literature is concerned with what could generally be referred to as *laws of large numbers*, where the main interest lies in describing the limiting objects upon

*This project was supported by NUS Research Grant R-155-000-198-114.
†National University of Singapore, Singapore. E-mail: gursharn.kaur24@gmail.com
‡National University of Singapore, Singapore. E-mail: adrian.roellin@nus.edu.sg

appropriate scaling as some number n that captures the size of the model — for example, the number of vertices of a graph — tends to infinity. In many applications, the limiting objects are deterministic, since the randomness in the model "averages out", like in the case of the fraction of heads in a sequence of independent fair coin tosses. And if the limiting objects are random, then typically because of a phenomenon related to *de Finetti's Theorem* in the sense that the randomness left in the limit can be thought of as being distinct from the randomness describing the fine details of the model. In the case of dense graph limit theory, this phenomenon is captured by the Aldous-Hoover theory of infinite exchangeable arrays; see Diaconis and Janson (2008).

In analogy to the classical *Law of Large Numbers* for sums of independent random variables, it is natural to ask about fluctuations *around* the limits, which in the classical case is captured by the *Central Limit Theorem*. This is of profound importance, since statistical inference is based on exactly this kind of fluctuations. But despite the large literature on dense graph limit theory, we are not aware of any attempts made to develop a higher-order fluctuation theory for random graph models, neither in the dense nor sparse regime.

There have been some recent efforts to understand the subgraphs counts in the context of graphons and dense graph limit theory. Hladký, Pelekis and Šileikis (2019) analysed the limiting distributions of r-clique counts of a random graph obtained through sampling from a graphon (which is our model $\mathbb{G}(n,\kappa)$ below), and Chatterjee and Bhattacharya (2021) generalised the results to arbitrary subgraphs. Maugis (2020) analysed localised versions, where the counts are not global, but only over one specific vertex. Their results yield in essence that the scaling and limiting distribution depends on specific properties of the graphon, and this is intimately related to the work of Janson and Nowicki (1991) on U-statistics. What makes subgraph counts problematic as test statistics is the fact that it is not immediately clear what is actually being tested (in other words, what aspects of the model the dominating fluctuations represent), and how different subgraph counts are related to each other, which is crucial when performing multiple test over different subgraph counts.

What we propose in this article is not to use subgraph counts as test statistics directly, but use more fundamental statistics — centred subgraph counts — which themselves completely determine the fluctuations of subgraph counts, which are orthogonal to each other, and which are jointly Gaussian in the limit with a straightforward covariance structure. The latter in particular allows for a proper correction when performing multiple tests. We give a rather complete description of these statistics in the dense case for models where vertices have independent labels, and conditionally on the vertex labels, edges are sampled independently of each other with probabilities given by a graphon. This model is the workhorse of dense graph limit theory, although in this article, we generalise this to sampling schemes where vertex labels need not be identically distributed. We believe the latter is an important contribution and covers the important case where vertex labels are fixed and arranged on an equally spaced lattice.

As mentioned, the key to understanding all fluctuations is to analyse *centred* subgraph counts rather than regular subgraph counts, and we are not the first to do so. Centred subgraph counts were studied in depth by Janson (1994), where the normal limits were shown using martingale methods, and by Janson (1997), who used the method of moments. Fang and Röllin (2015) studied statistics similar to centred subgraph counts to construct a test whether a given graph is compatible with a constant graphon, and Bubeck, Ding, Eldan, and Rácz (2016) used centred triangle counts to construct a test for dimensionality in geometric random graphs; see also Gao and Lafferty (2017a,b).

As we will argue in the next section, the mathematical framework of generalised U-statistics can be used to describe the fluctuations in dense graph sequences. Generalised

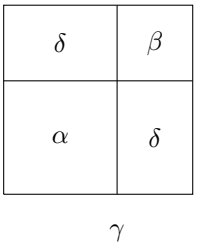

Figure 1: The 2×2 graphon κ defined in (1.1).

U-statistics were introduced by Janson and Nowicki (1991) to understand fluctuations of subgraph counts in the Erdős-Rényi random graph and related models, and a more comprehensive treatise was given by Janson (1994, 1997). In particular, using the framework of *Gaussian Hilbert spaces*, Janson (1997) was able to describe the Gaussian limiting objects arising from generalised U-statistics, although his description is rather abstract and not easily interpretable in the context of dense graph limit theory.

Our contribution is to modify the approach of Janson (1997) in such a way that it becomes clearer what the limiting Gaussian Hilbert spaces are and such that it applies to non-identically distributed vertex labels, and we complement the theory with a multivariate normal approximation theorem for smooth and non-smooth test functions, which is based on Stein's method. Incidentally, none of the existing approximation theorems in the literature seem to be applicable to the present situation due to the fact that the summands in our test statistics are uncorrelated, a case that has drawn surprisingly little attention in the literature so far. Although subgraph counts can be handled using Stein's method, as was shown by Barbour, Karoński, and Ruciński (1989) for smooth metrics, by Röllin and Ross (2015) for total variation and local limit metrics, by Röllin (2017), Krokowski, Reichenbachs, and Thäle (2017) and Privault and Serafin (2020) for the Kolmogorov metric, centred subgraph counts, which are sums of uncorrelated but not independent random variables, cannot be handled with these approaches. The only result in this direction we are aware of is that of Fang and Röllin (2015), who considered bi-variate normal approximation for related sums of uncorrelated random variables in the case of constant graphons.

1.1 The basic decomposition of subgraph counts — an example

Before elaborating on the general theory, we first illustrate what a decomposition of a subgraph into orthogonal components looks like in the simple case of a 2×2 block graphon, also called stochastic block model. This model is general enough to illustrate the main points, but also simple enough to work out the details, at least for simple subgraphs.

First, fix constants $\alpha, \beta, \delta \in [0,1]$ and $\gamma \in (0,1)$. Then let $\kappa : [0,1]^2 \to [0,1]$ be the graphon defined as

$$\kappa(x,y) = \begin{cases} \alpha & \text{if } x,y \leqslant \gamma, \\ \delta & \text{if } x \leqslant \gamma < y \text{ or } y \leqslant \gamma < x, \\ \beta & \text{if } x,y > \gamma, \end{cases} \qquad \text{for } x,y \in [0,1]. \tag{1.1}$$

This graphon is illustrated in Figure 1 and represents a graph with two communities with connection probability α and β within the respective communities, and δ across the two communities.

We now generate a random graph G_n on n vertices in the usual way. Let $U_1, \dots, U_n$

be independent random variables distributed uniformly on $[0,1]$, and conditionally on U_i and U_j, connect vertices i and j with probability $\kappa(U_i,U_j)$, independently of all else. It is clear that the probability of a vertex belonging to the first community is γ and the probability it belongs to the second community is $1-\gamma$. We denote by $Z_i = \mathrm{I}[U_i \leqslant \gamma]$ the indicator that vertex i belongs to the first community, and by Y_{ij} the indicator that i and j are connected.

Now, to start with, consider the so-called *edge density*

$$t_{\swarrow}^{\mathrm{inj}}(G_n) = \frac{1}{n(n-1)} \sum_{i_1,i_2}{}' Y_{i_1 i_2},$$

where the sum ranges over all vertices and the prime in the double sum indicates exclusion of the diagonal cases $i_1 = i_2$ as usual. With $\hat{Y}_{ij} = Y_{ij} - \kappa(U_i,U_j)$, it is straightforward to deduce the decomposition

$$t_{\swarrow}^{\mathrm{inj}}(G_n) = \frac{1}{n(n-1)} \sum_{i_1,i_2}{}' \hat{Y}_{i_1 i_2} + \frac{1}{n(n-1)} \sum_{i_1,i_2}{}' (\kappa(U_{i_1},U_{i_2}) - \bar{\kappa}) + \bar{\kappa}, \tag{1.2}$$

where $\bar{\kappa} = \mathbb{E}\kappa(U_1,U_2) = \alpha\gamma^2 + 2\delta\gamma(1-\gamma) + \beta(1-\gamma)^2$. Now, the second sum in (1.2) itself is a U-statistic, making further decomposition necessary. To this end, we write

$$\kappa(U_i,U_j) - \bar{\kappa} = \rho_1(\hat{Z}_i + \hat{Z}_j) + \rho_2 \hat{Z}_i \hat{Z}_j, \tag{1.3}$$

where

$$\hat{Z}_i = Z_i - \gamma, \qquad \rho_1 = \alpha\gamma - \beta(1-\gamma) + (1-2\gamma)\delta, \qquad \rho_2 = \alpha + \beta - 2\delta. \tag{1.4}$$

Using (1.3) on the second sum in (1.2) and a tedious exercise of adding, subtracting and rearranging terms, as well as observing that $\hat{Z}_i^2 = (1-2\gamma)\hat{Z}_i + \gamma(1-\gamma)$, we can write

$$\sum_{i_1,i_2}{}' (\kappa(U_{i_1},U_{i_2}) - \bar{\kappa}) = (\beta - \alpha + 2n\rho_1) \sum_i \hat{Z}_i + \rho_2 \left[\left(\sum_i \hat{Z}_i \right)^2 - n\gamma(1-\gamma) \right].$$

Thus, we arrive at a complete decomposition of the form

$$t_{\swarrow}^{\mathrm{inj}}(G_n) = \bar{\kappa} + \frac{2n^{1/2}\rho_1 W}{n-1} + \frac{\rho_2 \left(W^2 - \gamma(1-\gamma)\right)}{n-1} + \frac{2^{1/2} V_{\swarrow,1}}{n^{1/2}(n-1)^{1/2}} + \frac{(\beta-\alpha)W}{n^{1/2}(n-1)}, \tag{1.5}$$

where

$$W = n^{-1/2} \sum_i \hat{Z}_i, \qquad V_{\swarrow,1} = \tbinom{n}{2}^{-1/2} \sum_{i_1<i_2} \hat{Y}_{i_1 i_2}. \tag{1.6}$$

There are multiple reasons why this decomposition is useful. First, both W and $V_{\swarrow,1}$ are centred and uncorrelated random variables, and moreover, they themselves are sums of uncorrelated random variables; this is true if even the U_i are not identically distributed, so long as they are independent. Hence, the variance and covariance structure is straightforward to calculate, and with the normalisations given above and assuming the U_i are identically distributed, all variances are of order 1. Second, all quantities have Gaussian limits; for W, this follows easily from the classical central limit theorem, but it is also not difficult to prove for $V_{\swarrow,1}$ using Stein's method or the method of moments. This fact also implies the limit for W^2, namely a χ_1^2-distribution. Third, it is now straightforward to read off the limiting behaviour of $t_{\swarrow}^{\mathrm{inj}}(G)$ from this decomposition (upon appropriate scaling):

1. $\rho_1 \neq 0$: The second term in (1.5) dominates and the limit is Gaussian.

2. $\rho_1 = 0$ and $\rho_2 \neq 1$: The third and fourth terms in (1.5) dominate and the limit is the weighted sum of two independent random variables, one Gaussian and the other having a centred χ^2_1-distribution.
3. $\rho_1 = 0$ and $\rho_2 = 0$: The fourth term in (1.5) dominates and the limit is Gaussian.

These convergence results were also obtained by Hladký et al. (2019) and Chatterjee and Bhattacharya (2021). Note that even in the third case, the contribution of W in (1.5) does not vanish if $\alpha \neq \beta$, although the fluctuation only contributes to a scale that is smaller than the dominating fluctuation. It is also important to recognise that this decomposition is not unique; for example, since

$$\frac{1}{n-1} = \frac{1}{n} + \frac{1}{n(n-1)},$$

fluctuations at one scale can always be slightly rescaled and change the composition of fluctuations at another scale.

This simple example already reveals the subtle nature of subgraph counts under inhomogeneous sampling schemes, even for just the edge density. It also shows the main disadvantage of modelling vertex labels U_i as random: In general, the subgraph counts are dominated by the group labels, rather than the randomness in the edges. This is usually not a desired property if testing a graph model against network data.

The case of triangles $t^{\mathrm{inj}}_{\vartriangle}(G_n) = \frac{1}{(n)_3} \sum'_{i_1,i_2,i_3} Y_{i_1 i_2} Y_{i_2 i_3} Y_{i_1 i_3}$ is much more involved and tedious to deduce, and we therefore only give the final decomposition (for multiple sums, the prime indicates exclusion of any set of indices where at least two indices coincide). We have

$$t^{\mathrm{inj}}_{\vartriangle}(G_n) = c_1 + R_{0.5} + R_{1.0} + R_{1.5} + R_{2.0} + R_{2.5}$$

where

$$\begin{aligned}
R_{0.5} &= \frac{c_2 W}{n^{1/2}},\\
R_{1.0} &= \frac{c_3(W^2 - \gamma(1-\gamma))}{n-1} + \frac{c_4 V_{\diagup\!\diagdown,4} + c_5(V_{\diagup\!\diagdown,2} + V_{\diagup\!\diagdown,3}) + c_6 V_{\diagup\!\diagdown,1}}{n^{1/2}(n-1)^{1/2}},\\
R_{1.5} &= \frac{c_7 W}{n^{1/2}(n-1)} + \frac{n^{1/2} c_8 \left(W^3 - n^{-1/2}\gamma(1-\gamma)(1-2\gamma)\right)}{(n-1)(n-2)}\\
&\quad + \frac{c_9 \left(V_{\vartriangle} + V_{\vee,1} + V_{\vee,2} + V_{\vee,3}\right)}{n^{1/2}(n-1)^{1/2}(n-2)^{1/2}} + \frac{c_{10} V_{\diagup\!\diagdown,1} W + c_{11}(V_{\diagup\!\diagdown,2} + V_{\diagup\!\diagdown,3}) W + c_{12} V_{\diagup\!\diagdown,4} W}{(n-1)^{1/2}(n-2)},\\
R_{2.0} &= \frac{c_{13} V_{\diagup\!\diagdown,1} + c_{14}(V_{\diagup\!\diagdown,2} + V_{\diagup\!\diagdown,3}) + 2c_{15} V_{\diagup\!\diagdown,4}}{n^{1/2}(n-1)^{1/2}(n-2)} + \frac{c_{16}(W^2 - \gamma(1-\gamma))}{(n-1)(n-2)},\\
R_{2.5} &= \frac{c_{17} W}{n^{1/2}(n-1)(n-2)},
\end{aligned}$$

with $V_{\diagup\!\diagdown,2}$ and W as in (1.6) and with

$$\begin{gathered}
V_{\diagup\!\diagdown,2} = \tbinom{n}{2}^{-1/2} \sum_{i<j} \hat{Z}_i \hat{Y}_{ij}, \quad V_{\diagup\!\diagdown,3} = \tbinom{n}{2}^{-1/2} \sum_{i<j} \hat{Z}_j \hat{Y}_{ij}, \quad V_{\diagup\!\diagdown,4} = \tbinom{n}{2}^{-1/2} \sum_{i<j} \hat{Z}_i \hat{Z}_j \hat{Y}_{ij},\\
V_{\vee,1} = \tbinom{n}{3}^{-1/2} \sum_{i<j<k} \kappa(U_i, U_k) \hat{Y}_{ij} \hat{Y}_{jk}, \quad V_{\vee,2} = \tbinom{n}{3}^{-1/2} \sum_{i<j<k} \kappa(U_j, U_k) \hat{Y}_{ji} \hat{Y}_{ik},\\
V_{\vee,3} = \tbinom{n}{3}^{-1/2} \sum_{i<j<k} \kappa(U_i, U_j) \hat{Y}_{ik} \hat{Y}_{kj}, \quad V_{\vartriangle} = \tbinom{n}{3}^{-1/2} \sum_{1i<j<k} \hat{Y}_{ij} \hat{Y}_{jk} \hat{Y}_{ik}
\end{gathered}$$

(1.7)

(the values of the constants c_1 to c_{17} can be found in the Appendix); these results are again consistent with Hladký et al. (2019) and Chatterjee and Bhattacharya (2021). Note also that all these quantities are again uncorrelated and themselves sums of uncorrelated random variables, and they are scaled to be of order 1. We have arranged the terms so that R_α has standard deviation of order $n^{-\alpha}$. What our main result, Theorem 3.1, says is that all the quantities arising in such a decomposition are jointly close to a multivariate normal distribution that has a straightforward covariance structure, which is why we believe they are better suited for statistical applications than subgraph counts.

Note that explicit decompositions like the ones presented above are possible whenever $\kappa(U_i, U_j)$ can be written as sums and products of random variables involving the individual U_i like we did in (1.3) for the 2×2 block graphon. This is possible in particular whenever κ is piece-wise constant, that is, is of block form, and in principle one could construct an algorithm that derives such decompositions explicitly for any subgraph density and any block graphon. In general, though, (1.3) has to be replaced by an approximation, and that will be the content of Lemma 2.2.

In the case where the subgraph is not connected, one can always decompose the subgraph density into sums and products of subgraph densities of connected graphs. For example, if $\mathrm{inj}(F, G)$ denotes the number of injective homomorphisms from F to G, we have

$$\mathrm{inj}(\text{△ ⋮}, G) = \mathrm{inj}(\text{△}, G)\,\mathrm{inj}(\text{⋰}, G) - 6\,\mathrm{inj}(\text{△}, G) - 6\,\mathrm{inj}(\text{△·}, G)$$

so that

$$t^{\mathrm{inj}}_{\text{△ ⋮}}(G) = \frac{n(n-1)}{(n-4)(n-5)} t^{\mathrm{inj}}_{\text{△}}(G)\, t^{\mathrm{inj}}_{\text{⋰}}(G) - \frac{6}{(n-4)(n-5)} t^{\mathrm{inj}}_{\text{△}}(G) - \frac{6}{n-5} t^{\mathrm{inj}}_{\text{△·}}(G). \tag{1.8}$$

Hence, densities of connected subgraphs tell in essence the whole story.

