Electronic Colloquium on Computational Complexity, Report No. 33 (2009)

Random Graphs and the Parity Quantifier

Phokion G. Kolaitis* Swastik Kopparty†

April 16, 2009

Abstract

The classical zero-one law for first-order logic on random graphs says that for every first-order property φ in the theory of graphs and every $p \in (0,1)$, the probability that the random graph $G(n,p)$ satisfies φ approaches either 0 or 1 as n approaches infinity. It is well known that this law fails to hold for any formalism that can express the parity quantifier: for certain properties, the probability that $G(n,p)$ satisfies the property need not converge, and for others the limit may be strictly between 0 and 1.

In this work, we capture the limiting behavior of properties definable in first order logic augmented with the parity quantifier, $\mathsf{FO}[\oplus]$, over $G(n,p)$, thus eluding the above hurdles. Specifically, we establish the following "modular convergence law":

> For every $\mathsf{FO}[\oplus]$ sentence φ, there are two explicitly computable rational numbers a_0, a_1, such that for $i \in \{0,1\}$, as n approaches infinity, the probability that the random graph $G(2n+i,p)$ satisfies φ approaches a_i.

Our results also extend appropriately to FO equipped with Mod_q quantifiers for prime q.

In the process of deriving the above theorem, we explore a new question that may be of interest in its own right. Specifically, we study the joint distribution of the subgraph statistics modulo 2 of $G(n,p)$: namely, the number of copies, mod 2, of a fixed number of graphs $F_1, \ldots, F_\ell$ of bounded size in $G(n,p)$. We first show that every $\mathsf{FO}[\oplus]$ property φ is almost surely determined by subgraph statistics modulo 2 of the above type. Next, we show that the limiting joint distribution of the subgraph statistics modulo 2 depends only on $n \mod 2$, and we determine this limiting distribution completely. Interestingly, both these steps are based on a common technique using multivariate polynomials over finite fields and, in particular, on a new generalization of the Gowers norm.

The first step above is analogous to the Razborov-Smolensky method for lower bounds for $\mathsf{AC0}$ with parity gates, yet stronger in certain ways. For instance, it allows us to obtain examples of simple graph properties that are exponentially uncorrelated with every $\mathsf{FO}[\oplus]$ sentence, which is something that is not known for $\mathsf{AC0}[\oplus]$.

*UC Santa Cruz and IBM Almaden Research Center. kolaitis@cs.ucsc.edu

†Massachusetts Institute of Technology. swastik@mit.edu

ISSN 1433-8092

Contents

1 Introduction

For quite a long time, combinatorialists have studied the asymptotic probabilities of properties on classes of finite structures, such as graphs and partial orders. Assume that $\mathcal{C}$ is a class of finite structures and let $\Pr_n$, $n \geq 1$, be a sequence of probability measures on all structures in $\mathcal{C}$ with n elements in their domain. If Q is a property of some structures in $\mathcal{C}$ (that is, a decision problem on $\mathcal{C}$), then the *asymptotic probability* $\Pr(Q)$ *of* Q *on* $\mathcal{C}$ is defined as $\Pr(Q) = \lim_{n\to\infty} \Pr_n(Q)$, provided this limit exists. In this paper, we will be focusing on the case when $\mathcal{C}$ is the class $\mathcal{G}$ of all finite graphs, and $\Pr_n = G(n,p)$ for constant p; this is the probability distribution on n-vertex undirected graphs where between each pair of nodes an edge appears with probability p, independently of other pairs of nodes. For example, for this case, the asymptotic probabilities $\Pr(\textsc{Connectivity}) = 1$ and $\Pr(\textsc{Hamiltonicity}) = 1$; in contrast, if $\Pr_n = G(n,p(n))$ with $p(n) = 1/n$, then $\Pr(\textsc{Connectivity}) = 0$ and $\Pr(\textsc{Hamiltonicity}) = 0$.

Instead of studying separately one property at a time, it is natural to consider formalisms for specifying properties of finite structures and to investigate the connection between the expressibility of a property in a certain formalism and its asymptotic probability. The first and most celebrated such connection was established by Glebskii et al. [GKLT69] and, independently, by Fagin [Fag76], who showed that a *0-1 law* holds for first-order logic[1] FO on the random graph $G(n,p)$ with p a constant in $(0,1)$; this means that if Q is a property of graphs expressible in FO and $\Pr_n = G(n,p)$ with p a constant in $(0,1)$, then $\Pr(Q)$ exists and is either 0 or 1. This result became the catalyst for a series of investigations in several different directions. Specifically, one line of investigation [SS87, SS88] investigated the existence of 0-1 laws for first-order logic FO on the random graph $G(n,p(n))$ with $p(n) = n^{-\alpha}$, $0 < \alpha < 1$. Since first-order logic on finite graphs has limited expressive power (for example, FO cannot express $\textsc{Connectivity}$ and 2-$\textsc{Colorability}$), a different line of investigation pursued 0-1 laws for extensions of first-order logic on the random graph $G(n,p)$ with p a constant in $(0,1)$. In this vein, it was shown in [BGK85, KV87] that the 0-1 law holds for extensions of FO with fixed-point operators, such as least fixed-point logic LFP, which can express $\textsc{Connectivity}$ and 2-$\textsc{Colorability}$. As regards to higher-order logics, it is clear that the 0-1 law fails even for existential second-order logic ESO, since it is well known that $\mathsf{ESO} = \mathsf{NP}$ on finite graphs [Fag74]. In fact, even the *convergence law* fails for ESO, that is, there are ESO-expressible properties Q of finite graphs such that $\Pr(Q)$ does *not* exist. For this reason, a separate line of investigation pursued 0-1 laws for syntactically-defined subclasses of NP. Eventually, this investigation produced a complete classification of the quantifier prefixes of ESO for which the 0-1 law holds [KV87, KV90, PS89], and provided a unifying account for the asymptotic probabilities of such NP-complete problems as k-$\textsc{Colorability}$, $k \geq 3$.

Let L be a logic for which the 0-1 law (or even just the convergence law) holds on the random graph $G(n,p)$ with p a constant in $(0,1)$. An immediate consequence of this is that L cannot

[1] Recall that the formulas of first-order logic on graphs are obtained from atomic formulas $E(x,y)$ (interpreted as the adjacency relation) and equality formulas $x = y$ using Boolean combinations, existential quantification, and universal quantification; the quantifiers are interpreted as ranging over the set of vertices of the graph (and not over sets of vertices or sets of edges, etc.).

express any *counting* properties, such as EVEN CARDINALITY ("there is an even number of nodes"), since $\Pr_{2n}(\text{EVEN CARDINALITY}) = 1$ and $\Pr_{2n+1}(\text{EVEN CARDINALITY}) = 0$. In this paper, we turn the tables around and systematically investigate the asymptotic probabilities of properties expressible in extensions of FO with *counting quantifiers* Mod_q^i, where q is a prime number. The most prominent such extension is $\mathsf{FO}[\oplus]$, which is the extension of FO with the *parity quantifier* Mod_2^1. The syntax of $\mathsf{FO}[\oplus]$ augments the syntax of FO with the following formation rule: if $\varphi(y)$ is a $\mathsf{FO}[\oplus]$-formula, then $\oplus y\varphi(y)$ is also a $\mathsf{FO}[\oplus]$-formula; this formula is true if the number of y's that satisfy $\varphi(y)$ is odd (analogously, $\mathsf{Mod}_q^i y\varphi(y)$ is true if the number of y's that satisfy $\varphi(y)$ is congruent to $i \mod q$). A typical property on graphs expressible in $\mathsf{FO}[\oplus]$ (but not in FO) is $\mathcal{P} := \{G : \text{ every vertex of } G \text{ has odd degree}\}$, since a graph is in $\mathcal{P}$ if and only if it satisfies the $\mathsf{FO}[\oplus]$-sentence $\forall x \oplus y E(x,y)$.

Our main result (see Theorem 2.1) is a *modular convergence law* for $\mathsf{FO}[\oplus]$ on $G(n,p)$ with p a constant in $(0,1)$. This law asserts that if φ is a $\mathsf{FO}[\oplus]$-sentence, then there are two explicitly computable rational numbers a_0, a_1, such that, as $n \to \infty$, the probability that the random graph $G(2n+i,p)$ satisfies φ approaches a_i, for $i = 0, 1$. Moreover, a_0 and a_1 are computable and are of the form $r/2^s$, where r and s are non-negative integers. We also establish that an analogous modular convergence law holds for every extension $\mathsf{FO}[\mathsf{Mod}_q]$ of FO with the counting quantifiers $\{\mathsf{Mod}_q^i : i \in [q-1]\}$, where q is a prime. It should be noted that results in [HKL96] imply that the modular convergence law for $\mathsf{FO}[\oplus]$ does *not* generalize to extensions of $\mathsf{FO}[\oplus]$ with fixed-point operators. This is in sharp contrast to the aforementioned 0-1 law for FO which carries over to extensions of FO with fixed-point operators.

1.1 Methods

Earlier 0-1 laws have been established by a combination of standard methods and techniques from mathematical logic and random graph theory. In particular, on the side of mathematical logic, the tools used include the compactness theorem, Ehrenfeucht-Fraïssé games, and quantifier elimination. Here, we establish the modular convergence law by combining quantifier elimination with, interestingly, algebraic methods related to multivariate polynomials over finite fields. In what follows in this section, we present an overview of the methods and techniques that we will use.

1.1.1 The distribution of subgraph frequencies mod q, polynomials and Gowers norms

Let us briefly indicate the relevance of polynomials to the study of $\mathsf{FO}[\oplus]$ on random graphs. A natural example of a statement in $\mathsf{FO}[\oplus]$ is a formula φ such that G satisfies φ if and only if the number of copies of H in G is odd, for some graph H (where by copy we mean an induced subgraph, for now). Thus understanding the asymptotic probability of φ on $G(n,p)$ amounts to understanding the distribution of the number of copies (mod 2) of H in $G(n,p)$.

In this spirit, we ask: what is the probability that in $G(n,1/2)$ there is an odd number of triangles (where we count *unordered* triplets of vertices $\{a,b,c\}$ such that a,b,c are all pairwise adjacent[2])?

We reformulate this question in terms of the following "triangle polynomial", that takes the adjacency matrix of a graph as input and returns the parity of the number of triangles in the graph; $P_\triangle : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$, where

$$P_\triangle((x_e)_{e\in\binom{n}{2}}) = \sum_{\{e_1,e_2,e_3\} \text{ forming a } \triangle} x_{e_1}x_{e_2}x_{e_3},$$

where the arithmetic is mod 2. Note that for the random graph $G(n,1/2)$, each entry of the adjacency matrix is chosen independently and uniformly from $\{0,1\}$. Thus the probability that a random graph $G \in G(n,1/2)$ has an odd number of triangles is precisely equal to $\Pr_{x\in\mathbb{Z}_2^n}[P_\triangle(x) = 1]$. Thus we have reduced our problem to studying the distribution of the evaluation of a certain polynomial at a random point, a topic of much study in pseudorandomness and algebraic coding theory, and we may now appeal to tools from these areas.