1.2 Preliminaries on dense graph limit theory

In what follows, all graphs are assumed to be simple and finite, without loops. Consider a graph G_n on the vertex set $[n] \coloneqq \{1, \dots, n\}$. For any graph F on k vertices, the *homomorphism density of F in G_n* is defined as

$$t_F(G_n) = \frac{\hom(F, G_n)}{n^k},$$

where $\hom(F, G_n)$ is the number of graph homomorphisms from F to G_n. A sequence of graphs $G_1, G_2, \dots$ is called *dense*, if the number of edges $e(G_n) \asymp n^2$, and it is called convergent if $\lim_{n\to\infty} t_F(G_n)$ exists for all F. Lovász and Szegedy (2006) showed that if $G_1, G_2, \dots$ is a convergent dense graph sequence, then there exists a symmetric measurable function $\kappa : [0,1]^2 \to [0,1]$ such that

$$\lim_{n\to\infty} t_F(G_n) = t_F(\kappa) \coloneqq \int_{[0,1]^k} \prod_{v \overset{F}{\sim} w} \kappa(x_v, x_w) dx_1 \cdots dx_k, \tag{1.9}$$

where $\prod_{v \overset{F}{\sim} w}$ denotes the product over all pairs of vertices $\{v, w\}$ that are connected in F. Such functions are generally referred to as *graphons*, but (1.9) is only enough to determine κ up to measure-preserving transformations of $[0,1]$, so the actual space of limiting objects is the equivalence class of graphons with the same values of $t_F(\kappa)$ for all F. What makes the representation of the limits appealing is that finite graphs can easily be embedded in the space of graphons by representing the adjacency matrix of a

graph as a 0-1-valued function on $[0,1]^2$ in the canonical way. If G is a graph and κ the corresponding induced graphon, it is not difficult to see that $t_F(G) = t_F(\kappa)$ for all F.

In the context of graphs, the more natural objects to study are the *injective* homomorphism densities, defined as

$$t_F^{\mathrm{inj}}(G_n) = \frac{\mathrm{inj}(F, G_n)}{(n)_k},$$

where $\mathrm{inj}(F, G_n)$ is the number of *injective* homomorphisms from F to G_n and where $(n)_k = n(n-1)\cdots(n-k+1)$. An approximate relation between $t_F(G_n)$ and $t_F^{\mathrm{inj}}(G_n)$ is given by the inequality

$$\left| t_F^{\mathrm{inj}}(G_n) - t_F(G_n) \right| \leqslant \frac{\binom{k}{2}}{n}. \tag{1.10}$$

Thus, $\lim t_F(G_n) = t_F(\kappa)$ if and only if $\lim t_F^{\mathrm{inj}}(G_n) = t_F(\kappa)$, and so from the point of view of dense graph limits, there is no difference between considering $t_F^{\mathrm{inj}}(G_n)$ instead of $t_F(G_n)$. However, higher-order fluctuations of these statistics are of smaller order than n^{-1}, and so (1.10) is not informative for such purposes. In this article, we will only focus on $t_F^{\mathrm{inj}}(G_n)$, but results can be translated in principle via certain identities, relating the numbers of homomorphisms and injective homomorphisms although the formulas are not straightforward; see for example Lovász (2012, Section 5.2.3.).

1.3 A simple (and naive) central limit theorem

In order to motivate much of the remainder of this article, and in particular justify the expression "higher-order" in the title rather than just "second-order" as one would naturally expect from the analogy with the classical central limit theorem, we start with a heuristic analysis of the workhorse model of dense graph limit theory. Let κ be a graphon and $U = (U_1, U_2, \ldots, U_n)$ be a sequence of independent random variables, each distributed uniformly on $[0,1]$, and given U, let $Y_{ij} = 1$ with probability $\kappa(U_i, U_j)$ and $Y_{ij} = 0$ with probability $1 - \kappa(U_i, U_j)$ for all $1 \leqslant i < j \leqslant n$. Let G_n be the graph on the vertex set $[n]$, where $i < j$ are connected if $Y_{ij} = 1$ and left unconnected otherwise. We denote the resulting random graph model by $\mathbb{G}(n, \kappa)$. Lovász and Szegedy (2006) proved the basic law of large numbers of dense graph limit theory, which states that such G_n converges to κ almost surely as n tends to infinity.

The case of the Erdős-Rényi random graph is the special case $\kappa \equiv p$ for some $0 \leqslant p \leqslant 1$, which we assume for now. The first-order behaviour is then given by $t_F^{\mathrm{inj}}(G_n) \to p^{e(F)}$, where $e(F)$ is the number of edges in F. Moreover, it is easy to see from (1.9) that if F consists of connected components $F_1, \ldots, F_m$, then for any graph G we have

$$t_F(G) = \prod_{i=1}^{m} t_{F_i}(G); \tag{1.11}$$

hence, it is enough to consider the fluctuations of $t_F(G)$ and $t_F^{\mathrm{inj}}(G)$ for *connected* F (although for $t_F^{\mathrm{inj}}(G)$, a clean identity such as (1.11) that does not involve n does not exist, as is apparent from (1.8)). Now, the second order fluctuations of $t_F^{\mathrm{inj}}(G)$ are not difficult to describe (see Janson and Nowicki (1991) for the general statements and Reinert and Röllin (2010) for rates of convergence in some special cases). Let K_2 be the one-edge graph on two vertices; then,

$$\mathrm{Cor}\left(t_{K_2}^{\mathrm{inj}}(G_n), t_F^{\mathrm{inj}}(G_n) \right) \to 1, \qquad n \to \infty. \tag{1.12}$$

This means that the second-order behaviour of all subgraph counts is asymptotically determined by the total number of edges. More specifically,

$$c_F\, n \left(t_F^{\mathrm{inj}}(G_n) - p^{e(F)} \right) \approx n \left(t_{K_2}^{\mathrm{inj}}(G_n) - p \right) \approx \mathrm{N}(0, 2p(1-p)), \tag{1.13}$$

where c_F is some combinatorial constant depending only on F, and where $\mathrm{N}(\mu, \sigma^2)$ denotes the normal distribution with respective mean and variance. Note that (1.12) and (1.13) remain true for general F, not just connected F, since even if F is not connected, the quantities $t_{F_i}(G)$ in (1.11) are centred around positive constants, so that $t_F(G)$ is dominated by a linear combination of the $t_{F_i}(G)$.

Now, we can consider a more refined view of the normal distribution appearing in (1.13) as follows. If κ_n denotes the 0-1-graphon induced by the adjacency matrix of G_n, we can analyse the centred and scaled graphon measure $\hat{Z}_n(dz) = n(\kappa_n(z) - p)dz$ for $z \in \mathcal{D}_2$, and think of it as converging weakly to a white noise process Z_2 living on $\mathcal{D}_2 = \{(y_1, y_2) \in [0,1]^2 : y_1 < y_2\}$ and having infinitesimal variance $p(1-p)dy$ (see next section for exact definitions). That is, for any weight function $\varphi \in L_2(\mathcal{D}_2)$,

$$\int_{\mathcal{D}_2} \varphi(y)\, \hat{Z}_n(dy) \xrightarrow{\mathscr{L}} \int_{\mathcal{D}_2} \varphi(y)\, Z_2(dy) \sim N(0, \|\varphi\|_{p,2}^2), \tag{1.14}$$

where $\|\varphi\|_{p,2}^2 = \int_{\mathcal{D}_2} \varphi(y)^2\, p(1-p)\, dy$, and this result can easily be established for multiple φ simultaneously (see (2.1) for precise definition of the stochastic integral). We use the term "white noise" loosely here, but in the next section, we will refer to Z_2 more appropriately as "Gaussian stochastic measure", since the term "white noise" has a more specific meaning in Hilda calculus; see, for example, Di Nunno, Øksendal, and Proske (2009) for an excellent introduction. Further embellishments of this result could be considered, such as the convergence of the integrated process

$$\hat{Z}_n(x, y) = \int_0^x \int_y^1 \hat{Z}_n(du, dv), \qquad (x, y) \in \mathcal{D}_2,$$

to a corresponding Brownian sheet on $\mathcal{D}_2$ (this requires a consistent ordering of the vertices, though).

Even in this refined view, the main deficiency remains, namely that the only randomness surviving the limiting procedure is that of the number of edges, albeit now with a description at a local level. For general graphons κ, this phenomenon of loss of randomness becomes even more pronounced. As we will see, for non-constant graphons, subgraph counts are dominated by functions of the form $\sum_{i=1}^n \psi(U_i)$; that is, the randomness coming from the vertex labels dominates, and no information about the edges in the graph survives when taking limits.

While from the point of view of the classical Central Limit Theorem this could be seen as the end of the story, we have not taken into account the fact that underlying all of the graph statistics are so-called *generalised U-statistics*. Such statistics have a much richer structure of fluctuations than sums of independent random variables. And while there is no canonical third and even higher-order fluctuation theory for sums of independent random variables, since it is not possible to make probabilistic sense out of "subtracting the dominating effect and analyse what is left" for sums of independent random variables without making use of signed measures, the situation for generalised (and regular) U-statistics is different, since fluctuations can happen simultaneously at different scales, and these fluctuations can be studied separately from one another.

1.4 Summary of main findings

We now give a summary of the remainder of this article in the easier case where the vertex labels are fixed and lie on an equally spaced lattice. That is, $U_i \equiv i/n$ for $1 \leqslant i \leqslant n$, and the edges Y_{ij} are sampled independently with probability $\kappa(U_i, U_j) = \kappa(i/n, j/n)$, $1 \leqslant i < j \leqslant n$. We will denote this random graph model by $\mathbb{G}_{\mathrm{lat}}(n, \kappa)$. The picture that emerges from the fluctuations of subgraph counts is as follows.

For each $k \geqslant 2$ and each connected graph F on the vertex set $[k]$, consider the collection of *centred* subgraph indicators

$$X_{F,a} = \prod_{v \overset{F}{\sim} w} \left(Y_{a_v a_w} - \kappa(a_v/n, a_w/n) \right), \qquad a \in \mathcal{I}_k^n, \tag{1.15}$$

where $\mathcal{I}_k^n = \{(a_1, \dots, a_k) \in \mathbb{N}^k : 1 \leqslant a_1 < \cdots < a_k \leqslant n\}$, and the corresponding statistic

$$W_F = \tbinom{n}{k}^{-1/2} \sum_{a \in \mathcal{I}_k^n} X_{F,a}. \tag{1.16}$$

It turns out that W_F converges to a Gaussian distribution, or more generally, the collection of random variables $X_F = (X_{F,a})_{a \in \mathcal{I}_k^n}$, scaled and embedded appropriately, converges to a white noise process Z_F that lives on the space

$$\mathcal{D}_k = \left\{ (x_1, \dots, x_k) \in [0,1]^k : x_1 \leqslant \cdots \leqslant x_k \right\}.$$

and has infinitesimal variance $\prod_{v \overset{F}{\sim} w} \kappa(u_v, u_w) \left(1 - \kappa(u_v, u_w)\right) du$; the case of $F = K_2$ is given in (1.14). Moreover, the processes Z_F turn out to be independent of each other for different F, and convergence holds jointly for any finite collection of F.

Note that we consider ordered sums as in (1.16) in our main result, and the fields Z_F are independent of each other even if the F are isomorphic (but not identical). Hence, sums of the form

$$\sum_{a \in \mathcal{A}_k^n} X_{F,a},$$

where $\mathcal{A}_k^n \subset [n]^k$ is the set of k-tuples of pairwise different indices, can be analysed by considering ordered sums and then summing over all isomorphic copies of F.

Now, regular subgraph indicators

$$\prod_{v \overset{F}{\sim} w} Y_{vw} \tag{1.17}$$

can be approximated in L_2 by linear combinations of random variables of the form (1.15), and in that sense, centred connected subgraph counts in (1.15) are really at the heart of all fluctuations of (1.17). In some cases, the approximation is in fact an equality, which lead to the identify (1.2) based on weighted sums of (1.15), leading to the quantities (1.6) and (1.7). We will also show that the rate of convergence is $\mathrm{O}\big(n^{-1/2}\big)$ for smooth-enough test functions and of order $\mathrm{O}\big(n^{-1/(2(p+2))}\big)$ for the convex set distance, where p is the maximal size of centred subgraphs considered, although the latter result is unlikely optimal.

While the collection of fields $(Z_F)_{F \in \mathcal{F}}$, where $\mathcal{F}$ is an enumeration of all connected finite graphs, can be thought of as the limiting object of some sort of "centred and normalised" graph, it is important to keep in mind that the limiting white noise fields are really just Gaussian stochastic measures, or equivalently, Gaussian Hilbert spaces, which are collections of Gaussian random variables and not objects in an actual Polish space for which we could define weak convergence. Thus, the results in this article only lay the foundations for such considerations; concretely, we establish convergence of finite dimensional distributions with rates of convergence. Further work is needed to turn this into a full-fledged notion of weak convergence.

For the model $\mathbb{G}(n, \kappa)$, where the U_i are independent uniform random variables on $[0,1]$, the fields Z_F need to be augmented by additional dimensions to take into account randomness of the vertex labels, as can be seen in (1.6) and (1.7), where the U_i do not just appear in the quantity W, but also act as weights in the remaining quantities $V_{\nearrow,2}$ and so forth. We will elaborate on this in more detail in Section 2.

1.5 Statistical applications

We believe Janson and Nowicki's theory of generalised U-statistics along with our explicit multivariate normal approximation theorem open up new possibilities for inference in statistical network analysis. While subgraph counts have been used for inference (see, for instance, the discussion by Ospina-Forero, Deane, and Reinert (2019)), we now make a few points on how centred subgraph counts could be used in statistical applications.

First, in the light of the results discussed in this article, we believe the model $\mathbb{G}(n,\kappa)$ is not appropriate for statistical applications, since the randomness of the vertex labels and the randomness of the edges are conflated. It seems more natural to think of network data *conditionally* on the vertex labels, which is equivalent to using the model $\mathbb{G}_{\mathrm{lat}}(n,\kappa)$.

Second, in order to calculate centred subgraph count statistics and use them for testing, the values $\kappa(U_v,U_w)$ need to be hypothesised *a priori* for each pair of vertices v and w. As a result, a statistical procedure based on $\mathbb{G}_{\mathrm{lat}}(n,\kappa)$ and centred subgraph counts to test whether the network is compatible with a specific graphon is in fact nothing but a test of whether a sequence of independent Bernoulli random variables $Y=(Y_{ij})_{1\leqslant i<j\leqslant n}$ are compatible with a specific model of their respective success probabilities $(p_{ij})_{1\leqslant i<j\leqslant n}$, and the choice of subgraphs $F\in\mathcal{F}$ determines to what sort of deviations the test is sensitive to. For instance, a test based solely on the edge-count test statistic

$$T_{\edge}(Y)=\sigma_{\edge}^{-1}\sum_{1\leqslant i<j\leqslant n}(Y_{ij}-p_{ij})\,,\qquad \sigma_{\edge}^2=\sum_{1\leqslant i<j\leqslant n}p_{ij}\,(1-p_{ij})\,,$$

is sensitive only to deviations of the overall edge density from that of the postulated model. By adding the two stars statistic

$$T_{\vee}(Y)=\sigma_{\vee}^{-1}\sum_{1\leqslant i<j<k\leqslant n}(Y_{ij}-p_{ij})\,(Y_{jk}-p_{jk})\,,$$
$$\sigma_{\vee}^2=\sum_{1\leqslant i<j<k\leqslant n}p_{ij}\,(1-p_{ij})\,p_{jk}\,(1-p_{jk})\,,$$

as well as the analogously defined statistics $T_{\wedge}(Y)$ and $T_{\triangle}(Y)$, a test will also detect deviations in the form of elevated levels of simultaneous presence or absence of edges with a common end point (leading to larger positive value of $T_{\vee}(Y)$), but also the opposite, namely elevated presence of mutual suppression, where presence of one edge inhibits presence of another (leading to a larger negative value of $T_{\vee}(Y)$). Correspondingly, higher order statistics yields information about presence of higher order dependencies among edges. What is noteworthy is that these statistics are very easy to calculate, and in particular the expressions for the variances are straightforward.

Third, if values for p_{ij} cannot be obtained *a priori*, centred subgraph counts can still serve as diagnostic tests after a model has been fitted by means of any other procedure. In such a case, these statistics can detect which aspects of the network have not been adequately captured by a model. For example, algorithms based on stochastic block models (see Funke and Becker (2019) for a survey on inference methods) typically yield a community assignment for each vertex as well as connection probabilities between any two communities, and these values can serve as estimates for p_{ij}. One has to keep in mind, though, that in order to make a valid statistical inference such a procedure would require *post-hoc* Type I error correction, since the network data has already been used to estimate the p_{ij}.