In Section 3, via the above approach, we show that the probability that $G(n,1/2)$ has an odd number of triangles equals $1/2 \pm 2^{-\Omega(n)}$. Similarly, for any connected graph $F \neq \mathrm{K}_1$ (the graph consisting of one vertex), the probability that $G(n,1/2)$ has an odd number of copies[3] of F is also $1/2 \pm 2^{-\Omega(n)}$ (when $F = \mathrm{K}_1$, there is no randomness in the number of copies of F in $G(n,1/2)$!). In fact, we show that for any collection of distinct connected graphs $F_1,\ldots,F_\ell$ ($\neq \mathrm{K}_1$), the joint distribution of the number of copies mod 2 of $F_1,\ldots,F_\ell$ in $G(n,1/2)$ is $2^{-\Omega(n)}$-close to the uniform distribution on $\mathbb{Z}_2^\ell$, i.e., the events that there are an odd number of F_i are essentially independent of one another.

Generalizing the above to $G(n,p)$ and counting mod q for arbitrary $p \in (0,1)$ and arbitrary integers q motivates the study of new kinds of questions about polynomials, that we believe are interesting in their own right. For $G(n,p)$ with arbitrary p, we need to study the distribution of $P(x)$, for certain polynomials P, when $x \in \mathbb{Z}_2^m$ is distributed according to the *p-biased measure*. Even more interestingly, for the study of $\mathsf{FO}[\mathsf{Mod}_q]$, where we are interested in the distribution of the number of triangles mod q, one needs to understand the distribution of $P(x)$ (P is now a polynomial over $\mathbb{Z}_q$) where x is chosen uniformly from $\{0,1\}^m \subseteq \mathbb{Z}_q^m$ (as opposed to x being chosen uniformly from all of $\mathbb{Z}_q^m$, which is traditionally studied). In Section 4, we develop all the relevant polynomial machinery in order to answer these questions. This involves generalizing some classical results of Babai, Nisan and Szegedy [BNS89] on correlations of polynomials. The key technical innovation here is our definition of a μ-Gowers norm (where μ is a measure on $\mathbb{Z}_q^m$) that measures the correlation, under μ, of a given function with low-degree polynomials (letting μ be the uniform measure, we recover the standard Gowers norm). After generalizing several results about the standard Gowers norm to the μ-Gowers norm case, we can then use a technique of Viola and Wigderson [VW07] to establish the generalization of [BNS89] that we need.

[2]Counting the number of *unordered* triples is not expressible in $\mathsf{FO}[\oplus]$, we ask this question only for expository purposes (nevertheless, we do give an answer to this question in Section 3).

[3]with a certain precise definition of "copy".

1.1.2 Quantifier elimination

Although we studied the distribution of subgraph frequencies mod q as an attempt to determine the limiting behavior of only a special family of $\mathsf{FO}[\mathsf{Mod}_q]$ properties, it turns out that this case, along with the techniques developed to handle it, play a central role in the proof of the full modular convergence law. In fact, we reduce the modular convergence law for general $\mathsf{FO}[\mathsf{Mod}_q]$ properties to the above case. We show that for any $\mathsf{FO}[\mathsf{Mod}_q]$ sentence φ, with high probability over $G \in G(n,p)$, the truth of φ on G is determined by the number of copies in G, mod q, of each small subgraph. Then by the results described earlier on the equidistribution of these numbers (except for the number of K_1, which depends only on n mod q), the full modular convergence law for $\mathsf{FO}[\mathsf{Mod}_q]$ follows.

In Section 6, we establish such a reduction using the method of elimination of quantifiers. To execute this, we need to analyze $\mathsf{FO}[\mathsf{Mod}_q]$ formulas which may contain free variables (i.e., not every variable used is quantified). Specifically, we show that for every $\mathsf{FO}[\mathsf{Mod}_q]$ formula $\varphi(\alpha_1, \ldots, \alpha_k)$, with high probability over $G \in G(n,p)$, it holds that for all vertices $w_1, \ldots, w_k$ of G, the truth of $\varphi(w_1, \ldots, w_k)$ is entirely determined by the following data: (a) which of the w_i, w_j pairs are adjacent, (b) which of the w_i, w_j pairs are equal to one another, and (c) the number of copies "rooted" at $w_1, \ldots, w_k$, mod q, of each small *labelled graph*. This statement is a generalization of what we needed to prove, but lends itself to inductive proof (*this* is quantifier elimination). This leads us to studying the distribution (via the polynomial approach described earlier) of the number of copies of labelled graphs in G; questions of the form, given two specified vertices v, w (the "roots"), what is the probability that there are an odd number of paths of length 4 in $G \in G(n,p)$ from v to w? After developing the necessary results on the distribution of labelled subgraph frequencies, combined with some elementary combinatorics, we can eliminate quantifiers and thus complete the proof of the modular convergence law.

1.2 Comparison with $\mathsf{AC0}[\oplus]$

Every $\mathsf{FO}[\oplus]$ property naturally defines a family of boolean functions $f_n : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$, such that a graph G satisfies φ if and only if $f_n(A_G) = 1$, where A_G is the adjacency matrix of G. This family of functions is easily seen to be contained in $\mathsf{AC0}[\oplus]$, which is $\mathsf{AC0}$ with parity gates (each $\forall$ becomes an AND gate, $\exists$ becomes a OR gate and $\oplus$ becomes a parity gate). This may be summarized by saying that $\mathsf{FO}[\oplus]$ is a highly uniform version of $\mathsf{AC0}[\oplus]$.

Currently, all our understanding of the power of $\mathsf{AC0}[\oplus]$ comes from the Razborov-Smolensky [Raz87, Smo87] approach to proving circuit lower bounds on $\mathsf{AC0}[\oplus]$. At the heart of this approach is the result that for every $\mathsf{AC0}[\oplus]$ function f, there is a low-degree polynomial P such that for $1 - \epsilon(n)$ fraction of inputs, the evaluations of f and P are equal. Note that this result automatically holds for $\mathsf{FO}[\oplus]$ (since $\mathsf{FO}[\oplus] \subseteq \mathsf{AC0}[\oplus]$).

We show that for the special case when $f : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$ comes from an $\mathsf{FO}[\oplus]$ property φ, a significantly improved approximation may be obtained: *(i)* We show that the degree of P may be chosen to be a constant depending only on φ, whereas the Razborov-Smolensky approximation required P to be of polylog(n) degree, *(ii)* The error parameter $\epsilon(n)$ may be

chosen to be exponentially small in n, whereas the Razborov-Smolensky method only yields $\epsilon(n) = 2^{-\log^{O(1)} n}$. *(iii):* Finally, the polynomial P can be chosen to be symmetric under the action of S_n on the $\binom{n}{2}$ coordinates, while in general, the polynomial produced by the Razborov-Smolensky approach need not be symmetric (due to the randomness involved in the choices).

These strengthened approximation results allow us, using known results about pseudorandomness against low-degree polynomials, to show that *(i)* there exist explicit pseudorandom generators that fool $\mathsf{FO}[\oplus]$ sentences, and *(ii)* there exist explicit functions f such that for any $\mathsf{FO}[\oplus]$ formula φ, the probability over $G \in G(n,p)$ that $f(G) = \varphi(G)$ is at most $\frac{1}{2} + 2^{-\Omega(n)}$. The first result follows from the pseudorandom generators against low-degree polynomials due to Bogdanov-Viola [BV07], Lovett [Lov08] and Viola [Vio08]. The second result follows from the result of Babai, Nisan and Szegedy [BNS89], and our generalization of it, giving explicit functions that are uncorrelated with low degree polynomials.

Obtaining similar results for $\mathsf{AC0}[\oplus]$ is one of the primary goals of modern day "low-level" complexity theory.

Organization of this paper: In the next section, we formally state our main results and some of its corollaries. In Section 3, we determine the distribution of subgraph frequencies mod q. In Section 4, we introduce the μ-Gowers Norm use it to prove some technical results on the bias of polynomials needed for the previous section. In Section 5, we state the theorem which implements the quantifier elimination and describe the plan for its proof. This plan is then executed in Sections 6, 7 and 8. We conclude with some open questions.

2 The Modular Convergence Law

We now state our main theorem.

Theorem 2.1 *Let q be a prime. Then for every $\mathsf{FO}[\mathsf{Mod}_q]$-sentence φ, there exist rationals $a_0, \ldots, a_{q-1}$ such that for every $p \in (0,1)$ and every $i \in \{0,1,\ldots,q-1\}$,*

$$\lim_{\substack{n\to\infty \\ n \equiv i \bmod q}} \Pr_{G \in G(n,p)}[G \text{ satisfies } \varphi] = a_i.$$

Remark The proof of Theorem 2.1 also yields:

- Given the formula φ, the numbers $a_0, \ldots, a_{q-1}$ can be computed.
- Each a_i is of the form r/q^s, where r, s are nonnegative integers.
- For every sequence of numbers $b_0, \ldots, b_{q-1} \in [0,1]$, each of the form r/q^s, there is a $\mathsf{FO}[\mathsf{Mod}_q]$-sentence φ such that for each i, the number a_i given by the theorem equals b_i.

Before we describe the main steps in the proof of Theorem 2.1, we make a few definitions.

For graphs $F = (V_F, E_F)$ and $G = (V_G, E_G)$, an *(injective) homomorphism* from F to G is an (injective) map $\chi : V_F \to V_G$ that maps edges to edges, i.e., for any $(u, v) \in E_F$, we have $(\chi(u), \chi(v)) \in E_G$. Note that we do not require that χ maps non-edges to non-edges. We denote by $[F](G)$ the number of injective homomorphisms from F to G, and we denote by $[F]_q(G)$ this number $\mod q$. We let $\mathsf{aut}(F) := [F](F)$ be the number of automorphisms of F.

The following lemma (which follows from Lemma 6.5 in Section 6), shows that for some graphs F, as G varies, the number $[F](G)$ cannot be arbitrary.

Lemma 2.2 *Let F be a connected graph and G be any graph. Then $\mathsf{aut}(F) \mid [F](G)$.*

For the rest of this section, let q be a fixed prime. Let Conn^a be the set of connected graphs on at most a vertices. For any graph G, let the *subgraph frequency vector* $\mathsf{freq}^a_G \in \mathbb{Z}_q^{\mathsf{Conn}^a}$ be the vector such that its value in coordinate F ($F \in \mathsf{Conn}^a$) equals $[F]_q(G)$, the number of injective homomorphisms from F to $G \bmod q$. Let $\mathsf{FFreq}(a)$, the set of *feasible frequency vectors*, be the subset of $\mathbb{Z}_q^{\mathsf{Conn}^a}$ consisting of all f such that for all $F \in \mathsf{Conn}^a$, $f_F \in \mathsf{aut}(F) \cdot \mathbb{Z}_q := \{\mathsf{aut}(F) \cdot x \mid x \in \mathbb{Z}_q\}$. By Lemma 2.2, for every G and a, $\mathsf{freq}^a_G \in \mathsf{FFreq}(a)$, i.e., the subgraph frequency vector is always feasible.

We can now state the two main technical results that underlie Theorem 2.1.

The first states that on almost all graphs G, every $\mathsf{FO}[\mathsf{Mod}_q]$ formula can be expressed in terms of the subgraph frequencies, $[F]_q(G)$, over all small connected graphs F.

Theorem 2.3 (Subgraph frequencies $\bmod q$ determine $\mathsf{FO}[\mathsf{Mod}_q]$ formulae) *For every $\mathsf{FO}[\mathsf{Mod}_q]$-sentence φ of quantifier depth t, there exists an integer $c = c(t, q)$ and a function $\psi : \mathbb{Z}_q^{\mathsf{Conn}^c} \to \{0, 1\}$ such that for all $p \in (0, 1)$,*

$$\Pr_{G \in G(n,p)} [(G \text{ satisfies } \varphi) \Leftrightarrow (\psi(\mathsf{freq}^c_G) = 1)] \geq 1 - \exp(-n).$$

This result is complemented by the following result, that shows the distribution of subgraph frequencies $\bmod q$ in a random graph $G \in G(n, p)$ is essentially uniform in the space of all feasible frequency vectors, up to the obvious restriction that the number of vertices (namely the frequency of K_1 in G) should equal $n \mod q$.

Theorem 2.4 (Distribution of subgraph frequencies $\bmod q$ depends only on $n \bmod q$) *Let $p \in (0, 1)$. Let $G \in G(n, p)$. Then for any constant a, the distribution of freq^a_G is $\exp(-n)$-close to the uniform distribution over the set*

$$\{f \in \mathsf{FFreq}(a) : f_{K_1} \equiv n \mod q\}.$$

Theorem 2.4 is proved in Section 3 by studying the bias of multivariate polynomials over finite fields via a generalization of the Gowers norm. Theorem 2.3 is proved in Section 6 using two main ingredients:

1. A generalization of Theorem 2.4 that determines the joint distribution of the frequencies of "labelled subgraphs" with given roots (see Section 8).

2. A variant of quantifier elimination (that may be called quantifier conversion) designed to handle Mod_q quantifiers that crucially uses the probabilistic input from the previous ingredient (see Section 6).

Proof of Theorem 2.1: Follows by combining Theorem 2.3 and Theorem 2.4. □

2.1 Pseudorandomness against $\mathsf{FO}[\oplus]$

We now point out three simple corollaries of our study of $\mathsf{FO}[\oplus]$ on random graphs.

Corollary 2.5 ($\mathsf{FO}[\mathsf{Mod}_q]$ is well approximated by low-degree polynomials) *For every $\mathsf{FO}[\mathsf{Mod}_q]$-sentence φ, there is a constant d, such that for each $n \in \mathbb{N}$, there is a degree d polynomial $P((X_e)_{e\in\binom{n}{2}}) \in \mathbb{Z}_q[(X_e)_{e\in\binom{n}{2}}]$, such that for all $p \in (0,1)$,*

$$\Pr_{G\in G(n,p)}[(G \text{ satisfies } \varphi) \Leftrightarrow P(A_G) = 1] \geq 1 - 2^{-\Omega(n)},$$

where $A_G \in \{0,1\}^{\binom{n}{2}}$ is the adjacency matrix of G.

Proof Follows from Theorem 2.3 and the observation that for any graph F of constant size, there is a polynomial $Q((X_e)_{e\in\binom{n}{2}})$ of constant degree, such that $Q(A_G) = [F]_q(G)$ for all graphs G. □

Corollary 2.6 (PRGs against $\mathsf{FO}[\oplus]$) *For each $s \in \mathbb{N}$ and constant $\epsilon > 0$, there is a constant $c \geq 0$ such that for each n, there is a family $\mathcal{F}$ of $\Theta(n^c)$ graphs on n vertices, computable in time $\mathrm{poly}(n^c)$, such that for all $\mathsf{FO}[\oplus]$-sentences φ of size at most s, and for all $p \in (0,1)$,*

$$|\Pr_{G\in\mathcal{F}}[G \text{ satisfies } \varphi] - \Pr_{G\in G(n,p)}[G \text{ satisfies } \varphi]| < \epsilon.$$

Proof For $p = 1/2$, this follows from the previous corollary and the result of Viola [Vio08] (building on results of Bogdanov-Viola [BV07] and Lovett [Lov08]) constructing a pseudorandom generator fooling low-degree polynomials under the uniform distribution. For general p, note that the same family $\mathcal{F}$ from the $p = 1/2$ case works, since the distribution of subgraph frequencies given in Theorem 2.4 is independent of p. □

The analogue of the previous corollary for FO was proved in [GS71, BEH81] (see also [BR05, NNT05]).

Corollary 2.7 (Explicit functions exponentially hard for $\mathsf{FO}[\oplus]$) *There is an explicit function $f : \{0,1\}^{\binom{n}{2}} \to \{0,1\}$ such that for every $\mathsf{FO}[\oplus]$-sentence φ,*

$$\Pr_{G \in G(n,p)}[(G \text{ satisfies } \varphi) \Leftrightarrow (f(A_G) = 1)] < \frac{1}{2} + 2^{-\Omega(n)}.$$

Proof Follows from Corollary 2.5, and the result of Babai, Nisan, Szegedy [BNS89] (for $p = 1/2$) and its generalization, Lemma 4.1 (for general p), constructing functions exponentially uncorrelated with low degree polynomials under the p-biased measure. It actually follows from our proofs that, one may even choose a function f that is a graph property (namely, invariant under the action of S_n on the coordinates). □

3 The Distribution of Subgraph Frequencies mod q

In this section, we prove Theorem 2.4 on the distribution of subgraph frequencies in $G(n,p)$.

We first make a few definitions. If F is a connected graph and G is any graph, a *copy* of F in G is a set $E \subseteq E_G$ such that there exists an injective homomorphism χ from F to G such that $E = \chi(E_F) := \{(\chi(v), \chi(w)) \mid (v,w) \in E_F\}$. We denote the set of copies of F in G by $\mathsf{Cop}(F,G)$, the cardinality of $\mathsf{Cop}(F,G)$ by $\langle F \rangle(G)$, and this number mod q by $\langle F \rangle_q(G)$. We have the following basic relation (which follows from Lemma 6.5 in Section 6).

Lemma 3.1 *If F is a connected graph with $|E_F| \geq 1$, then*

$$[F](G) = \mathsf{aut}(F) \cdot \langle F \rangle(G).$$

For notational convenience, we view $G(n,p)$ as a graph whose vertex set is $[n]$ and whose edge set is a subset of $\binom{[n]}{2}$.

We can now state the general equidistribution theorem from which Theorem 2.4 will follow easily (We use the notation $\Omega_{q,p,d}(n)$ to denote the expression $\Omega(n)$, where the implied constant depends only on q, p and d). Note that this theorem holds for arbitrary integers q, not necessarily prime.

Theorem 3.2 (Equidistribution of graph copies) *Let $q > 1$ be an integer and let $p \in (0,1)$. Let $F_1, \ldots, F_\ell \in \mathsf{Conn}^a$ be distinct graphs with $1 \leq |E_{F_i}| \leq d$.*

Let $G \in G(n,p)$. Then the distribution of $(\langle F_1 \rangle_q(G), \ldots, \langle F_\ell \rangle_q(G))$ on $\mathbb{Z}_q^\ell$ is $2^{-\Omega_{q,p,d}(n)+\ell}$-close to uniform in statistical distance.

Using this theorem, we complete the proof of Theorem 2.4.

Proof of Theorem 2.4: Let $F_1, \ldots, F_\ell$ be an enumeration of the elements of Conn^a except for K_1. By Theorem 3.2, the distribution of $g = (\langle F_i \rangle_q(G))_{i=1}^\ell$ is $2^{-\Omega(n)}$ close to uniform over $\mathbb{Z}_q^\ell$. Given the vector g, we may compute the vector freq_G^a by:

- $(\mathsf{freq}_G^a)_{\mathrm{K}_1} = n \mod q$.
- For $F \in \mathsf{Conn}^a \setminus \{\mathrm{K}_1\}$, $(\mathsf{freq}_G^a)_F = g_F \cdot \mathsf{aut}(F)$ (by Lemma 3.1).

This implies that the distribution of freq_G^a is $2^{-\Omega(n)}$-close to uniformly distributed over $\{f \in \mathsf{FFreq}(a) : f_{\mathrm{K}_1} = n \mod q\}$. □

Towards proving Theorem 3.2, we now introduce some tools.

3.1 Preliminary lemmas

As indicated in the introduction, the distribution of subgraph frequencies is most naturally studied via the distribution of values of certain polynomials. The following lemma, which is used in the proof of Theorem 3.2 (and again in Section 8 to study the distribution of labelled subgraph frequencies), gives a simple sufficient criterion for the distribution of values of a polynomial to be "unbiased". The proof appears in Section 4.

Lemma 3.3 *Let $q > 1$ be an integer and let $p \in (0,1)$. Let[4] $\mathcal{F} \subseteq 2^{[m]}$. Let $d > 0$ be an integer. Let $Q(Z_1, \ldots, Z_m) \in \mathbb{Z}_q[Z_1, \ldots, Z_m]$ be a polynomial of the form*

$$\sum_{S \in \mathcal{F}} a_S \prod_{i \in S} Z_i + Q'(\mathbf{Z}),$$

where $\deg(Q') < d$. Suppose there exist $\mathcal{E} = \{E_1, \ldots, E_r\} \subseteq \mathcal{F}$ such that:

- *$|E_j| = d$ for each j,*
- *$a_{E_j} \neq 0$ for each j.*
- *$E_j \cap E_{j'} = \emptyset$ for each j, j',*
- *For each $S \in \mathcal{F} \setminus \mathcal{E}$, $|S \cap (\cup_j E_j)| < d$.*

Let $\mathbf{z} = (z_1, \ldots, z_m) \in \mathbb{Z}_q^m$ be the random variable where, independently for each i, we have $\Pr[z_i = 1] = p$ and $\Pr[z_i = 0] = 1 - p$. Then,

$$\left| \mathbb{E}\left[\omega^{Q(\mathbf{z})}\right] \right| \leq 2^{-\Omega_{q,p,d}(r)},$$

where $\omega \in \mathbb{C}$ is a primitive q^{th}-root of unity.

The lemma below is a useful tool for showing that a distribution on $\mathbb{Z}_q^\ell$ is close to uniform.

Lemma 3.4 (Vazirani XOR lemma) *Let $q > 1$ be an integer and let $\omega \in \mathbb{C}$ be a primitive q^{th}-root of unity. Let $\mathbf{X} = (X_1, \ldots, X_\ell)$ be a random variable over $\mathbb{Z}_q^\ell$. Suppose that for every nonzero $c \in \mathbb{Z}_q^\ell$,*

$$\left| \mathbb{E}\left[\omega^{\sum_{i \in [\ell]} c_i X_i}\right] \right| \leq \epsilon.$$

Then $\mathbf{X}$ is $q^\ell \cdot \epsilon$-close to uniformly distributed over $\mathbb{Z}_q^\ell$.