Last, an important, but difficult question is that of which centred subgraph counts should be used to determine whether a given network is compatible with a specific graphon, and this is related to the question of *forcibility* of graphons; see Lovász and

Table 1: Results of centred and standardized subgraph counts for fitted models using the function `BM_Bernoulli` from the R-package '`blockmodels`'. The numbers reported are defined in (1.18). The first data set is simulated from a 4×4 stochastic block model with 200 vertices, where each group has 50 vertices. The second data set is the hospital encounter network '`rfid`' from the R-package `igraphdata`, and consists of 75 vertices representing hospital staff along with encounter counts for each pair of staff. The bold columns represent the estimated number of groups as recommended by the function `BM_Bernoulli` using the *Integrated Classification Likelihood (ICL)* criterion proposed by Biernacki et al. (2000).

	Number of groups									
	1	2	3	4	5	6	7	8	9	10
	Data simulated from 4×4 stochastic block model									
$z_{\nearrow}$	0.00	0.01	0.00	**0.17**	0.08	−0.07	0.01	−0.02	0.10	0.04
$z_{\vee}$	4.65	2.47	2.27	**-0.57**	−0.51	−0.73	0.16	−1.30	−1.12	−0.95
$z_{\triangle}$	−18.57	−0.38	1.04	**0.03**	0.22	0.15	−0.23	−0.11	−0.32	−0.36
$z_{\boxplus}$	57.36	2.43	0.77	**-0.24**	−0.07	−0.04	−0.33	−0.33	−0.69	−0.60
$z_{\boxplus}$	−5.39	2.31	2.62	**-0.93**	−0.56	−0.39	−0.65	−0.36	−0.59	−0.35
$z_{\curlyvee}$	1.90	2.42	1.55	**1.29**	0.27	0.83	1.61	0.14	0.28	0.15
	Data set '`rfid`'									
$z_{\nearrow}$	0.00	−0.31	0.01	−0.33	**0.05**	−0.01	0.13	−0.07	−0.17	0.10
$z_{\vee}$	71.48	19.49	8.72	9.32	**9.87**	6.25	1.02	−0.15	2.31	0.68
$z_{\triangle}$	11.32	1.49	1.53	4.40	**7.96**	8.24	8.20	8.25	6.73	6.33
$z_{\boxplus}$	136.27	30.38	22.25	17.06	**13.43**	13.15	12.31	12.44	10.94	10.86
$z_{\boxplus}$	44.32	1.11	−1.74	−1.85	**-0.41**	0.50	−1.24	−1.32	−0.67	−1.18
$z_{\curlyvee}$	44.23	−3.36	4.55	5.87	**7.85**	2.97	−0.52	0.28	−0.67	−0.57

Szegedy (2011). For example, for constant graphons, it is enough to consider edge counts and four-cycle counts; see Fang and Röllin (2015) for a corresponding statistical procedure.

To illustrate how centred subgraph counts can be used in actual applications, we have analysed two data sets of small networks. The first is a simulated network with 200 vertices, drawn from a 4×4 stochastic block model with connection probabilities given by the matrix

$$K = \begin{pmatrix} 0.45 & 0.34 & 0.82 & 0.60 \\ 0.34 & 0.70 & 0.98 & 0.57 \\ 0.82 & 0.98 & 0.03 & 0.82 \\ 0.60 & 0.57 & 0.82 & 0.25 \end{pmatrix}.$$

Each group had 50 vertices assigned to it, after which the connections were sampled based on the probabilities K and the respective groups two vertices belonged to. We then used the function '`BM_bernoulli`' from the R-package '`blockmodels`' which, for each given number of groups, does both assignment of group labels to vertices (also called 'clustering', 'community detection' or 'community recovery') and estimation of connection probabilities $\hat{K}$. From this, the individual connection probabilities $\hat{p}_{ij}$ were be derived. The package `blockmodels` uses the *Integrated Classification Likelihood (ICL)* criterion proposed by Biernacki et al. (2000) to choose the optimal number of groups, but we have done the centred subgraph count analysis for all group sizes from 1 to 10; the results are given in Table 1. With $y = (y_{ij})_{1 \leqslant i < j \leqslant n}$ denoting the edge indicators and $(\hat{p}_{ij})_{1 \leqslant i < j \leqslant n}$ denoting the estimated connection probabilities (which are a function of y),

let $\hat{y}_{ij} = y_{ij} - \hat{p}_{ij}$, the numbers reported in Table 1 are defined as

$$\begin{aligned} z_{\cdot\!\cdot}(y) &= \hat{\sigma}_{\cdot\!\cdot}^{-1}(y) \sum_{1\leqslant i<j\leqslant n} \hat{y}_{ij}, \qquad z_{\triangle}(y) = \hat{\sigma}_{\triangle}^{-1}(y) \sum_{1\leqslant i<j<k\leqslant n} \hat{y}_{ij}\hat{y}_{jk}\hat{y}_{ik}, \\ z_{\vee} &= \hat{\sigma}_{\vee}^{-1}(y) \sum_{1\leqslant i<j<k\leqslant n} \left(\hat{y}_{ij}\hat{y}_{jk} + \hat{y}_{ji}\hat{y}_{ik} + \hat{y}_{ik}\hat{y}_{kj}\right), \end{aligned} \tag{1.18}$$

where

$$\begin{aligned} \hat{\sigma}^2_{\cdot\!\cdot}(y) &= \sum_{1\leqslant i<j\leqslant n} \hat{p}_{ij}\left(1-\hat{p}_{ij}\right), \quad \hat{\sigma}^2_{\triangle}(y) = \sum_{1\leqslant i<j<k\leqslant n} \hat{p}_{ij}\left(1-\hat{p}_{ij}\right)\hat{p}_{jk}\left(1-\hat{p}_{jk}\right)\hat{p}_{ik}\left(1-\hat{p}_{ik}\right) \\ \hat{\sigma}^2_{\vee}(y) &= \sum_{1\leqslant i<j<k\leqslant n} \left(\hat{p}_{ij}\left(1-\hat{p}_{ij}\right)\hat{p}_{jk}\left(1-\hat{p}_{jk}\right) + \hat{p}_{ji}\left(1-\hat{p}_{ji}\right)\hat{p}_{ik}\left(1-\hat{p}_{ik}\right)\right. \\ &\qquad \left. + \hat{p}_{ik}\left(1-\hat{p}_{ik}\right)\hat{p}_{kj}\left(1-\hat{p}_{kj}\right)\right); \end{aligned}$$

the quantities $z_{\square\!\square}$, $z_{\square\!\square}$ and $z_{\triangle\!\cdot}$ are defined analogously. For this simulated network data, it is evident that the models fitted by `BM_bernoulli` for four or more groups are consistent with the three centred subgraph count statistics reported, and that they are not consistent when only one, two or three groups are hypothesised.

The second data set is from a hospital, where 75 individuals of the staff were equipped with devices to record encounters between these individuals whenever the devices were in close proximity to each other. Since the network is edge-weighted, where each weight represents the number of encounters recorded over the time period of the experiment, we have converted this to a simple unweighted network. An edge is present between two individuals if they had at least one encounter. It is clear that for this real data, the fitted stochastic block models do not capture higher order dependence well. Even when allowing the clustering algorithm dividing the vertices into ten groups, the dependence captured by centred triangles is not what one would expect from a stochastic block model.

2 Subgraph counts and generalised U-statistics

In this section we review and discuss material from Janson and Nowicki (1991) and Janson (1994, 1997) in order to show that the existing literature on generalised U-statistics does provide a suitable framework to describe the fluctuations arising in standard dense graph models. The material in this section is not strictly necessary to state and prove the main results in Section 3, but it gives the motivating context and associated general limit theory.

2.1 Gaussian Hilbert Spaces

We follow Janson (1997) in essence. While Gaussian Hilbert spaces will serve as a form of limiting objects, it is important to keep in mind that at this level of abstraction, Gaussian Hilbert spaces are just collections of Gaussian random variables, and there is not really a single object taking values in a single space. Although, for instance, Brownian motion indexed by time can be seen as a Gaussian Hilbert space, it comes with additional properties such as almost-sure path-wise continuity, which is a statement about the joint distribution of uncountably many of the variables and goes beyond the general theory discussed here.

Now, let H be a Hilbert space, where we denote the inner product by $\langle\cdot,\cdot\rangle_H$ and the resulting norm by $\|h\|_H := \sqrt{\langle h,h\rangle_H}$, although we will drop the dependence on H for norms and inner products if there is no ambiguity. A *Gaussian Hilbert space* indexed by H is a collection of centred Gaussian variables $(Z_h)_{h\in H}$ defined on a common probability

space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\operatorname{Cov}(Z_h, Z_{h'}) = \mathbb{E}\{Z_h Z_{h'}\} = \langle h, h' \rangle_H, \quad h, h' \in H.$$

Clearly, $\mathbb{E}Z_h^2 = \operatorname{Var} Z_h = \|h\|_H^2$. It is known that such a family can be constructed for every Hilbert space in such a way that, if $h_n \to h$ in H as $n \to \infty$, then $Z_{h_n} \to Z_h$ in $L_2(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, any countable collection of the Z_h of a Gaussian Hilbert space is jointly Gaussian.

Gaussian stochastic measures and stochastic integrals. Let $(M, \mathcal{M}, \mu)$ be a measure space, and consider the Hilbert space $L_2(M)$ (we drop the σ-algebra and measure from the notation if it does not cause ambiguity). A Gaussian Hilbert space $(Z_\varphi)_{\varphi \in L_2(M)}$ can be interpreted as a *Gaussian stochastic integral* on M by setting

$$\int_M \varphi(x)\, Z(dx) := Z_\varphi, \qquad \varphi \in L_2(M). \tag{2.1}$$

Indeed, the family of random variables defined by $Z(A) := Z_{I_A}$, where $A \in \mathcal{M}$ with $\mu(A) < \infty$ so that the indicator function I_A is in $L_2(M)$, defines a *Gaussian stochastic measure* Z on M, which has the following properties:

(i) if $A \in \mathcal{M}$ with $\mu(A) < \infty$, then

$$Z(A) \sim N(0, \mu(A));$$

(ii) if $A_1, A_2, \ldots \in M$ are disjoint sets with $\mu(A_i) < \infty$ for all $i \geqslant 1$, then the random variables $Z(A_1), Z(A_2), \ldots$ are mutually independent and

$$Z\left(\bigcup_{i \geqslant 1} A_i\right) = \sum_{i \geqslant 1} Z(A_i)$$

(note that convergence on the right hand side is in L_2, but since the summands are independent, it is also almost sure by Kolmogorov's three-series theorem).

This justifies the notation $\int \varphi\, dZ$ in (2.1), since any Gaussian stochastic measure on M in turn uniquely defines a Gaussian stochastic integral via the standard procedure of approximating functions in $L_2(M)$ via simple functions and taking closure. Note that a Gaussian stochastic measure can be loosely interpreted as white noise, but we will avoid this terminology for the remainder of this article for the reasons given in the previous section.

We can extend the single stochastic integral to a multiple stochastic integral

$$\int_{M^k} \varphi(x)\, Z^k(dx), \qquad \varphi \in L_2(M^k, \mu^k), \tag{2.2}$$

where μ^k is the usual product measure on the product sigma-algebra $\mathcal{M}^{\otimes k}$. To this end, let $A_1, \ldots, A_n \subset M$ be measurable and pairwise disjoint, and consider simple functions of the form

$$\varphi(x) = \sum_{i_1, \ldots, i_k = 1}^{n} \varphi_{i_1, \ldots, i_k}\, \mathrm{I}[x_1 \in A_{i_1}, \ldots, x_k \in A_{i_k}], \tag{2.3}$$

where $\varphi_{i_1, \ldots, i_k}$ vanishes whenever any two of the indices coincide. For such functions, the multiple integral can be defined as

$$\int_{M^k} \varphi(x)\, Z^k(dx) = \sum_{i_1, \ldots, i_k = 1}^{n} \varphi_{i_1, \ldots, i_k} Z(A_{i_1}) \cdots Z(A_{i_k}),$$

and the general case $\varphi \in L_2(M^k)$ can be obtained by approximating such functions by functions of the form (2.3); we refer to Nualart (2006) for details. The integral (2.2) turns out to be an element of the kth *Wiener Chaos* $\mathcal{H}_k$, which is the L_2-closure of the space generated by the random variables $\{H_k(Z_h); h \in H, \|h\|_H = 1\}$, where H_k is the kth Hermite polynomial.

2.2 Gaussian limits related to sums of independent random variables

Before detailing on the results known for generalised U-statistics, it is illuminating to briefly review the different types of results known for independent random variables, and how these results can be formulated in the framework of Gaussian Hilbert spaces.

In what follows, let $X_1, X_2, \dots$, be independent and identically distributed random variables with $\mathbb{E}X_1 = 0$ and $\operatorname{Var} X_1 = 1$.

Central Limit Theorem and Donsker's theorem. Let $H = \mathbb{R}$; the corresponding Gaussian Hilbert space can be simply constructed by taking a standard Gaussian variable Z_1 and letting $Z_c = cZ_1$ for $c \in \mathbb{R}$. The standard CLT then yields

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} cX_i \xrightarrow{\mathscr{L}} Z_c, \qquad c \in \mathbb{R}. \tag{2.4}$$

We can generalise (2.4) and replace the constant c on the left hand side by a general weight function. To this end, consider the Hilbert space $H = L_2([0,1])$ with the usual inner product $\langle \varphi_1, \varphi_2 \rangle = \int_0^1 \varphi_1(x)\varphi_2(x)dx$, and let $(Z_\varphi)_{\varphi \in L_2([0,1])}$ be a corresponding Gaussian Hilbert space. For given $\varphi \in L_2([0,1])$, which we assume to be continuous almost everywhere to avoid certain technical difficulties which are irrelevant for this discussion, Donsker's theorem yields

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} \varphi(i/n)X_i \xrightarrow{\mathscr{L}} \int_{[0,1]} \varphi(x)\, Z(dx), \qquad \varphi \in L_2([0,1]), \tag{2.5}$$

and this holds jointly for any finite collection of such φ. Moreover, (2.4) follows from (2.5) if we choose $\varphi \equiv c$, where $c \in \mathbb{R}$. It is important to stress that Donsker's theorem gives a stronger result than that. In fact, if we take $\varphi_t(x) = \mathrm{I}[x \leqslant t]$, where $0 \leqslant t \leqslant 1$, we can construct the Gaussian Hilbert space in such a way that the process $(Z_{\varphi_t})_{0 \leqslant t \leqslant 1}$ is almost surely *continuous in* t, so that this process can be identified with standard Brownian motion $B_t = Z_{\varphi_t}$ for $0 \leqslant t \leqslant 1$. And so, what Donsker's theorem actually yields is that

$$\left(\frac{1}{n^{1/2}} \sum_{i=1}^{\lfloor nt \rfloor} X_i \right)_{0 \leqslant t \leqslant 1} \xrightarrow{\mathscr{L}} (B_t)_{0 \leqslant t \leqslant 1}, \tag{2.6}$$

where weak convergence is with respect to the Skorohod topology (or uniform topology if the process on the left hand side of (2.6) is interpolated between jumps).

***U*-statistics.** Before turning to U-statistics, we first consider real-valued functions of the X_i. To this end, we may assume that the X_i take values in a general measure space $\mathcal{S}$, and we denote the distribution of X_i by μ. Consider the Hilbert space $H = L_2([0,1] \times \mathcal{S}, dt \times \mu)$ with the canonical inner product that satisfies $\langle \varphi_1\psi_1, \varphi_2\psi_2 \rangle = \langle \varphi_1, \varphi_2 \rangle \langle \psi_1, \psi_2 \rangle_\mu$, where $(\varphi\psi)(x,y) := \varphi(x)\psi(y)$. Denoting by $L_2^\circ(\mathcal{S})$ the set of functions $\psi \in L_2(\mathcal{S})$ with $\mathbb{E}\psi(X_1) = 0$, and assuming again that φ is continuous almost everywhere, it can be shown that

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\varphi(i/n)\psi(X_i) \xrightarrow{\mathscr{L}} \int_{[0,1]\times\mathcal{S}}\varphi(t)\psi(x)\,Z(dt,dx),$$
$$\varphi\in L_2([0,1]),\,\psi\in L_2^\circ(\mathcal{S}). \quad (2.7)$$

Again, this statement is also true jointly for any finite collection of (φ_i,ψ_i). Note that Z has infinitesimal variance $\mu(dx)$, since the space $L_2^\circ(\mathcal{S})$ comes with inner product $\langle\psi_1,\psi_1\rangle=\int_{\mathcal{S}}\psi_1(x)\psi_2(x)\,\mu(dx)$, and the quantity $Z(dt\times dx)$ can be loosely interpreted as "the normalised number of times the value dx has been observed among the indices dt". Note also that restricting ψ to be in $L_2^\circ(\mathcal{S})$ and not in $L_2(\mathcal{S})$ is necessary, since the measure Z has more degrees of freedom than the finite-n system, and thus cannot be fully observed. For example, if $X_i\sim\mathrm{Be}(p)$, then $\mathcal{S}=\{0,1\}$, and all functions $\psi\in L_2^\circ(\mathcal{S})$ are multiples of the function $\psi(0)=-p$ and $\psi(1)=1-p$. However, the variables $Z_0=Z([0,1]\times\{0\})$ and $Z_1=Z([0,1]\times\{1\})$, while constructed to be independent, cannot be observed individually — only their weighted sum $-pZ_0+(1-p)Z_1$ can be observed. This stems from the fact that the number of X_i with value 1 must equal n minus the number of X_i with value 0, and in this sense, the Gaussian Hilbert space $L_2([0,1]\times\mathcal{S})$ is slightly too big.

The result for U-statistics can be stated without introducing a new Gaussian Hilbert space — we only need multiple integrals over the same space. To this end, let

$$L_2^\circ(\mathcal{S}^k)=\left\{\psi\in L_2(\mathcal{S}^k)\,:\,\int_{\mathcal{S}}\psi(x_1,\dots,x_k)\,\mu(dx_i)=0,\,1\leqslant i\leqslant k,\,(x_j)_{j\neq i}\in\mathcal{S}^{k-1}\right\}. \quad (2.8)$$

For $a=(a_1,\dots,a_k)\in\mathcal{I}_k^n$, write $a/n\coloneqq(a_1/n,\dots,a_k/n)$ and $X_a=(X_{a_1},\dots,X_{a_k})$. Then, for almost everywhere continuous $\varphi\in L_2(\mathcal{D}_k)$,

$$\frac{1}{n^{k/2}}\sum_{a\in\mathcal{I}_k^n}\varphi(a/n)\psi(X_a) \xrightarrow{\mathscr{L}} \int_{\mathcal{D}_k\times\mathcal{S}^k}\varphi(t)\psi(x)\,Z^k(dt,dx),$$
$$\varphi\in L_2(\mathcal{D}_k),\,\psi\in L_2^\circ(\mathcal{S}^k). \quad (2.9)$$

Again, this statement is true jointly for any finite collection of (φ_i,ψ_i), even with different k; see Janson (1997, Theorem 11.16). Of course (2.7) is just a special case of (2.9), but (2.9) *follows* in essence from (2.7) and the continuous mapping theorem. General functions $\psi\in L_2(\mathcal{S}^k)$ can be decomposed into orthogonal elements $\psi_i\in L_2^\circ(\mathcal{S}^i)$, and a corresponding limit result then follows from (2.9), depending on the lowest-order, non-vanishing element φ_i (which in turn also determines the correct scaling to obtain a non-trivial limit); see Janson (1997, Theorem 11.19).

Generalised U-statistics. The final extension we consider are generalised U-statistics, which were introduced by Janson and Nowicki (1991). To this end, assume the X_i now take values in a space $\mathcal{S}_1$ with distribution μ_1, and let $(Y_{ij})_{1\leqslant i<j}$ be independent and identically distributed random elements taking values in a space $\mathcal{S}_2$ with distribution μ_2. For $a\in\mathcal{I}_k^n$, we define X_a as before and we let $Y_a=(Y_{a_ia_j})_{1\leqslant i<j\leqslant k}$. We now consider functions defined on the space

$$\mathcal{T}_k=\mathcal{S}_1^k\times\mathcal{S}_2^{\binom{k}{2}},\qquad\text{with measure }\mu_1^k\times\mu_2^{\binom{k}{2}},$$

where it is understood that $\mathcal{T}_1=\mathcal{S}_1$. For any function $\psi\colon\mathcal{T}_k\to\mathbb{R}$, we will write

$$\psi(X_a,Y_a)=\psi(X_{a_1},\dots,X_{a_k},Y_{a_1a_2},\dots,Y_{a_{k-1}a_k}).$$

Let now

$$L_2^\circ(\mathcal{T}_1)=\{\psi\in L_2(\mathcal{T}_1):\mathbb{E}\psi(X_1)=0\}.$$

For $k > 1$, let

$$\mathcal{F}_{-l}^k = \sigma\left(X_1, \dots, X_k\right) \vee \sigma\left(Y_{ij} : \text{for all } 1 \leqslant i < j \leqslant k \text{ with } l \notin \{i, j\}\right)$$

Then, define

$$L_2^\circ(\mathcal{T}_k) = \left\{\psi \in L_2(\mathcal{T}_k) : \mathbb{E}\left\{\psi\left(X, Y\right) \mid \mathcal{F}_{-l}^k\right\} = 0 \text{ for all } 1 \leqslant l \leqslant k\right\}. \tag{2.10}$$

In words, $L_2^\circ(\mathcal{T}_k)$ consists of all those functions that, for every $1 \leqslant l \leqslant k$, vanish when being simultaneously integrated over all Y_{ij} with $i = l$ or $j = l$. Then, with Z_k a Gaussian stochastic measure on $L_2(\mathcal{D}_k \times \mathcal{T}_k)$,

$$\begin{aligned}\frac{1}{n^{k/2}} \sum_{a \in \mathcal{I}_k^n} \varphi(a/n)\psi(X_a, Y_a) \xrightarrow{\mathscr{L}} \int_{\mathcal{D}_k \times \mathcal{T}_k} \varphi(t)\psi(x, y)\, Z_k(dt, dx, dy), \\ \varphi \in L_2(\mathcal{D}_k),\ \psi \in L_2^\circ(\mathcal{T}_k)\end{aligned} \tag{2.11}$$

(here, t and x are k-dimensional vectors, while y is a $\binom{k}{2}$-dimensional vector); see Janson (1997, Theorem 11.28) for general $\psi \in L_2(\mathcal{T}_k)$ via orthogonal decomposition.

2.3 Application to centred subgraph counts

We first apply (2.11) to centred subgraph counts of $\mathbb{G}(n, \kappa)$. However, since the Y_{ij} in $\mathbb{G}(n, \kappa)$ are not independent of each other, we need to resort to an auxiliary representation. Let $(U_i)_{i \geqslant 1}$ and $(V_{ij})_{i,j \geqslant 1}$ be independent random variables uniformly distributed on $[0, 1]$, hence $\mathcal{S}_1 = \mathcal{S}_2 = [0, 1]$, endowed with the Lebesgue measure. We construct a graph G_n on the vertex set $[n]$ by connecting vertices i and j if $V_{ij} \leqslant \kappa(U_i, U_j)$. For a given connected graph F on k vertices, we consider the function

$$\psi_F\left(u, v\right) = \psi(u) \prod_{i \overset{F}{\sim} j} \left(\mathrm{I}[v_{ij} \leqslant \kappa(u_i, u_j)] - \kappa(u_i, u_j)\right), \qquad u \in [0, 1]^k,\ v \in [0, 1]^{\binom{k}{2}}. \tag{2.12}$$

It is easy to verify that $\psi_F \in L_2^\circ(\mathcal{T}_k)$, and hence

$$\begin{aligned}&\frac{1}{n^{k/2}} \sum_{a \in \mathcal{I}_k^n} \varphi(a/n)\psi(U_a) \prod_{i \overset{F}{\sim} j} \left(\mathrm{I}[V_{a_i a_j} \leqslant \kappa(U_{a_i}, U_{a_j})] - \kappa(U_{a_i}, U_{a_j})\right) \\ &\xrightarrow{\mathscr{L}} \int_{\mathcal{D}_k \times [0,1]^k \times [0,1]^{\binom{k}{2}}} \varphi(t)\psi(u) \prod_{i \overset{F}{\sim} j} \left(\mathrm{I}[v_{ij} \leqslant \kappa(u_i, u_j)] - \kappa(u_i, u_j)\right)\, Z_k(dt, du, dv).\end{aligned} \tag{2.13}$$

Two comments are in place. First, the quantity $Z_k(dt, du, dv)$, in particular the dv-part, does not admit an intuitive interpretation, since the uniform random variables V_{ij} are only used as an auxiliary tool to represent subgraph counts as generalised U-statistics. Evaluating the stochastic integral in (2.13) with respect to dv, it is not difficult to show that integral can be written as

$$\int_{\mathcal{D}_k \times [0,1]^k} \varphi(t)\psi(u)\, Z_k(dt, du),$$

where Z_k now is a Gaussian stochastic measure on the space

$$L_2\left(\mathcal{D}_k \times [0, 1]^k, dt \times \prod_{i \overset{F}{\sim} j} \kappa(u_i, u_j)\left(1 - \kappa(u_i, u_j)\right) du\right).$$

Then, instead of (2.12), we will consider functions of the form

$$\psi_F(u,y) = \psi(u) \prod_{i \overset{F}{\sim} j} (y_{ij} - \kappa(u_i,u_j)), \qquad u \in [0,1]^k,\ y \in [0,1]^{\binom{k}{2}},\ \psi \in L_2([0,1]^k),$$

which allows to avoid the auxiliary representation via the V_{ij} and use the Y_{ij} directly, despite their dependence, and also allows to consider weighted edges. Hence, what we will show in the main section is that for U as before, and random variables Y_{ij}, which are conditionally independent given U and which satisfy $\mathbb{E}\{Y_{ij} \mid U\} = \kappa(U_i, U_j)$,

$$\frac{1}{n^{k/2}} \sum_{a \in \mathcal{I}_k^n} \varphi(a/n)\psi(U_a) \prod_{i \overset{F}{\sim} j} \big(Y_{a_i a_j} - \kappa(U_{a_i}, U_{a_j})\big) \xrightarrow{\mathscr{L}} \int_{\mathcal{D}_k \times [0,1]^k} \varphi(t)\psi(u)\, Z_k(dt, du). \quad (2.14)$$

Note that the measure Z_k is homogeneous over $\mathcal{D}_k$; this is because the U_i 'average out' the differences in the variances of the Y_{ij}, so that the points in $\mathcal{D}_k$ only see the combined variance effect across all Y_{ij}. This is different in $\mathbb{G}_{\mathrm{lat}}(n,\kappa)$, which is not vertex-exchangeable; see Remarks 3.3 and 3.4 for further discussion.

Second, (2.13) does not cover the case where the U_i are not identically distributed. This is important in particular for the model $\mathbb{G}_{\mathrm{lat}}(n,\kappa)$, but our results hold in greater generality.

2.4 An orthogonal decomposition of subgraph counts

We now discuss how (2.14) can be used to understand fluctuations of subgraph counts of $\mathbb{G}(n,\kappa)$, and so fix a graphon κ, let $U = (U_1, \dots, U_n)$ be a sequence of independent random variables uniformly distributed on $[0,1]$, let $Y = (Y_{ij})_{1 \leqslant i < j \leqslant n}$ be conditionally independent given U and distributed as before, and construct G_n on n vertices as before. Let F be a graph on the vertex set $[k]$ and write

$$t_F^{\mathrm{inj}}(G_n) = \frac{1}{(n)_k} \sum_{a \in \mathcal{A}_k^n} \prod_{i \overset{F}{\sim} j} Y_{ij}; \quad (2.15)$$

without loss of generality, we may assume that F has no isolated vertices. Now, the following lemma states that $t_F^{\mathrm{inj}}(G_n)$ can be decomposed into a sum of mostly uncorrelated centred subgraph counts, weighted by functions of the vertex labels. Some notation is needed. For a graph H we denote by $|H|$ the number of vertices in H, and for two (vertex-labelled) graphs H and H', we denote by $H \cup H'$ the graph where two vertices are connected if they are connected in at least one of H and H'. For a subgraph $H \subseteq F$ and a subset $A \subseteq [k]$, we denote by $H \cup A$ the graph obtained by interpreting A as the empty graph on the vertex set A. Moreover, we denote by H_P the unique graph on the vertex set $\{1, \dots, |H|\}$ that is isomorphic to H and preserves the ordering of the vertex labels, and we denote by $H \subseteq' F$ that H is a subgraph of F and that it has no isolated vertices. Let

$$\vartheta_H(u,y) = \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i,u_j)). \quad (2.16)$$

Lemma 2.1. *Let F be a graph on the vertex set $[k]$ without isolated vertices. Then there are functions $\psi_{H,A} \in L_2^\circ\big([0,1]^{|A|}\big)$ for $H \subseteq' F$ and $A \subset [k]$, such that*

$$t_F^{\mathrm{inj}}(G_n) = \sum_{H \subseteq' F} \sum_{A \subseteq [k]} r_{H,A}(U,Y), \quad (2.17)$$

where, with $l = |H \cup A|$,

$$r_{H,A}(u,y) = \frac{1}{(n)_l} \sum_{a \in \mathcal{A}_l^n} \psi_{H,A}(u_{a_{l-|A|+1}}, \dots, u_{a_l})\, \vartheta_{H_P}(u_{a_1}, \dots, u_{a_{|H|}}),$$

and such that the following holds: If H *and* H' *are not isomorphic or if* $|A| \neq |A'|$, *then*

$$\operatorname{Cov}\left(r_{H,A}(U,Y), r_{H',A'}(U,Y)\right) = 0,$$

and if $\psi_{H,A} \not\equiv 0$, *then, again with* $l = |H \cup A|$,

$$\operatorname{Var} r_{H,A}(U,Y) \asymp n^{-l}.$$

The key of this decomposition is that terms with different scalings are uncorrelated, so that we can separate the different orders of fluctuations of $t_F^{\mathrm{inj}}(G_n)$. However, whether $r_{H,A}(U,Y)$ has a normal limit or not, depends on H and A. Simply put, if $H \cup A$ is connected (for which it is necessary that $A \subset H$), the limit is normal, otherwise the limit is an element from a higher-order Wiener chaos. However, since each Wiener chaos itself is obtained by taking products of the underlying Gaussian Hilbert space and taking limits, we can decompose $r_{H,A}$ further, but only in an approximate sense.

The following lemma makes this precise and states that the statistics on the right hand side of (2.17) can be approximated in L_2 by products and sums of simpler statistics to any prescribed level of accuracy.

Lemma 2.2. *Let* $H \subseteq' F$, *and let* $A \subset [k]$. *Let* $\psi_{H,A}$ *and* $r_{H,A}$ *be as in Lemma 2.1. Denote by* $C_1, \dots, C_r$ *the connected components of* $H \cup A$ *and* $k_1, \dots, k_r$ *their respective sizes, and assume* $r \geqslant 2$ *(note that* $k_1 + \dots + k_r = l$*). For each* j, *let* C_j' *be an isomorphic copy of* C_j *on* $[k_j]$. *Then, for each* $\varepsilon > 0$, *there exists* N *and there exist functions* $\psi_{i,j} \in L_2^\circ([0,1]^{k_j})$ *for* $1 \leqslant i \leqslant N$ *and* $1 \leqslant j \leqslant r$, *such that*

$$\mathbb{E}\left(r_{H,A}(U,Y) - \sum_{i=1}^{N} \prod_{j=1}^{r} \frac{1}{(n)_{k_j}} \sum_{a \in \mathcal{A}_{k_j}^n} \psi_{i,j}(U_a) \vartheta_{C_j'}(U_a, Y_a) \right)^2 \leqslant \frac{\varepsilon}{n^l}, \tag{2.18}$$

for all n.

In other words, the standardised statistics $n^{l/2} r_{H,A}$ can be approximated in L_2 by products and sums of centred *connected* subgraphs counts uniformly in n, and we will show that these statistics themselves all have Gaussian limits. While the overall quality of approximation of t_F^{inj} is only of order n^{-2} in general, the important point here is of course that the fluctuations at different scalings are uncorrelated and become independent in the limit.

3 Main result

We are now ready to formulate our main result, which provides bounds on the multivariate normal approximation of sums as they appear in (2.18). In order to have a cleaner framework, our result will be formulated for sums over the index set $\mathcal{I}_k^n$, which makes all summands uncorrelated, but sums over $\mathcal{A}_k^n$ as they appear in (2.18) can of course be easily computed from sums over $\mathcal{I}_k^n$.

3.1 Gaussian approximation of centred connected subgraph counts

Let $n \geqslant 1$, let κ be a graphon, and assume $\kappa \not\equiv 0$ and $\kappa \not\equiv 1$. Let $U = (U_v)_{v \in [n]}$ be independent (but not necessarily identically distributed) random variables, and given

U, let $(Y_{vw})_{1\leqslant v<w\leqslant n}$ be random variables that are conditionally independent given U and that satisfy $\mathbb{E}\{Y_{ij} \mid U\} = \kappa(U_i, U_j)$. Recall that we set $U_a = (U_{a_1}, \dots, U_{a_k})$ for $a \in \mathcal{I}_k^n$. Let $d \geqslant 1$, and for each $1 \leqslant i \leqslant d$, let F_i be a connected graph on the vertex set $[k_i]$, where $k_i \geqslant 1$, let $\varphi_i \in L_2(\mathcal{D}_{k_i})$ and $\psi_i \in L_2([0,1]^{k_i})$ (note that here we do not require that $\psi_i \in L_2^\circ([0,1]^{k_i})$, since centring is either done explicitly below if $k_i = 1$, or else is not necessary since the centred subgraphs provide the centring). For $1 \leqslant i \leqslant d$, define

$$W_i = \begin{cases} n^{-1/2} \sum_{a=1}^{n} \varphi_i(a/n) \left(\psi_i(U_a) - \mathbb{E}\psi_i(U_a)\right) & \text{if } k_i = 1, \\ \binom{n}{k_i}^{-1/2} \sum_{a \in \mathcal{I}_{k_i}^n} \varphi_i(a/n)\, \psi_i(U_a) \prod_{v \overset{F_i}{\sim} w} \left(Y_{a_v a_w} - \kappa(U_{a_v}, U_{a_w})\right) & \text{if } k_i \geqslant 2, \end{cases} \tag{3.1}$$

and let $W = (W_1, \dots, W_d)$. Then, for $\Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant d} = \operatorname{Var} W$, we have

$$\sigma_{ij} = \begin{cases} n^{-1} \sum_{a=1}^{n} \varphi_i(a/n)\varphi_j(a/n) \operatorname{Cov}\left(\psi_i(U_a), \psi_j(U_a)\right) & \text{if } k_i = k_j = 1, \\ \binom{n}{k_i}^{-1} \sum_{a \in \mathcal{I}_{k_i}^n} \varphi_i(a/n)\varphi_j(a/n) \mathbb{E}\left\{ \psi_i(U_a)\psi_j(U_a) \prod_{v \overset{F_i}{\sim} w} \operatorname{Var}\left(Y_{a_{ij}} \mid U\right) \right\} & \text{if } F_i = F_j, \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

(we emphasise that in (3.2), the condition '$F_i = F_j$' really means equality including vertex labels, not just that F_i and F_j are isomorphic). Before stating our main result, we need some more notation. For a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$ of non-negative integers and $z \in \mathbb{R}^d$, let

$$|\alpha| = \alpha_1 + \cdots + \alpha_d, \qquad \alpha! = \alpha_1! \cdots \alpha_d!, \qquad z^\alpha = z_1^{\alpha_1} \cdots z_d^{\alpha_d}$$

and

$$\partial^\alpha g(x) = \frac{\partial^{|\alpha|} g(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

For any multi-index $\alpha \in \mathbb{N}^d$, let

$$|h|_\alpha = \sup_{x \in \mathbb{R}^d} \left|\partial^\alpha h(x)\right|.$$

Moreover, for two d-dimensional random vectors X and Y, and with $\mathcal{K}$ the class of convex sets in $\mathbb{R}^d$, define the convex set distance

$$d_c\left(\mathscr{L}(X), \mathscr{L}(Y)\right) = \sup_{A \in \mathcal{K}} \left|\mathbb{P}[X \in A] - \mathbb{P}[Y \in A]\right|.$$

Theorem 3.1. *Let W be defined as in* (3.1)*, and let $Z = (Z_1, \dots, Z_d)$ be a centred Gaussian random vector with covariance matrix Σ as given by* (3.2)*, and assume the Y_{ij}, φ_i and ψ_i are all bounded. Let p be an odd integer such that $p \geqslant \max\{k_1, \dots, k_d\}$. Then, for any $(p+2)$-times partially differentiable function $h : \mathbb{R}^d \to \mathbb{R}$*

$$\left|\mathbb{E}h(W) - \mathbb{E}h(Z)\right| \leqslant \frac{C \sup_{\alpha:|\alpha|\leqslant p+2} |h|_\alpha}{n^{1/2}} \tag{3.3}$$

for some constant C that is independent of n. Moreover,

$$d_c\left(\mathscr{L}(W), \mathscr{L}(Z)\right) \leqslant C n^{-\frac{1}{2(p+2)}} \tag{3.4}$$

again for some constant C that is independent of n.