[4] If S is a set, we use the notation 2^S to denote its power set.

3.2 Proof of the equidistribution theorem

Proof of Theorem 3.2: By the Vazirani XOR Lemma (Lemma 3.4), it suffices to show that for each nonzero $c \in \mathbb{Z}_q^\ell$, we have $\left|\mathbb{E}\left[\omega^R\right]\right| \leq 2^{-\Omega_{q,p,d}(n)}$, where $R := \sum_{i\in[\ell]} c_i \langle F_i \rangle_q(G)$, and $\omega \in \mathbb{C}$ is a primitive q^{th}-root of unity.

We will show this by appealing to Lemma 3.3. Let $m = \binom{n}{2}$. Let $\mathbf{z} \in \{0,1\}^{\binom{[n]}{2}}$ be the random variable where, for each $e \in \binom{[n]}{2}$, $z_e = 1$ if and only if e is present in G. Thus, independently for each e, $\Pr[z_e = 1] = p$.

We may now express R in terms of the z_e. Let K_n denote the complete graph on the vertex set $[n]$. Thus $\mathsf{Cop}(F_i, \mathrm{K}_n)$ is the set of E that could potentially arise as copies of F_i in G. Then we may write,

$$\begin{aligned} R = \sum_{i\in[\ell]} c_i \langle F_i \rangle_q(G) &= \sum_{i\in[\ell]} c_i \sum_{E\in\mathsf{Cop}(F_i,\mathrm{K}_n)} \prod_{e\in E} z_e \\ &= \sum_{E\in\mathcal{F}} c_E \prod_{e\in E} z_e, \end{aligned}$$

where $\mathcal{F} \subseteq 2^{\binom{[n]}{2}}$ is the set $\bigcup_{i:c_i\neq 0} \mathsf{Cop}(F_i, \mathrm{K}_n)$, and for $E \in \mathcal{F}$, $c_E = c_i$ for the unique i satisfying $E \in \mathsf{Cop}(F_i, \mathrm{K}_n)$ (note that since the F_i are nonisomorphic connected graphs, the $\mathsf{Cop}(F_i, \mathrm{K}_n)$ are pairwise disjoint).

Let $Q(\mathbf{Z}) \in \mathbb{Z}_q[\mathbf{Z}]$, where $\mathbf{Z} = (Z_e)_{e\in\binom{[n]}{2}}$ be the polynomial $\sum_{E\in\mathcal{F}} c_E \prod_{e\in E} Z_e$. Then $R = Q(\mathbf{z})$. We wish to show that

$$\left|\mathbb{E}\left[\omega^{Q(\mathbf{z})}\right]\right| \leq 2^{-\Omega_{q,p,d}(n)}. \tag{1}$$

We do this by demonstrating that the polynomial $Q(\mathbf{Z})$ satisfies the hypotheses of Lemma 3.3.

Let $d^* = \max_{i:c_i\neq 0} |E_{F_i}|$. Let $i_0 \in [\ell]$ be such that $c_{i_0} \neq 0$ and $|E_{F_{i_0}}| = d^*$. Let $\chi_1, \chi_2, \ldots, \chi_r \in \mathsf{Inj}(F_{i_0}, \mathrm{K}_n)$ be a collection of homomorphisms such that for all distinct $j, j' \in [r]$, we have $\chi_j(V_{F_{i_0}}) \cap \chi_{j'}(V_{F_{i_0}}) = \emptyset$. Such a collection can be chosen greedily so that $r = \Omega(\frac{n}{d})$. Let $E_j \in \mathsf{Cop}(F_{i_0}, \mathrm{K}_n)$ be given by $\chi_j(E_{F_{i_0}})$. Let $\mathcal{E}$ be the family of sets $\{E_1, \ldots, E_r\} \subseteq \mathcal{F}$. We observe the following properties of the E_j:

1. For each $j \in [r]$, $|E_j| = d^*$ (since χ_j is injective).
2. For each $j \in [r]$, $c_{E_j} = c_{i_0} \neq 0$.
3. For distinct $j, j' \in [r]$, $E_j \cap E_{j'} = \emptyset$ (by choice of the χ_j).
4. For every $S \in \mathcal{F} \setminus \mathcal{E}$, $|S \cap (\cup_j E_j)| < d^*$. To see this, take any $S \in \mathcal{F} \setminus \mathcal{E}$ and suppose $|S \cap (\cup_j E_j)| \geq d^*$. Let $i' \in [\ell]$ be such that $c_{i'} \neq 0$ and $S \in \mathsf{Cop}(F_{i'}, \mathrm{K}_n)$. Let $\chi \in \mathsf{Inj}(F_{i'}, \mathrm{K}_n)$ with $\chi(E_{F_{i'}}) = S$. By choice of d^*, we know that $|S| \leq d^*$. Therefore, the only way that $|S \cap (\cup_j E_j)|$ can be $\geq d^*$ is if (1) $|S| = d^*$, and (2) $S \cap (\cup_j E_j) = S$, or in other words, $S \subseteq (\cup_j E_j)$. However, since the $\chi_j(V_{F_{i_0}})$ are all pairwise disjoint, this implies that $S \subseteq E_j$ for some j. But since $|E_j| = |S|$, we have $S = E_j$, contradicting our choice of S. Therefore, $|S \cap (\cup_j E_j)| < d^*$ for any $S \in \mathcal{F} \setminus \mathcal{E}$.

It now follows that $Q(\mathbf{Z})$, $\mathcal{F}$ and $\mathcal{E}$ satisfy the hypothesis of Lemma 3.3. Consequently, (recalling that $r = \Omega(n/d)$ and $d^* \leq d$) Equation (1) follows, completing the proof of the theorem. □

Remark We just determined the joint distribution of the number of injective homomorphisms, mod q, from all small connected graphs to $G(n,p)$. This information can be used in conjunction with Lemma 6.2 to determine the joint distribution of the number of injective homomorphisms, mod q, from *all* small graphs to $G(n,p)$.

4 The Bias of Polynomials

Our main goal in this section is to give a full proof of Lemma 3.3, which gives a criterion for a polynomial to be unbiased. Along the way, we will introduce the μ-Gowers norm and some of its useful properties.

Our proof of Lemma 3.3 will go through the following lemma (which is proved in the next subsection). It shows that "Generalized Inner Product" polynomials are uncorrelated with polynomials of lower degree. This generalizes a result of Babai Nisan and Szegedy [BNS89] (which dealt with the case $q = 2$ and $p = 1/2$).

Lemma 4.1 *Let $q > 1$ be an integer and let $p \in (0,1)$. Let $E_1, \ldots, E_r$ be pairwise disjoint subsets of $[m]$ each of cardinality d. Let $Q(Z_1, \ldots, Z_m) \in \mathbb{Z}_q[Z_1, \ldots, Z_m]$ be a polynomial of the form*

$$\left(\sum_{j=1}^{r} a_j \prod_{i \in E_j} Z_i \right) + R(\mathbf{Z}),$$

where each $a_j \neq 0$ and $\deg(R(\mathbf{Z})) < d$. Let $\mathbf{z} = (z_1, \ldots, z_m) \in \mathbb{Z}_q^m$ be the random variable where, independently for each i, we have $\Pr[z_i = 1] = p$ and $\Pr[z_i = 0] = 1 - p$. Then,

$$\left| \mathbb{E}\left[\omega^{Q(\mathbf{z})} \right] \right| \leq 2^{-\Omega_{q,p,d}(r)}.$$

Given Lemma 4.1, we may now prove Lemma 3.3.

Proof of Lemma 3.3: Let $U = \cup_{j=1}^{r} E_j$. Fix any $x \in \{0,1\}^{[m]\setminus U}$, and let $Q_x(\mathbf{Y}) \in \mathbb{Z}_q[(Y_i)_{i \in U}]$ be the polynomial

$$\sum_{S \in \mathcal{F}} a_S \left(\prod_{j \in S \cap ([m]\setminus U)} x_j \right) \left(\prod_{i \in S \cap U} Y_i \right) + Q'(x, \mathbf{Y})$$

so that $Q_x(y) = Q(x,y)$ for each $y \in \mathbb{Z}_q^U$. Notice that the degree (in $\mathbf{Y}$) of the term corresponding to $S \in \mathcal{F}$ is $|S \cap U|$. By assumption, unless $S = E_j$ for some j, we must have $|S \cap U| < d$.

Therefore the polynomial $Q_x(\mathbf{Y})$ is of the form:

$$\sum_{j=1}^{r} a_{E_i} \prod_{i \in E_i} Y_i + R(\mathbf{Y}),$$

where $\deg(R(\mathbf{Y})) < d$. By Lemma 4.1,

$$\left| \mathbb{E}\left[\omega^{Q_x(\mathbf{y})}\right] \right| < 2^{-\Omega_{q,p,d}(r)},$$

where $\mathbf{y} \in \{0,1\}^U$ with each $y_i = 1$ independently with probability p.

As $Q_x(y) = Q(x,y)$, we get

$$\left| \mathbb{E}\left[\omega^{Q(\mathbf{z}^x)}\right] \right| < 2^{-\Omega_{q,p,d}(r)},$$

where $\mathbf{z}^x \in \mathbb{Z}_q^n$ is the random variable $\mathbf{z}$ conditioned on the event $z_j = x_j$ for every $j \in [m] \backslash U$. Now, the distribution of $\mathbf{z}$ is a convex combination of the distributions of $\mathbf{z}^x$ as x varies over $\{0,1\}^{[m]\backslash U}$. This allows us to deduce that

$$\left| \mathbb{E}\left[\omega^{Q(\mathbf{z})}\right] \right| \leq 2^{-\Omega_{q,p,d}(r)},$$

as desired. □

4.1 The μ-Gowers norm

The proof of Lemma 4.1 will use a variant of the Gowers norms. Let $Q : \mathbb{Z}_q^m \to \mathbb{Z}_q$ be any function, and define $f : \mathbb{Z}_q^m \to \mathbb{C}$ by $f(x) = \omega^{Q(x)}$. The Gowers norm of f is an analytic quantity that measures how well Q correlates with degree d polynomials: the correlation of Q with polynomials of degree $d-1$ under the uniform distribution is bounded from above by the d^{th}-Gowers norm of f. Thus to show that a certain Q is uncorrelated with all degree $d-1$ polynomials under the uniform distribution, it suffices to bound the d^{th}-Gowers norm of f. In Lemma 4.1, we wish to show that a certain Q is uncorrelated with all degree $d-1$ polynomials under a distribution μ that need not be uniform. To this end, we define a variant of the Gowers norm, which we call the μ-Gowers norm, and show that if the $(d,\mu)^{\text{th}}$-Gowers norm of f is small, then Q is uncorrelated with all degree $d-1$ polynomials *under* μ. We then complete the proof of Lemma 4.1 by bounding the $(d,\mu)^{\text{th}}$-Gowers norm of the relevant f.