Remark 3.2. Consider again the example of the 2×2 graphon and $t^{\mathrm{inj}}_{\triangle}(G_n)$ from Section 1.1. For example, letting W_1 equal W from (1.6), we can write

$$W_1 = n^{-1/2} \sum_{a=1}^{n} \varphi_1(a/n) \left(\psi_1(U_a) - \mathbb{E}\psi_1(U_a)\right), \qquad \varphi_1 \equiv 1,\ \psi_1(u) = u;$$

letting W_2 equal $V_{\diagup,1}$ from (1.6), we can write

$$W_2 = \tbinom{n}{2}^{-1/2} \sum_{a \in \mathcal{I}_2^n} \varphi_2(a/n)\, \psi_2(U_a) \prod_{v \overset{\diagup}{\sim} w} \left(Y_{a_v a_w} - \kappa(U_{a_v}, U_{a_w})\right), \qquad \varphi_2 \equiv 1,\ \psi_2 \equiv 1;$$

letting W_3 equal $V_{\diagup,4}$ from (1.7), we can write

$$W_3 = \tbinom{n}{2}^{-1/2} \sum_{a \in \mathcal{I}_2^n} \varphi_3(a/n)\, \psi_3(U_a) \prod_{v \overset{\diagup}{\sim} w} \left(Y_{a_v a_w} - \kappa(U_{a_v}, U_{a_w})\right), \qquad \varphi_3 \equiv 1,\ \psi_3(u_1, u_2) = u_1 u_2;$$

finally, letting W_4 equal $V_{\vee,1}$ from (1.7), we can write

$$W_4 = \tbinom{n}{3}^{-1/2} \sum_{a \in \mathcal{I}_3^n} \varphi_4(a/n)\, \psi_4(U_a) \prod_{v \overset{\vee}{\sim} w} \left(Y_{a_v a_w} - \kappa(U_{a_v}, U_{a_w})\right), \qquad \varphi_4 \equiv 1,\ \psi_4(u_1, u_2, u_3) = \kappa(u_1, u_3).$$

The other quantities in (1.7) can be represented in a similar manner. Applying Theorem 3.1, we obtain that the quantities in (1.6) and (1.7) are jointly close in distribution to independent Gaussian random variables.

Remark 3.3. Consider the case of $\mathbb{G}(n, \kappa)$; that is, the U_i are independent and distributed uniformly on $[0, 1]$. Assume φ_i and ψ_i are continuous almost everywhere; then, with $\bar{\psi}_i(u) = \psi_i(u) - \mathbb{E}\psi_i(U_1)$ if $k_i = 1$,

$$\lim_{n\to\infty} \sigma_{ij} = \begin{cases} \displaystyle\int_{[0,1]} \varphi_i(t)\varphi_j(t)\, dt \int_{[0,1]} \bar{\psi}_i(u)\bar{\psi}_j(u) du & \text{if } k_i = k_j = 1, \\ \displaystyle\int_{\mathcal{D}_k} \varphi_i(t)\varphi_j(t)\, dt \int_{[0,1]^k} \psi_i(u)\psi_j(u) \prod_{v \overset{F_i}{\sim} w} \kappa(u_v, u_w)\left(1 - \kappa(u_v, u_w)\right) du & \text{if } F_i = F_j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Therefore, the corresponding Gaussian stochastic measures Z_F are determined by the measure spaces

$$\left(\mathcal{D}_k \times [0,1]^k, dt \times \prod_{v \overset{F}{\sim} w} \kappa(u_v, u_w)\left(1 - \kappa(u_v, u_w)\right) du\right), \qquad F \in \mathcal{F},\ k = |F|,$$

and these measures are independent of each other.

Remark 3.4. Now consider $\mathbb{G}_{\mathrm{lat}}(n, \kappa)$. The case $k_i = 1$ is not interesting since $\sigma_{ij} = 0$ for all j. Moreover, we can assume without loss of generality that $\psi_i \equiv 1$. Assume that φ and κ are continuous almost everywhere. Then, if $F_i = F_j$, we have

$$\lim_{n\to\infty} \sigma_{ij} = \int_{\mathcal{D}_k} \varphi_i(t)\varphi_j(t) \prod_{v \overset{F}{\sim} w} \kappa(t_v, t_w)\left(1 - \kappa(t_v, t_w)\right) dt \tag{3.6}$$

and $\sigma_{ij} = 0$ otherwise. Therefore, the corresponding Gaussian stochastic measure Z_F is determined by the measure space

$$\left(\mathcal{D}_k, \prod_{v \overset{F}{\sim} w} \kappa(t_v, t_w)\left(1 - \kappa(t_v, t_w)\right) \, dt \right), \qquad F \in \mathcal{F},\, k = |F|\,,$$

and these measures are independent of each other.

3.2 Connection to fourth moment theorem

One might wonder why the limits of the centred subgraph count statistics for connected F turn out to be Gaussian. We believe that this is connected to the *Fourth Moment Theorem*, first proved by Nualart and Peccati (2005); see Nourdin and Peccati (2012) for a comprehensive discussion and proofs based on Stein's method.

The theorem can be formulated as follows. Let $(Z_h)_{h \in H}$ be a Gaussian Hilbert space defined on some probability space Ω, and let $\mathcal{F}$ be the sigma-algebra generated by that space. Let $F_n \in L_2(\Omega, \mathcal{F})$ with $\mathbb{E}F_n = 0$ and $\operatorname{Var} F_n = 1$ for all $n \geqslant 1$, and assume the F_n are elements of a fixed Wiener chaos. Then, F_n converges to a standard Gaussian distribution if and only if $\mathbb{E}F_n^4$ converges to 3.

Consider a Gaussian stochastic measure Z_2 on $\mathcal{D}_2$, where $\mathcal{D}_k$ is as before — we can think of Z_2 as the Gaussian approximation of the centred and scaled "edge-field" $\hat{Z}_n$ in (1.14). For $1 \leqslant i < j \leqslant n$, let

$$X_{ij} = \int_{\left[\frac{i-1}{n}, \frac{i}{n}\right] \times \left[\frac{j-1}{n}, \frac{j}{n}\right]} Z_2(dx, dy).$$

It is easy to see that the X_{ij} are independent and that

$$X_{ij} \sim \mathrm{N}(0, n^{-2}).$$

We can think of X_{ij} as a Gaussian version of the centred and scaled edge indicator between vertices i and j, where $i < j$.

Let $\varphi \in L_2(\mathcal{D}_3)$ and assume φ is continuous almost everywhere, and define $\bar{\varphi}_n \in L_2(\mathcal{D}_2 \times \mathcal{D}_2)$ as

$$\bar{\varphi}_n(x, y, u, v) = \varphi\left(\frac{\lceil nx \rceil}{n}, \frac{\lceil ny \rceil}{n}, \frac{\lceil nv \rceil}{n}\right) \mathrm{I}[\lceil ny \rceil / n = \lceil nu \rceil / n]\,.$$

Note that

$$\begin{aligned} \int_{\mathcal{D}_2} \int_{\mathcal{D}_2} & \bar{\varphi}_n(x, y, u, v)\, Z_2(dx, dy) Z_2(du, dv) \\ & = \sum_{i<j<k} \varphi\left(\frac{i}{n}, \frac{j}{n}, \frac{k}{n}\right) X_{ij} X_{jk} + \text{small boundary term} \end{aligned} \tag{3.7}$$

lives in the second Wiener chaos and has variance

$$\frac{1}{n^4} \sum_{i<j<k} \varphi\left(\frac{i}{n}, \frac{j}{n}, \frac{k}{n}\right)^2 \approx \frac{1}{n} \int_{\mathcal{D}_3} \varphi(x)^2 dx.$$

The sum on the right hand side of (3.7) is just the two-star count of the X_{ij}, and so centred subgraph counts of random graphs are in essence multiple stochastic integrals of the centred edge indicators.

Now, assume that $\int_{\mathcal{D}_3} \varphi(x)^2 dx = 1$ and consider

$$F_n = \sum_{i<j<k} \sqrt{n}\varphi\left(\frac{i}{n}, \frac{j}{n}, \frac{k}{n}\right) X_{ij}X_{jk};$$

we have $\operatorname{Var} F_n = 1$. It is not difficult to see that, if φ is continuous almost everywhere,

$$\begin{aligned}\mathbb{E}F_n^4 &= \sum_{i<j<k} n^2\varphi\left(\frac{i}{n}, \frac{j}{n}, \frac{k}{n}\right)^4 \times \frac{3}{n^8} \\ &\quad + 3\sum_{i<j<k} \sum_{\substack{u<v<w \\ (u,v,w)\neq(i,j,k)}} n^2\varphi\left(\frac{i}{n}, \frac{j}{n}, \frac{k}{n}\right)^2 \varphi\left(\frac{u}{n}, \frac{v}{n}, \frac{w}{n}\right)^2 \times \frac{1}{n^8} \\ &= 3\left(\int_{\mathcal{D}_3} \varphi(x)^2 dx\right)^2 + \mathrm{o}(1) = 3 + \mathrm{o}(1),\end{aligned}$$

and so, by the fourth moment theorem, F_n converges to a standard normal. The corresponding multivariate convergence can be made with similar arguments for any finite collection of such φ. Formally, we can therefore identify a Gaussian Hilbert Space on $L_2(\mathcal{D}_3)$ with

$$\int_{\mathcal{D}_3} \varphi(x_1, x_2, x_3)\, Z_2(dx_1, dx_2) Z_2(dx_2, dx_3), \qquad \varphi \in L_2(\mathcal{D}_3). \tag{3.8}$$

This argument is general, and in the same manner, we can think of Z_2 giving rise to a Gaussian Hilbert space on $L_2(\mathcal{D}_k)$ for every connected graph F on k vertices and identify it with the integral

$$\int_{\mathcal{D}_k} \varphi(x) \prod_{i\overset{F}{\sim}j} Z_2(dx_i, dx_j), \qquad \varphi \in L_2(\mathcal{D}_k). \tag{3.9}$$

However, it is important to keep in mind that the convergence of F_n is only distributional, so it is not clear whether the Gaussian Hilbert spaces (3.9) can be coupled with the underlying space Z_2 in a non-trivial and meaningful manner.

4 Abstract approximation theorem

The following abstract multivariate normal approximation theorem is based on Stein's method and can yield informative bounds even in the case of vectors of sums of uncorrelated, but not necessarily independent random variables. Let e_i be the i-th unit-vector in $\mathbb{N}^d$. A triple of d-dimensional vectors (W, W', G) is called *Stein coupling* if

$$\mathbb{E}\left\{G^t g(W') - G^t g(W)\right\} = \mathbb{E}\left\{W^t g(W)\right\}. \tag{4.1}$$

for all $g : \mathbb{R}^d \to \mathbb{R}^d$ for which the expectations exists; see Chen and Röllin (2010) and Fang and Röllin (2015).

Theorem 4.1. *Let (W, W', G) be a Stein coupling, and let $\Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant d} = \operatorname{Cov}(W)$; set $D = W' - W$. Then, for any $(p+2)$-times differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ and with Z being a centred Gaussian vector with covariance matrix Σ,*

$$\begin{aligned}&|\mathbb{E}h(W) - \mathbb{E}h(Z)| \\ &\leqslant \sum_{i=1}^d \left(\sum_{\alpha:1\leqslant|\alpha|\leqslant p} \frac{|h|_{\alpha+e_i}}{(|\alpha|+1)\alpha!} \sqrt{\operatorname{Var}\mathbb{E}\left\{G_i D^\alpha \mid W\right\}} \right. \\ &\quad \left. + \sum_{\alpha:2\leqslant|\alpha|\leqslant p} \frac{|h|_{\alpha+e_i}}{(|\alpha|+1)\alpha!} \left|\mathbb{E}\left\{G_i D^\alpha\right\}\right| + \sum_{\alpha:|\alpha|=p+1} \frac{|h|_{\alpha+e_i}}{(p+2)\alpha!} \mathbb{E}\left|G_i D^\alpha\right| \right).\end{aligned} \tag{4.2}$$

Remark 4.2. Note that, by Young's inequality, we can upper bound the last term in (4.2) as

$$\mathbb{E}\left|G_i D^\alpha\right| \leqslant \left\|G_i\right\|_\infty \sum_{j=1}^d \frac{\alpha_j}{p+1} \mathbb{E}\left|D_j^{p+1}\right|. \tag{4.3}$$

Proof of Theorem 4.1. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a solution to the Stein's equation

$$\sum_{i,j=1}^d \sigma_{ij}\partial_{ij} g(z) - \sum_{i=1}^d z_i \partial_i g(z) = h(z) - \mathbb{E}h(Z), \qquad z \in \mathbb{R}^d. \tag{4.4}$$

From Meckes (2009, Eq. (10)), it is immediate that

$$\left|g\right|_\alpha \leqslant \frac{1}{|\alpha|}\left|h\right|_\alpha \tag{4.5}$$

and it is therefore enough to bound

$$\mathbb{E}\left\{\sum_{i,j=1}^d \sigma_{ij}\partial_{ij} g(W) - \sum_{i=1}^d W_i \partial_i g(W)\right\}$$

in order to bound the left hand side of (4.2). By Taylor's theorem for multivariate functions,

$$f(w') - f(w) = \sum_{l=1}^p \sum_{\alpha:|\alpha|=l} \frac{\partial^\alpha f(w)}{\alpha!}(w'-w)^\alpha + R^{(p+1)}(w', w)$$

and

$$R^{(p+1)}(w', w) = \sum_{\alpha:|\alpha|=p+1} \frac{p+1}{\alpha!}(w'-w)^\alpha \int_0^1 (1-s)^p \partial^\alpha f(w + s(w'-w))ds.$$

Now, using (4.1), we have

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{i=1}^d W_i \partial_i g(W)\right\} \\
&\quad = \mathbb{E}\left\{\sum_{i=1}^d G_i \left(\partial_i g(W') - \partial_i g(W)\right)\right\} \\
&\quad = \mathbb{E}\left\{\sum_{i=1}^d G_i \sum_{l=1}^p \sum_{\alpha:|\alpha|=l} \frac{\partial^{\alpha+e_i} g(W)}{\alpha!} D^\alpha\right\} \\
&\qquad + \mathbb{E}\left\{\sum_{i=1}^d G_i \sum_{\alpha:|\alpha|=p+1} \frac{p+1}{\alpha!} D^\alpha \int_0^1 (1-s)^p \partial^{\alpha+e_i} g(W + sD)ds\right\} =: r_1 + r_2.
\end{aligned}$$

Now,

$$\begin{aligned}
r_1 &= \sum_{l=1}^p \mathbb{E}\left\{\sum_{i=1}^d \sum_{\alpha:|\alpha|=l} \frac{\partial^{\alpha+e_i} g(W)}{\alpha!} G_i D^\alpha\right\} \\
&= \sum_{l=1}^p \mathbb{E}\left\{\sum_{i=1}^d \sum_{\alpha:|\alpha|=l} \frac{\partial^{\alpha+e_i} g(W)}{\alpha!} \left(\mathbb{E}\left\{G_i D^\alpha \mid W\right\} - \mathbb{E}\left\{G_i D^\alpha\right\}\right)\right\} \\
&\quad + \sum_{l=1}^p \mathbb{E}\left\{\sum_{i=1}^d \sum_{\alpha:|\alpha|=l} \frac{\partial^{\alpha+e_i} g(W)}{\alpha!} \mathbb{E}\left\{G_i D^\alpha\right\}\right\} =: r_{1,1} + r_{1,2}.
\end{aligned}$$

First,

$$|r_{1,1}| \leqslant \sum_{l=1}^{p} \sum_{i=1}^{d} \sum_{\alpha:|\alpha|=l} \frac{|g|_{\alpha+e_i}}{\alpha!} \mathbb{E}\left|\mathbb{E}\left\{G_i D^\alpha \mid W\right\} - \mathbb{E}\left\{G_i D^\alpha\right\}\right|.$$

Next, recalling that $\sigma_{ij} = \mathbb{E}\{G_i D_j\}$,

$$\left| r_{1,2} - \mathbb{E} \sum_{i,j=1}^{d} \sigma_{ij} \partial_{ij} g(W) \right| \leqslant \sum_{l=2}^{p} \sum_{i=1}^{d} \sum_{\alpha:|\alpha|=l} \frac{|g|_{\alpha+e_i}}{\alpha!} \left|\mathbb{E}\left\{G_i D^\alpha\right\}\right|,$$

A bound on r_2 can be obtained in a similar manner, and so

$$\begin{aligned} &\left| \mathbb{E}\left\{ \sum_{i,j=1}^{d} \sigma_{ij} \partial_{ij} g(W) - \sum_{i=1}^{d} W_i \partial_i g(W) \right\} \right| \\ &\quad \leqslant \sum_{i=1}^{d} \left(\sum_{\alpha:1\leqslant|\alpha|\leqslant p} \frac{|g|_{\alpha+e_i}}{\alpha!} \sqrt{\operatorname{Var} \mathbb{E}\left\{G_i D^\alpha \mid W\right\}} + \sum_{\alpha:2\leqslant|\alpha|\leqslant p} \frac{|g|_{\alpha+e_i}}{\alpha!} \left|\mathbb{E}\left\{G_i D^\alpha\right\}\right| \right. \\ &\qquad \left. + \sum_{\alpha:|\alpha|=p+1} \frac{|g|_{\alpha+e_i}}{\alpha!} \mathbb{E}\left|G_i D^\alpha\right| \right). \end{aligned} \tag{4.6}$$

Applying (4.5), the claim follows. □

5 Proof of Theorem 3.1

Fix $d \geqslant 1$, and for each $1 \leqslant i \leqslant d$, let F_i be a connected graph on the vertex set $[k_i]$. Let $U = (U_v)_{1\leqslant v\leqslant n}$ be independent random variables, let κ be a graphon, and let $Y = (Y_{vw})_{1\leqslant v<w\leqslant n}$ be random variables that are independent conditionally on U and such that $\mathbb{E}\{Y_{vw} \mid U_v, U_w\} = \kappa(U_v, U_w)$. For $1 \leqslant i \leqslant d$ and $a \in \mathcal{I}_{k_i}^n$, let

$$T_{i,a} = \prod_{v \overset{F_i}{\sim} w} \left(Y_{a_v a_w} - \kappa(U_{a_v}, U_{a_w})\right) \tag{5.1}$$

if $k_i \geqslant 2$, and for convenience, set $T_{i,a} = 1$ if $k_i = 1$. For each $1 \leqslant i \leqslant d$, let $\psi_i : [0,1]^{k_i} \to \mathbb{R}$ be a bounded function, and for $a \in \mathcal{I}_{k_i}^n$, let

$$\Phi_{i,a} = \begin{cases} \psi_i(U_{a_1}) - \mathbb{E}\psi_i(U_{a_1}) & \text{if } k_i = 1, \\ \psi_i(U_a) & \text{if } k_i \geqslant 2. \end{cases}$$

Now, for $a \in \mathcal{I}_{k_i}^n$, let

$$X_{i,a} = \binom{n}{k_i}^{-1/2} \varphi_i(a/n) \Phi_{i,a} T_{i,a}.$$

Recalling the definition of $W = (W_1, \ldots, W_d)$ from (3.1), we have

$$W_i = \sum_{a \in \mathcal{I}_{k_i}^n} X_{i,a}.$$

5.1 Stein Coupling

Let $1 \leqslant i \leqslant d$ and $a \in \mathcal{I}_{k_i}^n$. For $1 \leqslant j \leqslant d$, let

$$N_j^{i,a} = \{b \in \mathcal{I}_{k_j}^n : |a \cap b| \geqslant 2 \wedge k_i\},$$

where in the expression $a \cap b$, the ordered tuples a and b are interpreted as unordered sets. Note that if $k_i \geqslant 2$ and $k_j = 1$ then $N_j^{i,a} = \emptyset$. Let

$$W_j^{i,a} = W_j - \sum_{b \in N_j^{i,a}} X_{i,b}.$$

Let I be uniformly distributed on $[d]$ and independent of all else, and given I, let A be uniformly distributed on $\mathcal{I}_{k_I}^n$. Let

$$W' = W^{I,A} = \left(W_1^{I,A}, \ldots, W_d^{I,A}\right), \qquad G = -d\binom{n}{k_I} X_{I,A}\, e_I,$$

where e_i is the i-th unit vector in $\mathbb{R}^d$.

Lemma 5.1. *(W, W', G) is a d-dimensional Stein coupling.*

Proof. Write $g(x) = (g_1(x), \ldots, g_d(x))$; averaging over I and A,

$$\mathbb{E}\left\{G^t g(W')\right\} = -\sum_{i=1}^{d} \sum_{a \in \mathcal{I}_{k_i}^n} \binom{n}{k_i}^{-1/2} \varphi_i(a/n) \mathbb{E}\left\{\Phi_{i,a} T_{i,a} g_i(W^{i,a})\right\}. \tag{5.2}$$

If $k_i = 1$, then $W^{i,a}$ does not contain any information about U_a, and since $T_{i,a} = 1$ and $\mathbb{E}\Phi_{i,a} = 0$, it follows that $\mathbb{E}\left\{\Phi_{i,a} T_{i,a} g_i(W^{i,a})\right\} = 0$. If $k_i \geqslant 2$, then conditionally on U, $W^{i,a}$ does not contain any information about $(Y_{vw})_{v,w \in a}$. Since $\mathbb{E}\left\{T_{i,a} \mid U\right\} = 0$ it again follows that $\mathbb{E}\left\{\Phi_{i,a} T_{i,a} g_i(W^{i,a})\right\} = 0$. Hence $\mathbb{E}\left\{G^t g(W')\right\} = 0$. It is straightforward to check that $-\mathbb{E}\left\{G^t g(W)\right\} = \mathbb{E}\left\{W^t g(W)\right\}$. □

5.2 Estimates on mixed moments

Before proving the main theorem, we present some lemmas, which will be used in the proof of Theorem 3.1. For a graph F on the vertex set $[k]$ and $b \in \mathcal{I}_k^n$, denote by $F(b)$ the graph on the vertex set $\{b_1, \ldots, b_k\}$ where b_v and b_w are connected in $F(b)$ if and only if v and w are connected in F. In other words, $F(b)$ is the induced graph when mapping vertex v to vertex b_v for all $v \in [k]$. We assume throughout that, for each $1 \leqslant i \leqslant d$, F_i is a connected graph on the vertex set $[k_i]$, where $k_i \geqslant 1$. The reader should keep in mind that the bounds obtained in Lemmas 5.3–5.5 are *worst-case* bounds, and will typically be sharp if all graphs involved are line graphs, but depending on the combinatorics of the F_i, the bounds could be much smaller. Phrases like "there are $\mathrm{O}(n^k)$ choices" have to be understood in the context of the usual Bachmann–Landau notation, which in this case means that the number of choices can be "of order n^k or of *smaller order*".

Lemma 5.2 (c.f. Janson and Nowicki (1991, Lemma 5))**.** *Let $m \geqslant 2$, and for each $1 \leqslant l \leqslant m$, let $1 \leqslant i_l \leqslant d$, and let $b_l \in \mathcal{I}_{k_{i_l}}^n$. Assume*

$$\mathbb{E}\left\{\prod_{l=1}^{m} \Phi_{i_l,b_l} \prod_{l=1}^{m} T_{i_l,b_l}\right\} \neq 0. \tag{5.3}$$

Then, every vertex and every edge belong to at least two of the subgraphs

$$F_{i_1}(b_1), \cdots, F_{i_m}(b_m). \tag{5.4}$$

Moreover, the subgraphs (5.4) *either coincide in $m/2$ disjoint pairs (m necessarily even) or there is a vertex that belongs to at least three of them.*

Proof. Without loss of generality, assume there is $m' \leqslant m$ such that $k_{i_l} = 1$ for all $l > m'$ (if there are no such indices, set $m' = m$). So, assume

$$\mathbb{E}\left\{\prod_{l=1}^{m} \Phi_{i_l,b_l} \mathbb{E}\left\{\prod_{l=1}^{m'} T_{i_l,b_l} \mid U\right\}\right\} \neq 0.$$

Suppose there is an edge between v and w in a subgraph that is not in any other subgraph, so that the factor $Y_{vw} - \kappa(U_v, U_w)$ appears exactly once in $\prod_{l=1}^{m'} T_{i_l,b_l}$. Since the Y_{vw} are conditionally independent given U and since $\mathbb{E}\{Y_{vw} - \kappa(U_v, U_w) \mid U\} = 0$, it would follow that $\mathbb{E}\left\{\prod_{l=1}^{m'} T_{i_l,b_l} \mid U\right\} = 0$, which contradicts the claim. Also, as a consequence, every vertex among the subgraphs that has at least one edge attached to it, must also appear in another subgraph. Suppose now there is an isolated vertex v in a subgraph, say $F_{i_l}(b_l)$ for some $l > m'$, that is not in any other subgraph. In that case, U_v only appears in Φ_{i_l,b_l} and $\mathbb{E}\left\{\prod_{l=1}^{m'} T_{i_l,b_l} \mid U\right\}$ does not depend on U_v. Due to the fact that $\mathbb{E}\Phi_{i_l,b_l} = 0$ for such $F_{i_l}(b_l)$ and independence, the left hand side of (5.3) would equal zero, again in contradiction to the claim. This concludes the proof of the first assertion.

To prove the second assertion, assume each vertex appears in exactly two of the $F_{i_l}(b_l)$. If a vertex is in $F_{i_l}(b_l)$ and $F_{i_{l'}}(b_{l'})$, say, then all edges attached to it, must also be in $F_{i_l}(b_l)$ and $F_{i_{l'}}(b_{l'})$, and so forth. Since both graphs are connected, they must coincide. Hence, the $F_{i_l}(b_l)$ must come in identical pairs. □

Lemma 5.3. *Let $1 \leqslant i, i_1, \ldots, i_m \leqslant d$ for some $m \geqslant 2$. Then there exists a constant $C > 0$ that is independent of n such that*

$$\left|\mathbb{E} \sum_{a \in \mathcal{I}_{k_i}^n} \sum_{b_1 \in N_{i_1}^{i,a}} \cdots \sum_{b_m \in N_{i_m}^{i,a}} X_{i,a} X_{i_1,b_1} \cdots X_{i_m,b_m}\right| \leqslant C n^{-(m-1)/2}. \tag{5.5}$$

Proof. First, write the expectation on the left hand side of (5.5) as

$$\begin{aligned} \xi \coloneqq & \frac{1}{\binom{n}{k_i}^{1/2} \times \binom{n}{k_{i_1}}^{1/2} \times \cdots \times \binom{n}{k_{i_m}}^{1/2}} \\ & \times \sum_{a \in \mathcal{I}_{k_i}^n} \sum_{b_1 \in N_{i_1}^{i,a}} \cdots \sum_{b_m \in N_{i_m}^{i,a}} \mathbb{E}\left\{\Phi_{i,a} T_{i,a} \Phi_{i_1,b_1} T_{i_1,b_1} \cdots \Phi_{i_m,b_m} T_{i_m,b_m}\right\}. \end{aligned} \tag{5.6}$$

Fix $a, b_1, \ldots, b_m$ and consider the induced subgraphs

$$F_i(a), F_{i_1}(b_1), \ldots, F_{i_m}(b_m), \tag{5.7}$$

which are subgraphs on the vertex set $[n]$. By Lemma 5.2, if the corresponding expectation of the summand in (5.6) is non-zero, then either these subgraphs coincide in pairs of disjoint subgraphs, or all vertices and edges appear in at least two subgraphs while at least one vertex appears in three. Note that all the subgraphs share vertices with $F_i(a)$ by the definition of $N_j^{i,a}$, and thus can coincide in distinct pairs only if $m = 1$, which is excluded.

Assume $k_i \geqslant 2$, and recall that each of the $F_{i_l}(b_l)$ shares at least two vertices with $F_i(a)$. Also, note that $F_i(a)$ has k_i vertices and so $\sum_{l=1}^m k_{i_l}$ must be at least k_i in order for every vertex in $F_i(a)$ to also be in one of the other subgraphs. However, if $m \geqslant 2$, $\sum_{l=1}^m k_{i_l}$ must be larger than k_i to also cover all *edges* of $F_i(a)$, of which there are at least $k_i - 1$; indeed, if a vertex of $F_i(a)$ has two edges attached to it and the two edges are contained in different subgraphs, say one in $F_{i_1}(b_1)$ and the other in $F_{i_2}(b_2)$, then

that vertex must belong to all three subgraphs. Therefore, if $\sum_{l=1}^m k_{i_l} < k_i + m - 1$, it is not possible that each edge of $F_i(a)$ also belongs to one of the other subgraphs, and so all terms in (5.6) vanish, that is, $\xi = 0$, and the claim is trivially true.

If $\sum_{l=1}^m k_{i_l} = k_i + m - 1$, the sum (5.6) contains at most $\mathrm{O}(n^{k_i})$ non-zero terms, since all vertices of $F_{i_1}(b_1), \dots, F_{i_m}(b_m)$ must coincide with vertices of $F_i(a)$ to cover all of the latter, and this arrangement contributes only a combinatorial factor to the sum that is independent of n. Thus,

$$|\xi| \leqslant \frac{Cn^{k_i}}{n^{k_i/2} n^{(k_{i_1}+\cdots+k_{i_m})/2}} \leqslant \frac{C}{n^{(m-1)/2}},$$

where the second inequality follows from the fact that $\sum_{l=1}^m k_{i_l} = k_i + m - 1$.

If $\sum_{l=1}^m k_{i_l} > k_i + m - 1$, let $q := \sum_{l=1}^m k_{i_l} - (k_i + m - 1)$. Note that q is the maximal number of vertices available among $F_{i_1}(b_1), \dots, F_{i_m}(b_m)$ that do not need to overlap with $F_i(a)$ (there might be fewer that can be chosen outside of $F_i(a)$, but in any case, never *more*). Assume first q is even. Since every vertex must be contained in at least two subgraphs, there are $q/2$ additional free choices in (5.6), contributing a factor of $\mathrm{O}(n^{q/2})$ to the sum, so that

$$|\xi| \leqslant \frac{Cn^{k_i+q/2}}{n^{k_i/2} n^{(k_{i_1}+\cdots+k_{i_m})/2}} \leqslant \frac{C}{n^{(m-1)/2}},$$

where the second inequality follows from the fact that $\sum_{l=1}^m k_{i_l} = k_i + q + m - 1$. If q is odd, one of the q vertices cannot be chosen freely, so that the additional factor appearing in $\mathrm{O}(n^{(q-1)/2})$, and we obtain

$$|\xi| \leqslant \frac{Cn^{k_i+(q-1)/2}}{n^{k_i/2} n^{(k_{i_1}+\cdots+k_{i_m})/2}} \leqslant \frac{C}{n^{(m-1)/2+1/2}} \leqslant \frac{C}{n^{(m-1)/2}}.$$

If $k_i = 1$, coverage of $F_i(a)$ is always guaranteed, since every $F_{i_l}(b_l)$ overlaps with the one vertex of $F_i(a)$. Hence, with $q = \sum_{l=1}^m k_{i_l} - m$, there are at most $q/2$ vertices which can be chosen freely if q is even and $(q-1)/2$ if q is odd, contributing a factor of no more than $\mathrm{O}(n^{q/2})$ to (5.6), so that

$$|\xi| \leqslant \frac{Cn^{1+q/2}}{n^{1/2} n^{(k_{i_1}+\cdots+k_{i_m})/2}} \leqslant \frac{Cn^{1/2}}{n^{m/2}}.$$

This concludes the proof. □

Lemma 5.4. *Let $1 \leqslant i, j \leqslant d$ and let $m \geqslant 2$. Then there exists a constant $C > 0$ that is independent of n such that*

$$\left| \mathbb{E} \sum_{a \in \mathcal{I}_{k_i}^n} \sum_{b_1 \in N_j^{i,a}} \cdots \sum_{b_m \in N_j^{i,a}} X_{j,b_1} \cdots X_{j,b_m} \right| \leqslant \begin{cases} Cn^{1-m/2} & \text{if } k_i = 1, \\ Cn^{k_i - m} & \text{if } k_i \geqslant 2. \end{cases} \tag{5.8}$$

Proof. First, write

$$\xi := \frac{1}{\binom{n}{k_j}^{m/2}} \sum_{a \in \mathcal{I}_{k_i}^n} \sum_{b_1 \in N_j^{i,a}} \cdots \sum_{b_m \in N_j^{i,a}} \mathbb{E}\left\{\Phi_{j,b_1} T_{j,b_1} \cdots \Phi_{j,b_m} T_{j,b_m}\right\} \tag{5.9}$$

Fix $a, b_1, \dots, b_m$ and consider the induced subgraphs

$$F_j(b_1), \dots, F_j(b_m), \tag{5.10}$$

which are subgraphs on the set $[n]$. By Lemma 5.2, if the corresponding summand in (5.9) is non-zero, every vertex must appear in at least two of these subgraphs.