We first define the μ-Gowers norm and develop some of its basic properties.

Let H be an abelian group and let μ be a probability distribution on H. For each $d \geq 0$, define a probability distribution $\mu^{(d)}$ on H^{d+1} inductively by $\mu^{(0)} = \mu$, and, for $d \geq 1$, let $\mu^{(d)}(x, t_1, \ldots, t_d)$ equal

$$\frac{\mu^{(d-1)}(x, t_1, \ldots, t_{d-1}) \cdot \mu^{(d-1)}(x + t_d, t_1, \ldots, t_{d-1})}{\sum_{z \in H} \mu^{(d-1)}(z, t_1, \ldots, t_{d-1})}.$$

Equivalently, to sample $(x, t_1, \ldots, t_d)$ from $\mu^{(d)}$, first take a sample $(x, t_1, \ldots, t_{d-1})$ from $\mu^{(d-1)}$, then take a sample $(y, t'_1, \ldots, t'_{d-1})$ from $\mu^{(d-1)}$ conditioned on $t'_i = t_i$ for each $i \in [d-1]$, and finally set $t_d = y - x$ (our sample is then $(x, t_1, \ldots, t_{d-1}, t_d)$). Notice that the distribution of a sample $(x, t_1, \ldots, t_d)$ from $\mu^{(d)}$ is such that for each $S \subseteq [d]$, the distribution of the point $x + \sum_{i \in S} t_i$ is precisely μ.

For a function $f : H \to \mathbb{C}$ and $\mathbf{t} \in H^d$, we define its d^{th}-*derivative in directions* $\mathbf{t}$ to be the function $D_{\mathbf{t}} f : H \to \mathbb{C}$ given by

$$D_{\mathbf{t}} f(x) = \prod_{S \subseteq [d]} f(x + \sum_{i \in S} t_i)^{\circ S},$$

where $a^{\circ S}$ equals the complex conjugate $\bar{a}$ if $|S|$ is odd, and $a^{\circ S}$ equals a otherwise. From the definition it immediately follows that $D_{(\mathbf{t},u)} f(x) = D_{\mathbf{t}} f(x) \overline{D_{\mathbf{t}} f(x+u)}$ (where $(\mathbf{t}, u)$ denotes the vector $(t_1, \ldots, t_d, u) \in H^{d+1}$).

We now define the μ-Gowers norm.

Definition 4.2 (μ-Gowers Norm) *If μ is a distribution on H, and $f : H \to \mathbb{C}$, we define its (d, μ)-Gowers norm by*

$$\|f\|_{U^d,\mu} = \left| \mathbb{E}_{(x,\mathbf{t}) \sim \mu^{(d)}} \left[(D_{\mathbf{t}} f)(x) \right] \right|^{\frac{1}{2^d}}.$$

When H is of the form $\mathbb{Z}_q^m$, then the (d, μ)-Gowers norm of a function is supposed to estimate the correlation, under μ, of that function with polynomials of degree $d-1$. Intuitively, this happens because the Gowers norm of f measures how often the d^{th} derivative of f vanishes.

The next few lemmas enumerate some of the useful properties that μ-Gowers norms enjoy.

Lemma 4.3 *Let $f : H \to \mathbb{C}$. Then,*

$$|\mathbb{E}_{x \sim \mu} [f(x)]| \leq \|f\|_{U^d,\mu}.$$

Proof We prove that for every d, $\|f\|_{U^d,\mu} \leq \|f\|_{U^{d+1},\mu}$. The lemma follows by noting that $\|f\|_{U^0,\mu} = |\mathbb{E}_{x \sim \mu} [f(x)]|$.

The proof proceeds (following Gowers [Gow01] and Green-Tao [GT08]) via the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|f\|_{U^d,\mu}^{2^{d+1}} &= \left| \mathbb{E}_{(x,\mathbf{t}) \sim \mu^{(d)}} \left[D_{\mathbf{t}} f(x) \right] \right|^2 \\
&\leq \mathbb{E}_{\mathbf{t}} \left[\left| \mathbb{E}_x \left[D_{\mathbf{t}} f(x) \right] \right|^2 \right] && \text{by Cauchy-Schwarz} \\
&= \mathbb{E}_{\mathbf{t}} \mathbb{E}_{x,y} \left[D_{\mathbf{t}} f(x) \overline{D_{\mathbf{t}} f(y)} \right] && \text{where } y \text{ is an independent sample of } x \text{ given } \mathbf{t}. \\
&= \mathbb{E}_{x,\mathbf{t},u} \left[D_{\mathbf{t}} f(x) \overline{D_{\mathbf{t}} f(x+u)} \right] && \text{where } u = y - x \\
&= \mathbb{E}_{(x,\mathbf{t},u) \sim \mu^{(d+1)}} \left[D_{\mathbf{t}} f(x) \overline{D_{\mathbf{t}} f(x+u)} \right] && \text{by definition of } \mu^{(d+1)} \\
&= \mathbb{E}_{(x,\mathbf{t},u) \sim \mu^{(d+1)}} \left[D_{(\mathbf{t},u)} f(x) \right] \\
&= \|f\|_{U^{d+1},\mu}^{2^{d+1}}.
\end{aligned}$$

This proves the lemma. □

Definition 4.4 *For each $i \in [r]$, let $g_i : H \to \mathbb{C}$. We define $(\bigotimes_{i=1}^r g_i) : H^r \to \mathbb{C}$ by*

$$\left(\bigotimes_{i=1}^{r} g_i\right)(x_1, \ldots, x_r) = \prod_{i=1}^{r} g_i(x_i).$$

For each $i \in [r]$, let μ_i be a probability measure on H. We define the probability measure $\bigotimes_{i=1}^r \mu_i$ on H^r by

$$\left(\bigotimes_{i=1}^{r} \mu_i\right)(x_1, \ldots, x_r) = \prod_{i=1}^{r} \mu_i(x_i).$$

Lemma 4.5 $\| \bigotimes_{i=1}^r g_i \|_{U^d, \bigotimes_{i=1}^r \mu_i} = \prod_{i=1}^r \|g_i\|_{U^d, \mu_i}$.

Proof Follows by expanding both sides and using the fact that $(\bigotimes_{i=1}^r \mu_i)^{(d)} = \bigotimes_{i=1}^r \left(\mu_i^{(d)}\right)$. □

Lemma 4.6 *Let $q > 1$ be an integer and let $\omega \in \mathbb{C}$ be a primitive q^{th}-root of unity. For all $f : \mathbb{Z}_q^n \to \mathbb{C}$, all probability measures μ on $\mathbb{Z}_q^n$, and all polynomials $h \in \mathbb{Z}_q[Y_1, \ldots, Y_n]$ of degree $< d$,*

$$\|f\omega^h\|_{U^d, \mu} = \|f\|_{U^d, \mu}.$$

The above lemma follows from the fact that $(D_{\mathbf{t}} f) = (D_{\mathbf{t}}(f \cdot \omega^h))$.

Lemma 4.7 *Let $a \in \mathbb{Z}_q \setminus \{0\}$ and let $g : \mathbb{Z}_q^d \to \mathbb{C}$ be given by $g(y) = \omega^{a \prod_{i=1}^d y_i}$. Let μ be a probability distribution on $\mathbb{Z}_q^d$ with $\mathsf{supp}(\mu) \supseteq \{0,1\}^d$. Then $\|g\|_{U^d, \mu} < 1 - \epsilon$, where $\epsilon > 0$ depends only on q, d and μ.*

Proof As $\{0,1\} \subseteq \mathsf{supp}(\mu)$, the distribution $\mu^{(d)}$ give some positive probability $\delta > 0$ to the point $(x_0, \mathbf{e}) = (x_0, e_1, \ldots, e_d)$, where $x_0 = 0 \in \mathbb{Z}_q^d$, and $e_i \in \mathbb{Z}_q^d$ is the vector with 1 in the ith coordinate and 0 in all other coordinates (and δ depends only on q, d and μ). Then $(D_{\mathbf{e}} g)(x_0) = \prod_{S \subseteq [d]} g(\sum_{i \in S} e_i)^{\circ S} = \omega^{\pm a} \neq 1$ (since whenever $S \neq [d]$, we have $g(\sum_{i \in S} e_i) = 1$). On the other hand, whenever $\mathbf{t} \in (\mathbb{Z}_q^d)^d$ has some coordinate equal to 0, which also happens with positive probability depending only on d, μ and q, we have $(D_{\mathbf{t}} g(x)) = 1$. Thus in the expression

$$\|g\|_{U^d, \mu} = \left| \mathbb{E}_{(x, \mathbf{t}) \sim \mu^{(d)}} \left[(D_{\mathbf{t}} f)(x) \right] \right|^{\frac{1}{2^d}},$$

since every term in the expectation has absolute value at most 1, and we just found two terms with positive probability with values 1 and $\omega^{\pm a} \neq 1$, we conclude that $\|g\|_{U^d, \mu} < 1 - \epsilon$ for some ϵ depending only on q, μ and d. □

We now put together the above ingredients.

Theorem 4.8 *Let $f : (\mathbb{Z}_q^d)^r \to \mathbb{C}$ be given by*

$$f(x_1, \ldots, x_r) = \omega^{\sum_{j=1}^r a_j \prod_{i=1}^d x_{ij}},$$

where $a_j \in \mathbb{Z}_q \setminus \{0\}$ for all $j \in [r]$. Let μ be a probability distribution on $\mathbb{Z}_q^d$ with $\mathsf{supp}(\mu) \supseteq \{0,1\}^d$. Then for all polynomials $h \in \mathbb{Z}_q[(Y_{ij})_{i\in[d],j\in[r]}]$, with $\deg(h) < d$, we have

$$\left|\mathbb{E}_{x\sim\mu^{\otimes r}}\left[f(x)\omega^{h(x)}\right]\right| \leq c^r,$$

where $c < 1$ depends only on q, d and μ.

Proof Let $g_j : \mathbb{Z}_q^d \to \mathbb{C}$ be given by $g_j(y) = \omega^{a_j \prod_{i=1}^d y_i}$ (as in in Lemma 4.7), and take $c = 1 - \epsilon$ from that Lemma. Notice that $f = \otimes_{j=1}^r g_j$. Therefore by Lemma 4.5, we have

$$\|f\|_{U^d,\mu^{\otimes r}} = \prod_{j=1}^r \|g_j\|_{U^d,\mu} \leq c^r.$$

As the degree of h is at most $d-1$, Lemma 4.6 implies that

$$\|f\omega^h\|_{U^d,\mu^{\otimes r}} = \|f\|_{U^d,\mu^{\otimes r}} \leq c^r.$$

Lemma 4.3 now implies that

$$\left|\mathbb{E}_{x\sim\mu^{\otimes r}}\left[f(x)\omega^{h(x)}\right]\right| \leq c^r,$$

as desired. □

We can now complete the proof of Lemma 4.1.