Assume $k_i \geqslant 2$, and $k_j \geqslant 2$ (if $k_j = 1$, then $N_j^{i,a} = \emptyset$ and the claim is trivially true). There are $\mathrm{O}(n^{k_i})$ choices for a and since each of the $F_j(b_l)$ must have two vertices in the set a, we can assign $k_j - 2$ vertices freely for each such subgraph, subject to the condition that each vertex appears twice. With $q = m(k_j - 2)$, there are $\mathrm{O}(n^{q/2})$ choices if q is even. Hence

$$|\xi| \leqslant \frac{Cn^{k_i+q/2}}{n^{mk_j/2}} \leqslant Cn^{k_i-m}. \tag{5.11}$$

If q is odd, there are $\mathrm{O}(n^{(q-1)/2})$ choices, hence

$$|\xi| \leqslant \frac{Cn^{k_i+(q-1)/2}}{n^{mk_j/2}} \leqslant Cn^{k_i-m-1/2} \leqslant Cn^{k_i-m}. \tag{5.12}$$

In the case $k_i = 1$, similar arguments lead to the estimate

$$|\xi| \leqslant \frac{Cn^{1+m(k_j-1)/2}}{n^{mk_j/2}} \leqslant Cn^{1-m/2}, \tag{5.13}$$

if $m(k_j - 1)$ is even, and similarly if it is odd. This concludes the proof. □

Lemma 5.5. *Let $1 \leqslant i, i_1, \dots, i_m \leqslant d$ for some $m \geqslant 1$. Then there exists a constant $C > 0$ that is independent of n such that*

$$\Bigg| \sum_{a,a' \in \mathcal{I}_{k_i}^n} \sum_{b_1 \in N_{i_1}^{i,a}} \sum_{b_1' \in N_{i_1}^{i,a'}} \cdots \sum_{b_m \in N_{i_m}^{i,a}} \sum_{b_m' \in N_{i_m}^{i,a'}} \operatorname{Cov}\left(X_{i,a}X_{i_1,b_1}\cdots X_{i_m,b_m}, X_{i,a'}X_{i_1,b_1'}X_{i_m,b_m'}\right)\Bigg| \leqslant Cn^{-m}. \tag{5.14}$$

Proof. First, let ξ equal the left hand side of (5.14) without modulus. Consider first the case $k_i \geqslant 2$, in which case again we may assume $k_{i_l} \geqslant 2$ for all $1 \leqslant l \leqslant m$, since otherwise $\xi = 0$, using the same arguments as in the previous lemmas. Now, it is easy to verify that the covariances are zero if the two sets

$$a \cup \bigcup_{l=1}^m b_l, \quad \text{and} \quad a' \cup \bigcup_{l=1}^m b_l' \tag{5.15}$$

do not overlap (independence). Fix $a, a', b_1, b_1', \dots, b_m, b_m'$, consider the induced subgraphs

$$F_i(a), F_{i_1}(b_1), \dots, F_{i_m}(b_m), F_i(a'), F_{i_1}(b_1'), \dots, F_{i_m}(b_m'), \tag{5.16}$$

which are subgraphs on the set $[n]$, and also consider

$$\mathbb{E}\left\{X_{i,a}X_{i_1,b_1}\cdots X_{i_m,b_m} \cdot X_{i,a'}X_{i_1,b_1'}\cdots X_{i_m,b_m'}\right\}. \tag{5.17}$$

By Lemma 5.2, if (5.17) is non-zero, every vertex in (5.16) must appear in at least two of these subgraphs. Now, let $r = |a \cap a'|$.

Case $1 \leqslant r \leqslant k_i$: Since r vertices in $F_i(a)$ are also in $F_i(a')$, both $F_i(a)$ and $F_i(a')$ have $k_i - r$ more vertices each that need to be in any of the other subgraphs and, since F_i is connected, also at least $k_i - r$ more edges each. We proceed similarly as in the proof of Lemma 5.3. If $\sum_{l=1}^m k_{i_l} < k_i - r + m$, it is not possible for all edges of $F_i(a \setminus a')$ and those connecting $F_i(a \setminus a')$ with $F_i(a \cap a')$, to be covered, and so $\xi = 0$. Otherwise, let $q = \sum_{l=1}^m k_{i_l} - (k_i - r + m)$. There are $\mathrm{O}(n^{2k_i-r})$ choices for the vertices of $F_i(a)$ and $F_i(a')$ together, and there are $\mathrm{O}(n^{(2q)/2})$ choices for the remaining vertices. Hence,

$$|\xi| \leqslant \frac{Cn^{2k_i-r+q}}{n^{k_i+k_{i_1}+\cdots k_{i_m}}} \leqslant \frac{Cn^{2k_i-r+q}}{n^{k_i+q+k_i-r+m}} \leqslant \frac{C}{n^m},$$

where we have used that $\sum_{l=1}^m k_{i_l} = q + k_i - r + m$

Case $r = 0$: $F_i(a)$ and $F_i(a')$ are not overlapping and each of $F_i(a)$ and $F_i(a')$ have k_i vertices that need to be in any of the other subgraphs and, since F_i is connected, also at least $k_i - 1$ edges. If $\sum_{l=1}^m k_{i_l} < k_i - 1 + m$, it is not possible for all edges of $F_i(a)$ and $F_i(a')$, respectively, to be covered, and so $\xi = 0$. Otherwise, let $q = \sum_{l=1}^m k_{i_l} - (k_i - 1 + m)$. There are $\mathrm{O}(n^{2k_i})$ choices for the vertices of $F_i(a)$ and $F_i(a')$ together, and there are $\mathrm{O}(n^{(2q)/2-1})$ choices for the remaining vertices, since at least one vertex from $\bigcup_{l=1}^m b_l$ must overlap with $\bigcup_{l=1}^m b'_l$. Hence,

$$|\xi| \leqslant \frac{Cn^{2k_i+q-1}}{n^{k_i+k_{i_1}+\cdots k_{i_m}}} \leqslant \frac{Cn^{2k_i+q-1}}{n^{k_i+q+k_i-1+m}} \leqslant \frac{C}{n^m},$$

where we have used that $\sum_{l=1}^m k_{i_l} = q + k_i - 1 + m$.

Now suppose $k_1 = 1$. If $F_i(a) = F_i(a')$ there are n choices for this one vertex, and since every subgraph must share a vertex with $F_i(a)$ and $F_i(a')$, respectively, there are $\mathrm{O}\big(n^{2\times\sum_{l=1}^m (k_{i_l}-1)/2}\big)$ choices for the remaining vertices. Hence,

$$|\xi| \leqslant \frac{Cn^{1+k_{i_1}+\cdots+k_{i_m}-m}}{n^{1+k_{i_1}+\cdots k_{i_m}}} \leqslant \frac{C}{n^m}.$$

If $F_i(a) \neq F_i(a')$ there are $\mathrm{O}(n^2)$ choices for the two vertices, and since every subgraph must share a vertex with $F_i(a)$ and $F_i(a')$, respectively, there are $\mathrm{O}\big(n^{2\times\sum_{l=1}^m (k_{i_l}-1)/2-1}\big)$ choices for the remaining vertices, since at least one vertex from $\bigcup_{l=1}^m b_l$ must overlap with $\bigcup_{l=1}^m b'_l$. Hence

$$|\xi| \leqslant \frac{Cn^{2+k_{i_1}+\cdots+k_{i_m}-m-1}}{n^{1+k_{i_1}+\cdots k_{i_m}}} \leqslant \frac{C}{n^m}.$$

This concludes the proof. □

Lemma 5.6. *Let* $m \geqslant 2$. *Then*

$$\big|\mathbb{E}D_j^m\big| \leqslant Cn^{-m}, \qquad |\mathbb{E}\{G_i D^\alpha\}| \leqslant Cn^{-(|\alpha|-1)/2}, \quad \operatorname{Var}\mathbb{E}\{G_i D^\alpha \mid U, Y\} \leqslant Cn^{-|\alpha|}.$$

Proof. Note that

$$D_j = W'_j - W_j = -\sum_{b\in N_j^{I,A}} X_{j,b},$$

and hence,

$$\mathbb{E}\{G_i D^\alpha \mid U, Y\} = (-1)^{|\alpha|+1} \sum_{a\in\mathcal{I}_{k_i}^n} X_{i,a} \prod_{j=1}^d \left(\sum_{b\in N_j^{i,a}} X_{j,b}\right)^{\alpha_j}$$

and

$$\mathbb{E}\big\{D_j^m \mid U, Y\big\} = \frac{(-1)^m}{d} \sum_{i=1}^d \binom{n}{k_i}^{-1} \sum_{a\in\mathcal{I}_{k_i}^n} \left(\sum_{b\in N_j^{i,a}} X_{j,b}\right)^m.$$

The bounds are now a direct consequence of Lemmas 5.3–5.5. □

5.3 Proof of Theorem 3.1

Proof. The variance expressions (3.2) are straightforward to establish. The proof of the bounds (3.3) for $(p+2)$-times differentiable functions is a consequence of Theorem 4.1 and Remark 4.2 with the Stein coupling from Lemma 5.1, along with the moment estimates of Lemma 5.6 with the choice $m = p + 1$.

We use the smoothing technique of Gan, Röllin, and Ross (2017) in order to approximate the indicator function I_A by a $(p+2)$-times partially differentiable function. Fix $A \in \mathcal{K}$ and $\varepsilon > 0$, define

$$A^{\varepsilon} = \{y \in \mathbb{R}^d : d(y, A) < \varepsilon\}, \quad \text{and} \quad A^{-\varepsilon} = \{y \in \mathbb{R}^d : B(y; \varepsilon) \subseteq A\},$$

where $d(y, A) = \inf_{x \in A} |x - y|$ and $B(y; \epsilon)$ is the closed ball of radius ε around y. Let $\{h_{\varepsilon,A} : \mathbb{R}^d \to [0,1]; A \in \mathcal{K}\}$ be a class of functions, such that $h_{\varepsilon,A}(x) = 1$ for $x \in A$ and 0 for $x \notin A^{\varepsilon}$. Then, by Lemma 2.1 of Bentkus (2003), we have for any $\varepsilon > 0$ that

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)| \leqslant 4d^{1/4}\varepsilon + \sup_{A \in \mathcal{K}} |\mathbb{E}h_{\varepsilon,A}(W) - \mathbb{E}h_{\varepsilon,A}(Z)|. \tag{5.18}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a bounded and Lebesgue measurable function, and for $\delta > 0$, consider the smoothing operator S_δ defined as

$$(S_\delta f)(x) = \frac{1}{(2\delta)^d} \int_{x_1-\delta}^{x_1+\delta} \cdots \int_{x_d-\delta}^{x_d+\delta} f(z) dz_d \dots dz_1.$$

Choose $\delta = \frac{\varepsilon}{(p+3)^2\sqrt{d}}$, let $h_{\varepsilon,A} = S_\delta^{p+3} I_{A^{\varepsilon/(p+3)}}$; then by Lemma 3.9 of Gan et al. (2017), $h_{\varepsilon,A}$ is $(p+2)$-times partially differentiable and

$$\|h_{\varepsilon,A}\|_\infty \leqslant 1, \qquad |h_{\varepsilon,A}|_\alpha \leqslant \frac{1}{\varepsilon^{|\alpha|}}, \qquad 1 \leqslant |\alpha| \leqslant p+2.$$

Note that $h_{\varepsilon,A}(x) = 1$ for $x \in A$ and $h_{\varepsilon,A}(x) = 0$ for $x \notin A^{\varepsilon}$. Therefore, from (3.3),

$$|\mathbb{E}h_{\varepsilon,A}(W) - \mathbb{E}h_{\varepsilon,A}(Z)| \leqslant \frac{C \sup_{\alpha:|\alpha|\leqslant p+2} |h_{\varepsilon,A}|_\alpha}{n^{1/2}} \leqslant \frac{C}{n^{1/2}\varepsilon^{(p+2)}} \tag{5.19}$$

for some constant C. Now, using (5.18), we have

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W \in A) - \mathbb{P}(Z \in A)| \leqslant 4d^{1/4}\varepsilon + C\frac{1}{\varepsilon^{p+2}n^{1/2}}. \tag{5.20}$$

The final order $n^{-1/(2(p+2))}$ is then established by taking $\varepsilon = n^{-1/(2(p+2))}$. □

6 Proof of Lemmas 2.1 and 2.2

Consider the graph F on the vertex set $[k]$ as fixed. In what follows, for any subgraph $H \subseteq F$, H^c denotes the 'edge complement' of H and is the graph obtained by removing from F all edges which are present in H and then removing all, if any, resulting isolated vertices.

Lemma 6.1. *Recalling* (2.16)*, and with F any graph on the vertex set $[k]$, we can write*

$$\prod_{i \overset{F}{\sim} j} y_{ij} = \sum_{H \subseteq' F} \rho_{F,H}(u, y), \tag{6.1}$$

where

$$\rho_{F,H}(u, y) = \prod_{i \overset{H^c}{\sim} j} \kappa(u_i, u_j) \times \vartheta_H(u, y), \qquad u \in [0,1]^k,\ y \in \mathbb{R}^{\binom{k}{2}}.$$

and where empty products are understood to equal 1*.*

Proof. We use induction over the number of edges in the graph F. If F has no edges, the claim is clearly true, since $H = \emptyset \subseteq F$ is the only subgraph of F without isolated vertices and in that case, $\prod_{i \overset{F}{\sim} j} y_{ij} = 1 = \rho_{F,\emptyset}(u, y)$.

Now, assume the assertion is true for all graphs with $e-1$ or fewer edges. Let F be a graph on k vertices with e edges. Fix an edge in F, say the edge between vertices l and m, where $1 \leqslant l < m \leqslant k$, and let F_{lm} be the subgraph of F obtained by removing that edge and any isolated vertex after the edge removal. Then

$$\prod_{i \overset{F}{\sim} j} y_{ij} = y_{lm} \prod_{i \overset{F_{lm}}{\sim} j} y_{ij} = (y_{lm} - \kappa(u_l, u_m)) \prod_{i \overset{F_{lm}}{\sim} j} y_{ij} + \kappa(u_l, u_m) \prod_{i \overset{F_{lm}}{\sim} j} y_{ij}$$

and by our assumption the decomposition holds for the subgraph F_{lm}, that is

$$\prod_{i \overset{F_{lm}}{\sim} j} y_{ij} = \sum_{H \subseteq' F_{lm}} \prod_{\{i,j\} \in E(F_{lm}) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)).$$

Thus we get

$$\begin{aligned}
\prod_{i \overset{F}{\sim} j} y_{ij} &= (y_{lm} - \kappa(u_l, u_m)) \sum_{H \subseteq' F_{lm}} \prod_{\{i,j\} \in E(F_{lm}) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)) \\
&\quad + \kappa(u_l, u_m) \sum_{H \subseteq' F_{lm}} \prod_{\{i,j\} \in E(F_{lm}) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)) \\
&= \sum_{H \subseteq' F_{lm}} \prod_{\{i,j\} \in E(F_{lm}) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)) (y_{lm} - \kappa(u_l, u_m)) \\
&\quad + \sum_{H \subseteq' F_{lm}} \prod_{\{i,j\} \in E(F_{lm}) \setminus E(H)} \kappa(u_i, u_j) \times \kappa(u_l, u_m) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)) \qquad (6.2) \\
&= \sum_{\substack{H \subseteq' F: \\ \{l,m\} \in E(H)}} \prod_{\{i,j\} \in E(F) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)) \\
&\quad + \sum_{\substack{H \subseteq' F: \\ \{l,m\} \notin E(H)}} \prod_{\{i,j\} \in E(F) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)) \\
&= \sum_{H \subseteq' F} \prod_{\{i,j\} \in E(F) \setminus E(H)} \kappa(u_i, u_j) \times \prod_{i \overset{H}{\sim} j} (y_{ij} - \kappa(u_i, u_j)).
\end{aligned}$$

Hence, the assertion is true for F, which completes the proof. □

Proof of Lemma 2.1. By Lemma 6.1,

$$t_F^{\mathrm{inj}}(G_n) = \sum_{H \subseteq' F} s_H(U, Y),$$

where

$$s_H(U, Y) = \frac{1}{(n)_k} \sum_{a \in \mathcal{A}_k^n} \rho_H(U_a, Y_a)$$

(we drop dependence on F, since it is fixed). Now, for $A \subset [k]$ (including the empty set), let

$$M_A = \{\psi \in L_2([0,1]^k) : \psi(u) \text{ depends on } (u_i)_{i \in A} \text{ only}\}$$

(in particular, $M_\emptyset$ consists of all constants) and

$$M_A^0 = \{\psi \in M_A : \mathbb{E}\{\psi(U)\varphi(U)\} = 0 \text{ for all } B \subsetneq A \text{ and all } \varphi \in M_B\}.$$

From Janson (1997, Lemma 11.17), it follows that, for any $\psi \in L_2([0,1]^k)$, there exists a unique orthogonal decomposition

$$\psi(u) = \sum_{A \subseteq [k]} \psi_A(u), \qquad \psi_A \in M_A^0,\ A \subseteq [k]. \qquad (6.3)$$