Proof of Lemma 4.1: By fixing the variables Z_i for $i \notin \cup_j E_j$, and then averaging over all such fixings, it suffices to consider the case $[m] = \cup_j E_j$. Then the polynomial $Q(Z_1, \ldots, Z_m) = \left(\sum_{j=1}^r a_j \prod_{i\in E_j} Z_i\right) + R(Z)$ can be rewritten in the form (after renaming the variables):

$$\sum_{j=1}^r a_j \prod_{i=1}^d X_{ij} + h(\mathbf{X}),$$

where $\deg(h) < d$. Let μ be the p-biased probability measure on $\{0,1\}^d \subseteq \mathbb{Z}_q^d$. Theorem 4.8 now implies that

$$\left|\mathbb{E}_{x\sim\mu^{\otimes r}}\left[\omega^{Q(x)}\right]\right| \leq 2^{-\Omega_{q,p,d}(r)},$$

as desired. □

5 Proof of Theorem 2.3

The proof of Theorem 2.3 will be via a more general theorem amenable to inductive proof, Theorem 5.8. Just as Theorem 2.3 states that for almost all $G \in G(n,p)$, the truth of any $\mathsf{FO}[\mathsf{Mod}_q]$ sentence on G is determined by subgraph frequencies, freq^c_G, Theorem 5.8 states that for almost all graphs $G \in G(n,p)$, for any $w_1, \ldots, w_k \in V_G$ the truth of any $\mathsf{FO}[\mathsf{Mod}_q]$ *formula* $\varphi(w_1, \ldots, w_k)$ on G is determined by the adjacency and equality information about $w_1, \ldots, w_k$ (which we will call the *type*), and the *labelled subgraph frequencies at* $\mathbf{w}$. In the next subsection, we formalize these notions.

5.1 Labelled graphs and labelled subgraph frequencies

Let I be a finite set. We begin with some preliminaries on I-labelled graphs.

Definition 5.1 (I-labelled graphs) *An* I-labelled graph *is a graph $F = (V_F, E_F)$ where some vertices are labelled by elements of I, such that (a) for each $i \in I$, there is exactly one vertex labelled i. We denote this vertex $F(i)$, and (b) the graph induced on the set of labelled vertices is an independent set. We denote the set of labelled vertices of F by $\mathcal{L}(F)$.*

Definition 5.2 (Homomorphisms and Copies) *A* homomorphism *from an I-labelled graph F to a pair $(G, \mathbf{w})$, where G is a graph and $\mathbf{w} \in V_G^I$, is a homomorphism $\chi \in \mathsf{Hom}(F, G)$ such that for each $i \in I$, χ maps $F(i)$ to w_i. A homomorphism from F to $(G, \mathbf{w})$ is called* injective *if for any distinct $v, w \in V_F$, such that $\{v, w\} \not\subseteq \mathcal{L}(F)$, we have $\chi(v) \neq \chi(w)$. A* copy *of F in $(G, \mathbf{w})$ is a set $E \subseteq E_G$ such that there exists an injective homomorphism χ from F to $(G, \mathbf{w})$ such that $E = \chi(E_F) := \{(\chi(v), \chi(w)) \mid (v, w) \in E_F\}$. An* automorphism *of F is an injective homomorphism from F to $(F, \mathbf{w})$, where $w_i = F(i)$ for each $i \in I$.*

Definition 5.3 ($\mathsf{Hom}, \mathsf{Inj}, \mathsf{Cop}, \mathsf{Aut}$ for labelled graphs) *Let F be an I-labelled graph, and G be any graph. Let $\mathbf{w} \in V_G^I$. We define $\mathsf{Hom}(F, (G, \mathbf{w}))$ to be the set of homomorphisms from F to $(G, \mathbf{w})$. We define $\mathsf{Inj}(F, (G, \mathbf{w}))$ to be the set of injective homomorphisms from F to $(G, \mathbf{w})$. We define $\mathsf{Cop}(F, (G, \mathbf{w}))$ to be the set of copies of F in $(G, \mathbf{w})$. We define $\mathsf{Aut}(F)$ to be the set of automorphisms of F. We let $[F](G, \mathbf{w})$ (respectively $\langle F \rangle (G, \mathbf{w})$, $\mathsf{aut}(F)$) be the cardinality of $\mathsf{Inj}(F, (G, \mathbf{w}))$ (respectively $\mathsf{Cop}(F, (G, \mathbf{w}))$, $\mathsf{Aut}(F)$).*

Finally, let $[F]_q(G, \mathbf{w}) = [F](G, \mathbf{w}) \mod q$ and $\langle F \rangle_q(G, \mathbf{w}) = \langle F \rangle (G, \mathbf{w}) \mod q$.

Definition 5.4 (Label-connected) *For F an I-labelled graph, we say F is* label-connected *if $F \setminus \mathcal{L}(F)$ is connected. Define Conn^t_I to be the set of all I-labelled label-connected graphs with at most t unlabelled vertices. For $i \in I$, we say an I-labelled graph F is* **dependent on label** *i if $F(i)$ is not an isolated vertex.*

Definition 5.5 (Partitions) *If I is a set, an* I-partition *is a set of subsets of I that are pairwise disjoint, and whose union is I. If Π is an I partition, then for $i \in I$ we denote the unique element of Π containing i by $\Pi(i)$. If V is any set and $\mathbf{w} \in V^I$, we say $\mathbf{w}$ respects Π if for all $i, i' \in I$, $w_i = w_{i'}$ iff $\Pi(i) = \Pi(i')$.*

The collection of all partitions of I is denoted $\mathsf{Partitions}(I)$.

If $I \subseteq J$, $\Pi \in \mathsf{Partitions}(I)$ and $\Pi' \in \mathsf{Partitions}(J)$, we say Π' extends Π if for all $i_1, i_2 \in I$, $\Pi(i_1) = \Pi(i_2)$ if and only if $\Pi'(i_1) = \Pi'(i_2)$.

Definition 5.6 (Types) *An I-*type *τ is a pair (Π_τ, E_τ) where $\Pi_\tau \in \mathsf{Partitions}(I)$ and $E_\tau \subseteq \binom{\Pi_\tau}{2}$. For a graph G and $\mathbf{w} \in V_G^I$, we define the* type *of $\mathbf{w}$ in G, denoted $\mathsf{type}_G(\mathbf{w})$, to be the I-*type *τ, where $\mathbf{w}$ respects Π_τ, and for all $i, i' \in I$, $\{\Pi_\tau(i), \Pi_\tau(i')\} \in E_\tau$ if and only if w_i and $w_{i'}$ are adjacent in G.*

The collection of all I-types is denoted $\mathsf{Types}(I)$.

If $I \subseteq J$, and $\tau \in \mathsf{Types}(I)$ and $\tau' \in \mathsf{Types}(J)$, we say τ' *extends* τ if $\Pi_{\tau'}$ extends Π_τ and for each $i_1, i_2 \in I$, $\{\Pi_\tau(i_1), \Pi_\tau(i_2)\} \in E_\tau$ if and only if $\{\Pi_{\tau'}(i_1), \Pi_{\tau'}(i_2)\} \in E_{\tau'}$.

Definition 5.7 (Labelled subgraph frequency vector) *Let G be a graph and I be any set. Let $\mathbf{w} \in V_G^I$. We define the* labelled subgraph frequency vector *at $\mathbf{w}$, $\mathsf{freq}_G^a(\mathbf{w}) \in \mathbb{Z}_q^{\mathsf{Conn}_I^a}$, to be the vector such that for each $F \in \mathsf{Conn}_I^a$,*

$$(\mathsf{freq}_G^a(\mathbf{w}))_F = [F]_q(G, \mathbf{w}).$$

Remark We will often deal with $[k]$-labelled graphs. By abuse of notation we will refer to them as k-labelled graphs. If $\mathbf{w} \in V^{[k]}$ and $u \in V$, when we refer to the tuple $(\mathbf{w}, v)$, we mean the $[k+1]$-tuple whose first k coordinates are given by $\mathbf{w}$ and whose $k+1$st coordinate is v. Abusing notation even further, when we deal with a $[k+1]$-labelled graph F, then by $[F](G, \mathbf{w}, v)$, we mean $[F](G, (\mathbf{w}, v))$. Similarly Conn_k^t denotes $\mathsf{Conn}_{[k]}^t$.

5.2 The quantifier eliminating theorem

We now state Theorem 5.8, from which Theorem 2.3 follows easily. Informally, it says that an $\mathsf{FO}[\mathsf{Mod}_q]$-formula $\varphi(\mathbf{w})$ is essentially determined by the type of $\mathbf{w}$, $\mathsf{type}_G(\mathbf{w})$, and the labelled subgraph frequencies at $\mathbf{w}$, $\mathsf{freq}_G^c(\mathbf{w})$.

Theorem 5.8 *For all primes q and integers $k, t > 0$, there is a constant $c = c(k, t, q)$ such that for every $\mathsf{FO}[\mathsf{Mod}_q]$ formula $\varphi(\alpha_1, \ldots, \alpha_k)$ with quantifier depth t, there is a function $\psi : \mathsf{Types}(k) \times \mathbb{Z}_q^{\mathsf{Conn}_k^c} \to \{0, 1\}$ such that for all $p \in (0, 1)$, the quantity*

$$\Pr_{G \in G(n,p)} \left[\begin{array}{l} \forall w_1, \ldots, w_k \in V_G, \\ \quad (G \text{ satisfies } \varphi(w_1, \ldots, w_k)) \Leftrightarrow (\psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1) \end{array} \right] \geq 1 - 2^{-\Omega(n)}.$$

Putting $k = 0$, we recover Theorem 2.3.

We now give a brief sketch of the proof of Theorem 5.8 (the detailed proof appears in Section 6). The proof is by induction on the size of the formula φ. When the formula φ

has no quantifiers, then the truth of $\varphi(\mathbf{w})$ on G is completely determined by $\mathsf{type}_G(\mathbf{w})$. The case where φ is of the form $\varphi_1(\alpha_1,\ldots,\alpha_k) \wedge \varphi_2(\alpha_1,\ldots,\alpha_k)$ is easily handled via the induction hypothesis. The case where $\varphi(\alpha_1,\ldots,\alpha_k) = \neg\varphi_1(\alpha_1,\ldots,\alpha_k)$ is similar.

The key cases for us to handle are thus *(i)* $\varphi(\alpha_1,\ldots,\alpha_k)$ is of the form $\mathsf{Mod}_q^i\beta, \varphi'(\alpha_1,\ldots,\alpha_k,\beta)$, and *(ii)* $\varphi(\alpha_1,\ldots,\alpha_k)$ is of the form $\exists\beta, \varphi'(\alpha_1,\ldots,\alpha_k,\beta)$. We now give a sketch of how these cases may be handled.

For case (i), let $\psi' : \mathsf{Types}(k+1)\times\mathbb{Z}_q^{\mathsf{Conn}_{k+1}^b}$ be the function given by the induction hypothesis for the formula φ'. Thus for most graphs $G \in G(n,p)$ (namely the ones for which ψ' is good for φ'), $\varphi(w_1,\ldots,w_k)$ is true if and only the number of vertices $v \in V_G$ such that $\psi'(\mathsf{type}_G(\mathbf{w},v), \mathsf{freq}_G^b(\mathbf{w},v)) = 1$ is congruent to i mod q. In Theorem 6.1 (whose proof appears in Section 7), we show that the number of such vertices v can be determined solely as a function of $\mathsf{type}_G(\mathbf{w})$ and $\mathsf{freq}_G^a(\mathbf{w})$ for suitable a. This fact allows us to define ψ in a natural way, and this completes case (i).

Case (ii) is the most technically involved case. As before, we get a function ψ' corresponding to φ' by the induction hypothesis. We show that one can define ψ essentially as follows: define $\psi(\tau,f) = 1$ if there exists some $(\tau',f') \in \mathsf{Types}(k+1)\times\mathbb{Z}_q^{\mathsf{Conn}_{k+1}^b}$ that "extends" (τ,f) for which $\psi'(\tau',f') = 1$; otherwise $\psi(\tau,f) = 0$. Informally, we show that if it is conceivable that there is a vertex v such that $\varphi'(\mathbf{w},v)$ is true, then $\varphi(\mathbf{w})$ is almost surely true. Proving this statement requires us to get a characterization of the distribution of labelled subgraph frequencies, significantly generalizing Theorem 2.4. This is done in Theorem 6.12 (whose proof appears in Section 8).

6 Quantifier Elimination

In this section, we give a full proof of Theorem 5.8. Before doing so, we state the main technical theorems: Theorem 6.1 (which is needed for eliminating Mod_q quantifiers), and Theorem 6.12 (which is needed for eliminating $\exists$ quantifiers). We do this in the following two subsections.

6.1 Counting extensions

The next theorem plays a crucial role in the elimination of the Mod_q quantifiers. This is the only step where the assumption that q is a prime plays a role in the modular convergence law.

Theorem 6.1 *Let q be a prime, let $k,b > 0$ be integers and let $a \geq (q-1)\cdot b\cdot|\mathsf{Conn}_{k+1}^b|$. There is a function*

$$\lambda : \mathsf{Types}(k+1)\times\mathbb{Z}_q^{\mathsf{Conn}_{k+1}^b}\times\mathsf{Types}(k)\times\mathbb{Z}_q^{\mathsf{Conn}_k^a}\to\mathbb{Z}_q$$

such that for all $\tau' \in \mathsf{Types}(k+1)$, $f' \in \mathbb{Z}_q^{\mathsf{Conn}_{k+1}^b}$, $\tau \in \mathsf{Types}(k)$, $f \in \mathbb{Z}_q^{\mathsf{Conn}_k^a}$, it holds that for every graph G, and every $w_1,\ldots,w_k \in V_G$ with $\mathsf{type}_G(\mathbf{w}) = \tau$ and $\mathsf{freq}_G^a(\mathbf{w}) = f$, the

cardinality of the set

$$\{v \in V_G : \mathsf{type}_G(\mathbf{w}, v) = \tau' \wedge \mathsf{freq}^b_G(\mathbf{w}, v) = f'\}$$

is congruent to $\lambda(\tau', f', \tau, f) \mod q$.

The proof appears in Section 7. The principal ingredient in its proof is the following lemma, which states that the numbers $[F](G, \mathbf{w})$, as F varies over small label-connected graphs, determine the number $[F'](G, \mathbf{w})$ for all small graphs F'.

Lemma 6.2 (Label-connected subgraph frequencies determine all subgraph frequencies) *For every k-labelled graph F' with $|V_{F'} \setminus \mathcal{L}(F')| \leq t$, there is a polynomial $\delta_{F'} \in \mathbb{Z}[(X_F)_{F \in \mathsf{Conn}^t_k}]$ such that for all graphs G and $\mathbf{w} \in V_G^k$,*

$$[F'](G, \mathbf{w}) = \delta_{F'}(x),$$

where $x \in \mathbb{Z}^{\mathsf{Conn}^t_k}$ is given by $x_F = [F](G, \mathbf{w})$.

6.2 The distribution of labelled subgraph frequencies mod q

In this subsection, we state the theorem that will help us eliminate $\exists$ quantifiers. Let us first give an informal description of the theorem. We are given a tuple $\mathbf{w} \in [n]^k$, and distinct $u_1, \ldots, u_s \in [n] \setminus \{w_1, \ldots, w_k\}$. Let G be sampled from $G(n,p)$ (recall that we think of $G(n,p)$ as a random graph whose vertex set is $[n]$: thus the w_i and u_j are vertices of G). The theorem completely describes the joint distribution of the labelled subgraph frequency vectors at all the tuples $\mathbf{w}$, $(\mathbf{w}, u_1), \ldots, (\mathbf{w}, u_s)$; namely it pins down the distribution of $(\mathsf{freq}^a_G(\mathbf{w}), \mathsf{freq}^b_G(\mathbf{w}, u_1), \ldots, \mathsf{freq}^b_G(\mathbf{w}, u_s))$. We first give a suitable definition of the set of *feasible frequency vectors*, and then claim that (a) the $\mathsf{freq}^a_G(\mathbf{w})$ is essentially uniformly distributed over the set of its feasible frequency vectors, and (b) conditioned on $\mathsf{freq}^a_G(\mathbf{w})$, the distributions of $\mathsf{freq}^b_G(\mathbf{w}, u_1), \ldots, \mathsf{freq}^b_G(\mathbf{w}, u_s)$ are all essentially independent and uniformly distributed over the set of those feasible frequency vectors that are "consistent" with freq^a_G.

To define the set of feasible frequency vectors (which will equal the set of all possible values that $\mathsf{freq}^a_G(\mathbf{w})$ may assume), there are two factors that come into play. The first factor, one that we already encountered while dealing with unlabelled graphs, is a divisibility constraint: the number $[F](G, \mathbf{w})$ is always divisible by a certain integer depending on F, and hence for some F, it cannot assume arbitrary values mod q. The second factor is a bit subtler: when $w_1, \ldots, w_k$ are not all distinct, for certain pairs F, F' of label-connected k-labelled graphs, $[F](G, \mathbf{w})$ is forced to equal $[F'](G, \mathbf{w})$. Let us see a simple example of such a phenomenon. Let $k = 2$ and let $w_1 = w_2$. Let the 2-labelled graph F be a path of length 2 with ends labelled 1 and 2. Let the 2-labelled graph F' be the disjoint union of an edge, one of whose ends is labelled 1, and an isolated vertex labelled 2. Then in any graph G, $[F](G, \mathbf{w}) = [F'](G, \mathbf{w}) =$ the degree of w_1.

In the rest of this subsection, we will build up some notation and results leading up to a definition of feasible frequency vectors and the statement of the main technical theorem describing the distribution of labelled subgraph frequency vectors.

Definition 6.3 (Quotient of a labelled graph by a paritition) *Let F be a I-labelled graph and let $\Pi \in \mathsf{Partitions}(I)$. We define F/Π to be the Π-labelled graph obtained from F by (a) for each $J \in \Pi$, identifying all the vertices with labels in J and labelling this new vertex J, and (b) deleting duplicate edges. If F and F' are I-labelled graphs and $\Pi \in \mathsf{Partitions}(I)$, we say F and F' are Π-equivalent if $F/\Pi \cong F'/\Pi$.*

Let $\mathbf{w} \in V_G^I$. Let $\Pi \in \mathsf{Partitions}(I)$ be such that $\mathbf{w}$ respects Π. Define $(\mathbf{w}/\Pi) \in V_G^\Pi$ by: for each $J \in \Pi$, $(\mathbf{w}/\Pi)_J = w_j$, where j is any element of J (this definition is independent of the choice of $j \in J$). Observe that as J varies over Π, the vertices $(w/\Pi)_J$ are all distinct.

The next two lemmas show that the numbers $[F](G, \mathbf{w})$ must satisfy certain constraints. These constraints will eventually motivate our definition of feasible frequency vectors.

Lemma 6.4 *If G is a graph and $\mathbf{w} \in V_G^I$, with $\mathbf{w}$ respecting $\Pi \in \mathsf{Partitions}(I)$, then for any I-labelled F,*

$$[F](G, \mathbf{w}) = [F/\Pi](G, (\mathbf{w}/\Pi)). \tag{2}$$

Proof We define a bijection $\alpha : \mathsf{Inj}(F/\Pi, (G, \mathbf{w}/\Pi)) \to \mathsf{Inj}(F, (G, \mathbf{w}))$. Let $\pi \in \mathsf{Hom}(F, F/\Pi)$ be the natural homomorphism sending each unlabelled vertex in V_F to its corresponding vertex in $V_{F/\Pi}$, and, for each $i \in I$ sending $F(i)$ to $(F/\Pi)(\Pi(i))$. We define $\alpha(\chi)$ to be $\chi \circ \pi$.

Take distinct $\chi, \chi' \in \mathsf{Inj}(F/\Pi, (G, \mathbf{w}/\Pi))$. Let $u \in V_{F/\Pi}$ with $\chi(u) \neq \chi'(u)$. Note that u cannot be an element of $\mathcal{L}(F/\Pi)$, for if $u = (F/\Pi)(\Pi(i))$, then $\chi(u) = \chi'(u) = w_i$. Thus $u \notin \mathcal{L}(F/\Pi)$. Let $v \in V_F$ be the vertex $\pi^{-1}(u)$ (which is uniquely specified since $u \notin \mathcal{L}(F/\Pi)$). Thus we have $\chi(\pi(v)) = \chi(u) \neq \chi'(u) = \chi'(\pi(v))$. Thus $\alpha(\chi) \neq \alpha(\chi')$, and α is one-to-one.

To show that α is onto, take any $\chi \in \mathsf{Inj}(F, (G, \mathbf{w}))$. Define $\chi' \in \mathsf{Inj}(F/\Pi, (G, \mathbf{w}/\Pi))$ by:

1. $\chi'(u) = \chi(\pi^{-1}(u))$ if $u \notin \mathcal{L}(F/\Pi)$.
2. $\chi'(u) = w_j$ for any $j \in J$, if $u = (F/\Pi)(J)$ with $J \in \Pi$.

Then $\alpha(\chi') = \chi$. □

Lemma 6.5 *Let G be a graph and $\mathbf{w} \in V_G^I$. Suppose all the $(w_i)_{i \in I}$ are distinct. Let F be an I-labelled label-connected graph with $|E_F| \geq 1$. Then*

$$[F](G, \mathbf{w}) = \mathsf{aut}(F) \cdot \langle F \rangle (G, \mathbf{w}).$$

Proof We give a bijection $\alpha : \mathsf{Aut}(F) \times \mathsf{Cop}(F, (G, \mathbf{w})) \to \mathsf{Inj}(F, (G, \mathbf{w}))$.

For each $E \in \mathsf{Cop}(F, (G, \mathbf{w}))$, we fix a $\chi_E \in \mathsf{Inj}(F, (G, \mathbf{w}))$ such that $\chi_E(E_F) = E$. Then we define $\alpha(\sigma, E) = \chi_E \circ \sigma$.