Applying this to $\psi_H = \prod_{i \overset{H^c}{\sim} j} \kappa(u_i, u_j) = \sum_{A \subseteq [k]} \tilde{\psi}_{H,A}(u)$, we can decompose s_H further into a sum of the form

$$s_{H,A}(u,y) = \frac{1}{(n)_k} \sum_{a \in \mathcal{A}_k^n} \tilde{\psi}_{H,A}(u_a) \prod_{i \overset{H}{\sim} j} \left(y_{a_i a_j} - \kappa(u_{a_i}, u_{a_j}) \right).$$

Let l be the number of vertices in $H \cup A$; we can rewrite $s_{H,A}$ as $r_{H,A}$ where

$$r_{H,A}(u,y) = \frac{1}{(n)_l} \sum_{a \in \mathcal{A}_l^n} \psi_{H,A}\left(u_{a_{A_P}}\right) \prod_{i \overset{H_P}{\sim} j} \left(y_{a_i a_j} - \kappa(u_{a_i}, u_{a_j}) \right),$$

with $\psi_{H,A}(u)$, $u \in [0,1]^{|A|}$, being the function obtained from $\tilde{\psi}_{H,A}(u)$, $u \in [0,1]^k$, by a change of coordinates from the (now ordered) set A to $(1, \dots, |A|)$. The claims about covariances and variances are straightforward to check. □

Proof of Lemma 2.2. Note that, for $|A| \geqslant 2$, M_A^0 is the L_2-closure of the linear space spanned by

$$\left\{ \prod_{i \in A} \psi_i(u_i) : \psi_i \in L_2^\circ([0,1]) \right\}.$$

Hence, for any $\varepsilon > 0$, there are $N_{H,A}$ and $\psi_{H,A,p,v} \in L_2^\circ([0,1])$, $v \in [|A|]$, $1 \leqslant p \leqslant N$, such that

$$\mathbb{E}\left(\psi_{H,A}(U) - \sum_{p=1}^{N_{H,A}} \prod_{i=1}^{|A|} \psi_{H,A,p,i}(U_i) \right)^2 \leqslant \varepsilon,$$

and hence, for any $a \in \mathcal{A}_k^n$,

$$\mathbb{E}\left(\psi_{H,A}(U_{a_{A_P}}) \prod_{i \overset{H_P}{\sim} j} (Y_{ij} - \kappa(U_i, U_j)) - \sum_{p=1}^{N_{H,A}} \prod_{i=1}^{|A|} \psi_{H,A,p,i}(U_{a_i}) \prod_{i \overset{H_P}{\sim} j} (Y_{ij} - \kappa(U_i, U_j)) \right)^2 \leqslant \varepsilon,$$

since $\left| \left(Y_{a_i a_j} - \kappa(U_{a_i}, U_{a_j}) \right) \right| \leqslant 1$. With

$$\tilde{r}_{H,A}(u,y) = \frac{1}{(n)_l} \sum_{a \in \mathcal{A}_l^n} \sum_{p=1}^{N_{H,A}} \prod_{i=1}^{|A|} \psi_{H,A,p,i}(u_{a_i}) \prod_{i \overset{H_P}{\sim} j} \left(y_{a_i a_j} - \kappa(u_{a_i}, u_{a_j}) \right)$$

we obtain

$$\begin{aligned}
&\mathbb{E}\left(r_{H,A}(U,Y) - \tilde{r}_{H,A}(U,Y) \right)^2 \\
&\leqslant \mathbb{E}\left(\frac{1}{(n)_l} \sum_{a \in \mathcal{A}_l^n} \left(\psi_{H,A}(U_a) - \sum_{p=1}^{N_{H,A}} \prod_{i=1}^{|A|} \psi_{H,A,p,i}(U_{a_i}) \right) \prod_{i \overset{H_P}{\sim} j} \left(Y_{a_i a_j} - \kappa(U_{a_i}, U_{a_j}) \right) \right)^2 \\
&\leqslant \frac{l!\,\varepsilon}{(n)_l},
\end{aligned}$$

where we have used that $\prod_{i=1}^{|A|} \psi_{H,A,p,i} \in L_2^\circ\left([0,1]^{|A|}\right)$, so that all cross terms with $|a \cap a'| > l$ vanish. The final claim now follows from Lemma 6.2. □

Lemma 6.2. *Let H be a graph on the vertex set $[l]$, and let $C_1, \dots, C_r$ denote the connected components of H. For each $1 \leqslant i \leqslant r$, let l_i be the size of C_i, let C_i' be a graph on $[l_i]$ that is isomorphic to C_i, and let $\psi_i \in L_2^\circ([0,1]^{l_i})$. Let*

$$S_i(u,y) = \sum_{a \in \mathcal{A}_{l_i}^n} \psi_i(u_a) \prod_{v \overset{C_i'}{\sim} w} (y_{a_v a_w} - \kappa(u_{a_v}, u_{a_w}))$$

Then

$$\mathbb{E}\left(\sum_{a \in \mathcal{A}_l^n} \prod_{i=1}^r \psi_i(U_{a_{V(C_i)}}) \prod_{v \overset{C_i}{\sim} w} (y_{a_v a_w} - \kappa(U_{a_v}, U_{a_w})) - \prod_{i=1}^r S_i(U,Y) \right)^2 \leqslant C n^{l-1} \tag{6.4}$$

for all $u_v \in [0,1]$ and $y_{vw} \in \{0,1\}$, $1 \leqslant v < w < n$.

Proof. When expanding the term $\prod_{i=1}^r \sum_{a \in \mathcal{A}_{l_i}^n}$, consider two cases: either the different tuples of indices are all disjoint, or they overlap by at least one index. The first case easily yields the second expression in the difference (6.4). For the second case, the size of the union of the indices can be at most $l-1$ which gives the order of the error in the approximation (6.4). □

A Complete orthogonal decomposition of the triangle density for 2×2 block graphon

Under the assumptions of Section 1.1, we have

$$t_{\vartriangle}^{\mathrm{inj}}(G_n) = R_{0.0} + R_{0.5} + R_{1.0} + R_{1.5} + R_{2.0} + R_{2.5}$$

where

$$\begin{aligned}
R_{0.0} &= (\alpha\gamma + \beta(1-\gamma)) \left(\gamma \left(\alpha^2\gamma + 3(1-\gamma)\delta^2 \right) - \alpha\beta\gamma(1-\gamma) + \beta^2(1-\gamma)^2 \right), \\
R_{0.5} &= \frac{1}{n^{1/2}} \times \left(3(\alpha\gamma(\alpha^2\gamma + (2-3\gamma)\delta^2) - \beta^3(1-\gamma)^2 + \beta(3\gamma^2 - 4\gamma + 1)\delta^2) \right) W, \\
R_{1.0} &= \frac{1}{n-1} \times 3 \left(\alpha^3\gamma + \delta^2(\alpha(1-3\gamma) - \beta(2-3\gamma)) + \beta^3(1-\gamma) \right) (W^2 - \gamma(1-\gamma)) \\
&\quad + \frac{1}{n^{1/2}(n-1)^{1/2}} \\
&\quad \times 18^{1/2} \Big((\alpha\gamma(\alpha - 2\delta) + \beta^2(1-\gamma) - 2\beta(1-\gamma)\delta + \delta^2) V_{\diagup,4} \\
&\quad\quad + (\alpha\gamma + \beta(1-\gamma))(\alpha\gamma - \beta(1-\gamma) + (1-2\gamma)\delta)(V_{\diagup,2} + V_{\diagup,3}) \\
&\quad\quad + \left(\gamma(\alpha^2\gamma^2 - 2\alpha(\gamma-1)\gamma\delta - (\gamma-1)\delta^2) - \beta^2(\gamma-1)^3 + 2\beta\gamma(\gamma-1)^2\delta \right) V_{\diagup,1} \Big), \\
R_{1.5} &= \frac{1}{n^{1/2}(n-1)} \times 3 \left(\alpha^3\gamma(3\gamma - 2) - \alpha \left(9\gamma^2 - 8\gamma + 1 \right) \delta^2 \right. \\
&\quad \left. -\beta^3 \left(3\gamma^2 - 4\gamma + 1 \right) + \beta \left(9\gamma^2 - 10\gamma + 2 \right) \delta^2 \right) W \\
&\quad + \frac{n^{1/2}}{(n-1)(n-2)} \times \left(\alpha^3 + 3\delta^2(\beta - \alpha) - \beta^3 \right) \left(W^3 - n^{-1/2}\gamma(1-\gamma)(1-2\gamma) \right) \\
&\quad + \frac{1}{n^{1/2}(n-1)^{1/2}(n-2)^{1/2}} \times 6^{1/2} \left(V_{\vartriangle} + V_{\diagup,1} + V_{\diagup,2} + V_{\diagup,3} \right) \\
&\quad + \frac{1}{(n-1)^{1/2}(n-2)} \\
&\quad \times 18^{1/2} \Big((\alpha^2\gamma^2 + 2\alpha\gamma(1-\gamma)\delta - \beta^2(1-\gamma)^2 - 2\beta\gamma(1-\gamma)\delta + (1-2\gamma)\delta^2) V_{\diagup,1} W
\end{aligned}$$

$$
\begin{aligned}
&+ (\alpha^2\gamma + \alpha(1-2\gamma)\delta + \beta^2(1-\gamma) - \beta(1-2\gamma)\delta - \delta^2)(V_{\swarrow,2} + V_{\swarrow,3})W \\
&+ (\alpha-\beta)(\alpha+\beta-2\delta)V_{\swarrow,4}W\Big),
\end{aligned}
$$

$$
\begin{aligned}
R_{2.0} = &\frac{1}{n^{1/2}(n-1)^{1/2}(n-2)} \\
&\times 18^{1/2}\big(2(1-\gamma)\gamma\left(\alpha^2(-\gamma) + \alpha(2\gamma-1)\delta + \beta^2(\gamma-1) + \beta(\delta-2\gamma\delta) + \delta^2\right)V_{\swarrow,1} \\
&+ (\beta-\alpha)\left(2\alpha\gamma(1-\gamma) + 2\beta\gamma(1-\gamma) + (1-2\gamma)^2\delta\right)(V_{\swarrow,2} + V_{\swarrow,3}) \\
&+ 2\left(\alpha^2(\gamma-1) + \alpha(\delta-2\gamma\delta) - \beta^2\gamma + \beta(2\gamma-1)\delta + \delta^2\right)V_{\swarrow,4}\big) \\
&+ \frac{1}{(n-1)(n-2)} \times 3(2\gamma-1)(\alpha-\beta)\left(\alpha^2 + \alpha\beta + \beta^2 - 3\delta^2\right)(W^2 - \gamma(1-\gamma)),
\end{aligned}
$$

$$
R_{2.5} = \frac{1}{n^{1/2}(n-1)(n-2)} \times 2\left(6\gamma^2 - 6\gamma + 1\right)(\alpha-\beta)\left(\alpha^2 + \alpha\beta + \beta^2 - 3\delta^2\right)W.
$$

Supplementary Material

Code to reproduce Table 1 (DOI: 10.1214/21-EJP708SUPP; .zip). The zip file contains code to reproduce Table 1. The code can also be found at https://github.com/aroellin/csgc.

References

S. Athreya, F. den Hollander, and A. Röllin (2019). Graphon-valued stochastic processes from population genetics. *arXiv preprint* arXiv:1908.06241. MR4312844

A. D. Barbour, M. Karoński, and A. Ruciński (1989). A central limit theorem for decomposable random variables with applications to random graphs. *J. Combin. Theory Ser. B* **47**, 125–145. MR1047781

V. Bentkus (2003). On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Inference* **113**, 385–402. MR1965117

C. Biernacki, G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 719–725.

B. Bollobás and O. Riordan (2009). Metrics for sparse graphs. In *Surveys in Combinatorics 2009*, Cambridge University Press, vol. 365 of *London Math. Soc. Lecture Note Ser.*, 211–287. MR2588543

C. Borgs, J. T. Chayes, H. Cohn, and N. Holden (2017). Sparse exchangeable graphs and their limits via graphon processes. *J. Mach. Learn. Res.* **18**, 7740–7810. MR3827098

C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao (2014a). An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *arXiv preprint* arXiv:1401.2906. MR3758733

C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao (2014b). An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *arXiv preprint* arXiv:1408.0744. MR3758733

S. Bubeck, J. Ding, R. Eldan, and M. Z. Rácz (2016). Testing for high-dimensional geometry in random graphs. *Random Struct. Algorithms* **49**, 503–532. MR3545825

F. Caron and E. B. Fox (2017). Sparse graphs using exchangeable random measures. *J. Roy. Statist. Soc. Ser. B* **79**, 1295–1366. MR3731666

Chatterjee, A., and Bhattacharya, B. B. (2021). Fluctuations of Subgraph Counts in Random Graphons. *arXiv preprint* arXiv:2104.07259.

L. H. Y. Chen and A. Röllin (2010). Stein couplings for normal approximation. *arXiv preprint* arXiv:1003.6039.

G. Di Nunno, B. K. Øksendal, and F. Proske (2009). *Malliavin calculus for Lévy processes with applications to finance*. vol. 2, Springer. MR2460554

P. Diaconis and S. Janson (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)* **28**, 33–61. MR2463439

X. Fang and A. Röllin (2015). Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics. *Bernoulli* **21**, 2157–2189. MR3378463

T. Funke and T. Becker (2019). Stochastic block models: A comparison of variants and inference methods. *PLoS one* **14**.

H. L. Gan, A. Röllin, and N. Ross (2017). Dirichlet approximation of equilibrium distributions in Cannings models with mutation. *Adv. Appl. Probab.* **49**, 927–959. MR3694323

C. Gao and J. Lafferty (2017a). Testing for global network structure using small subgraph statistics. *arXiv preprint* arXiv:1710.00862.

C. Gao and J. Lafferty (2017b). Testing network structure using relations between small subgraph probabilities. *arXiv preprint* arXiv:1704.06742.

J. Hladký, C. Pelekis, and M. Šileikis (2019). A limit theorem for small cliques in inhomogeneous random graphs. *arXiv preprint* arXiv:1903.10570. MR4313198

C. Hoppen, Y. Kohayakawa, C. G. Moreira, and R. M. Sampaio (2011). Limits of permutation sequences through permutation regularity. *arXiv preprint* arXiv:1106.1663. MR2995721

S. Janson (1994). Coupling and Poisson approximation. *Acta Appl. Math.* **34**, 7–15. MR1273843

S. Janson (1997). *Gaussian Hilbert Spaces*. Cambridge University Press. MR1474726

S. Janson and K. Nowicki (1991). The asymptotic distributions of generalized U-statistics with applications to random graphs. *Probab. Theory Related Fields* **90**, 341–375. MR1133371

K. Krokowski, A. Reichenbachs, and C. Thäle (2017). Discrete Malliavin-Stein method: Berry-Esseen bounds for random graphs and percolation. *Ann. Probab.* **45**. MR3630293

L. Lovász (2012). *Large Networks and Graph Limits*. American Mathematical Society. MR3012035

L. Lovász and B. Szegedy (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96**, 933–957. MR2274085

L. Lovász and B. Szegedy (2011). Finitely forcible graphons. *Journal of Combinatorial Theory, Series B* **101**, 269 – 301. MR2802882

P. A. Maugis (2020). Central limit theorems for local network statistics. *arXiv preprint* arXiv:2006.15738.

E. Meckes (2009). On Stein's method for multivariate normal approximation. In *High dimensional probability V: the Luminy volume*, Beachwood, OH: Inst. Math. Statist., vol. 5 of *Inst. Math. Stat. Collect.*, 153–178. MR2797946

I. Nourdin and G. Peccati (2012). *Normal approximation with Malliavin calculus: from Stein's method to universality*. vol. 192 of *Cambridge Tracts in Mathematics*, Cambridge: Cambridge University Press. MR2962301

D. Nualart (2006). *The Malliavin calculus and related topics*. vol. 1995, Springer. MR1344217

D. Nualart and G. Peccati (2005). Central limit theorems for sequences of multiple stochastic integrals. *Ann. Probab.* MR2118863

L. Ospina-Forero, C. M. Deane, and G. Reinert (2019). Assessment of model fit via network comparison methods based on subgraph counts. *J. Complex Netw.* **7**, 226–253.

N. Privault and G. Serafin (2020). Normal approximation for sums of weighted U-statistics — application to Kolmogorov bounds in random subgraph counting. *Bernoulli* **26**, 587–615. MR4036045

B. Ráth (2012). Time evolution of dense multigraph limits under edge-conservative preferential attachment dynamics. *Random Struct. Algorithms* **41**, 365–390. MR2967178

B. Ráth and L. Szakács (2012). Multigraph limit of the dense configuration model and the preferential attachment graph. *Acta Math. Hungar.* **136**, 196–221. MR2945218

G. Reinert and A. Röllin (2010). Random subgraph counts and U-statistics: Multivariate normal approximation via exchangeable pairs and embedding. *J. Appl. Probab.* **47**, 378–393. MR2668495

A. Röllin (2017). Kolmogorov bounds for the normal approximation of the number of triangles in the Erdős-Rényi random graph. *arXiv preprint* arXiv:1704.00410.

A. Röllin and N. Ross (2015). Local limit theorems via Landau-Kolmogorov inequalities. *Bernoulli* **21**, 851–880. MR3338649

Acknowledgments. We thank Siva Athreya and Matas Sileikis for helpful discussions. We also thank the referee for helpful suggestions.