First notice that $\alpha(\sigma, E)(E_F) = \chi_E(\sigma(E_F)) = \chi_E(E_F) = E$. Thus if $\alpha(\sigma, E) = \alpha(\sigma', E')$, then $E = E'$. But since χ_E is injective, for any $\sigma \neq \sigma'$, we have $\chi_E \circ \sigma \neq \chi_E \circ \sigma'$. Thus α is one-to-one.

To show that α is onto, take any $\chi \in \mathsf{Inj}(F, (G, \mathbf{w}))$. Let $E = \chi(E_F)$. As F is label-connected and $\chi_E(E_F) = \chi(E_F)$, we have $\chi_E(V_F) = \chi(V_F)$. We may now define $\sigma \in \mathsf{Aut}(F)$ by $\sigma(u) = \chi_E^{-1}(\chi(u))$ for each $u \in V_F$. Clearly, $\alpha(\sigma, E) = \chi$, and so α is onto.

Thus α is a bijection, and the lemma follows. □

Note that Lemma 2.2 and Lemma 3.1 follow formally from the above lemma.

Let $\mathrm{K}_1(I)$ be the I-labelled graph with $|I|+1$ vertices: $|I|$ labelled vertices and one isolated unlabelled vertex. The role of $\mathrm{K}_1(I)$ in the I-labelled theory is similar to the role of K_1 in the unlabelled case.

Definition 6.6 (Feasible frequency vectors) *We define the set of* feasible frequency vectors*,* $\mathsf{FFreq}(\tau, I, a)$ *to be the set of* $f \in \mathbb{Z}_q^{\mathsf{Conn}_I^a}$ *such that*

(a) for any $F \in \mathsf{Conn}_I^a$*, we have* $f_F \in \mathsf{aut}(F/\Pi_\tau) \cdot \mathbb{Z}_q$*.*

(b) for any $F, F' \in \mathsf{Conn}_I^a$ *that are* Π_τ*-equivalent, we have* $f_F = f_{F'}$*.*

Let $\mathsf{FFreq}_n(\tau, I, a)$ *be the set* $\{f \in \mathsf{FFreq}(\tau, I, a) : f_{K_1(I)} = n - |\Pi_\tau| \mod q\}$*. Note that if* $n = n' \mod q$*, then* $\mathsf{FFreq}_n(\tau, I, a) = \mathsf{FFreq}_{n'}(\tau, I, a)$*.*

Observe that for any $\mathbf{w} \in V_G^I$ with $\mathsf{type}_G(\mathbf{w}) = \tau$, the vector $\mathsf{freq}_G^a(\mathbf{w})$ is an element of $\mathsf{FFreq}(\tau, I, a)$. This follows from Lemma 6.4 and Lemma 6.5, which allow us to deduce (recall that $(\mathbf{w}/\Pi_\tau)_J$ are all distinct for $J \in \Pi_\tau$) that for any $F \in \mathsf{Conn}_I^a$,

$$[F](G, \mathbf{w}) = \mathsf{aut}(F/\Pi_\tau) \cdot \langle F/\Pi_\tau \rangle (G, \mathbf{w}/\Pi_\tau). \tag{3}$$

Observe also that if $|V_G| = n$, then $\mathsf{freq}_G^a(\mathbf{w}) \in \mathsf{FFreq}_n(\tau, I, a)$, since $[K_1(I)](G, \mathbf{w}) = |V_G \setminus \{w_1, \ldots, w_k\}| = n - |\Pi_{\mathsf{type}(\mathbf{w})}|$, as required by the definition.

Definition 6.7 (Extending) *Let* I *be a set and let* $J = I \cup \{i^*\}$*. Let* $a \geq b > 0$ *be positive integers. We say* $(\tau', f') \in \mathsf{Types}(J) \times \mathsf{FFreq}(\tau', J, b)$ extends $(\tau, f) \in \mathsf{Types}(I) \times \mathsf{FFreq}(\tau, I, a)$ *if* τ' *extends* τ*, and for every* $F \in \mathsf{Conn}_I^b$*, we have*

1. *if* $\{i^*\} \notin \Pi_{\tau'}$*,*

$$f_F = f'_{\widetilde{F}}, \tag{4}$$

where $\widetilde{F}$ *is the graph obtained from* F *by introducing an isolated vertex labelled* i^**.*

2. *if* $\{i^*\} \in \Pi_{\tau'}$*, letting* $\delta_H : \mathbb{Z}_q^{\mathsf{Conn}_J^b} \to \mathbb{Z}_q$ *be the function given by Lemma 6.2,*

$$f_F = f'_{\widetilde{F}} + \sum_{u \in V_F \setminus \mathcal{L}(F)} c_u \delta_{F_u}(f'), \tag{5}$$

where

- *$\widetilde{F}$ is the graph obtained from F by introducing an isolated vertex labelled i^*.*
- *c_u equals 1 if for all $i \in I$, if u is adjacent to $F(i)$, then $\{\Pi_{\tau'}(i^*), \Pi_{\tau'}(i)\} \in E_{\tau'}$. Otherwise, $c_u = 0$.*
- *F_u is the graph obtained from F by labelling the vertex u by i^* and deleting all edges between u and the other labelled vertices of F.*

The crux of the above definition is captured in the following lemma.

Lemma 6.8 *Let G be a graph. Let $a \geq b > 0$ be integers. Let $\mathbf{w} \in V^k$ and $v \in V$. Let $\tau = \mathsf{type}_G(\mathbf{w})$, $\tau' = \mathsf{type}_G(\mathbf{w}, v)$, $f = \mathsf{freq}^a_G(\mathbf{w})$ and $f' = \mathsf{freq}^b_G(\mathbf{w}, v)$. Then (τ', f') extends (τ, f).*

Proof We keep the notation of the previous definition. First observe that τ' extends τ.

If $\{k+1\} \notin \Pi_{\tau'}$, then we need to show that $[F]_q(G, \mathbf{w}) = [\widetilde{F}]_q(G, \mathbf{w}, v)$ for each $F \in \mathsf{Conn}^b_k$. This is immediate from the definitions.

If $\{k+1\} \in \Pi_{\tau'}$, then we need to show that $[F]_q(G, \mathbf{w}) = [\widetilde{F}]_q(G, \mathbf{w}, v) + \sum_{u \in V_F \setminus \mathcal{L}(F)} c_u [F_u]_q(G, \mathbf{w}, v)$. We do this by counting the $\chi \in \mathsf{Inj}(F, (G, \mathbf{w}))$ based on its image $\chi(V_F)$ as follows:

1. Category 1: $v \notin \chi(V_F)$. There are precisely $[\widetilde{F}](G, \mathbf{w}, v)$ such χ.
2. Category 2: $v = \chi(u)$ (in this case u is uniquely specified). Note that $u \notin \mathcal{L}(F)$. Then it must be the case that for any $i \in [k]$ such that u is adjacent to $F(i)$, w_i is adjacent to v. Thus $\{\Pi_{\tau'}(i), \Pi_{\tau'}(k+1)\} \in E_{\tau'}$, and so $c_u = 1$. The number of such χ is $[F_u](G, \mathbf{w}, v)$.

This proves the desired relation. □

We now state and prove two key uniqueness properties enjoyed by the notion of extension.

Lemma 6.9 *Let $a \geq b > 0$ be integers. Let $\mathbf{w} \in V_G^k$. Let $u \in V_G \setminus \{w_1, \ldots, w_k\}$. Let $\tau = \mathsf{type}_G(\mathbf{w})$ and $\tau' = \mathsf{type}_G(\mathbf{w}, u)$. Let $f = \mathsf{freq}^a_G(\mathbf{w})$. Then $\mathsf{freq}^b_G(\mathbf{w}, u)$ is the unique $f' \in \mathbb{Z}_q^{\mathsf{Conn}^b_{k+1}}$ such that:*

- *for each $H \in \mathsf{Conn}^b_{k+1}$ that is dependent on label $k+1$, we have $f'_H = [H]_q(G, \mathbf{w}, u)$.*
- *(τ', f') extends (τ, f).*

Proof By Lemma 6.8, the vector $\mathsf{freq}^b_G(\mathbf{w}, u)$ is such an f'.

To prove uniqueness, it suffices to show that any f' satisying these two properties equals $\mathsf{freq}^b_G(\mathbf{w}, u)$. Thus it suffices to show that for any $H \in \mathsf{Conn}^b_{k+1}$ not dependent on label $k+1$, $f'_H = (\mathsf{freq}^b_G(\mathbf{w}, u))_H$.

We prove this by induction on $|V_H \setminus \mathcal{L}(H)|$. Let $H \in \mathsf{Conn}^b_{k+1}$ not dependent on label $k+1$. Thus H is of the form $\widetilde{F}$ for some graph $F \in \mathsf{Conn}^b_k$ (as in the previous lemma, for a $[k]$-labelled graph F, we let $\widetilde{F}$ be the $[k+1]$-labelled graph obtained by adjoining an isolated vertex labelled $k+1$ to F). By Equation (5), we see that f'_H is *uniquely* determined by τ, τ', f_F and the numbers $(f'_{H'})_{H' \in \mathsf{Conn}^{|V_H \setminus \mathcal{L}(H)|-1}_{k+1}}$ (since each c_u is determined by τ' and each of the graphs F_u have $|F_u \setminus \mathcal{L}(F_u)| \leq |V_H \setminus \mathcal{L}(H)| - 1$). By induction hypothesis, all the $f'_{H'} = (\mathsf{freq}^b_G(\mathbf{w}, u))_{H'}$. Thus, since $\mathsf{freq}^b_G(\mathbf{w}, u)$ also satisfies Equation (5), we have $f'_H = (\mathsf{freq}^b_G(\mathbf{w}, u))_H$, as required. □

Lemma 6.10 *Let $a \geq b > 0$ be integers. Let $(\tau, f) \in \mathsf{Types}(k) \times \mathsf{FFreq}(\tau, [k], a)$. Let $\tau' \in \mathsf{Types}(k+1)$ extend τ with $\{k+1\} \notin \Pi_{\tau'}$. Then there is at most one $f' \in \mathsf{FFreq}(\tau', [k+1], b)$ such that (τ', f') extends (τ, f).*

Proof As in the previous lemma, for a $[k]$-labelled graph F, we let $\widetilde{F}$ be the $[k+1]$-labelled graph obtained by adjoining an isolated vertex labelled $k+1$ to F. For any $F \in \mathsf{Conn}^b_k$, we must have $f'_{\widetilde{F}} = f_F$. Now we claim that any $H \in \mathsf{Conn}^b_k$ is Π-equivalent to some graph of the form $\widetilde{F}$. To prove this, let $j \in [k]$ be such that $\Pi_{\tau'}(j) = \Pi_{\tau'}(k+1)$. Let H^* be the graph obtained from H by adding, for each neighbor u of $H(k+1)$, an edge between u and the $H(j)$, and then removing (a) all edges incident on $H(k+1)$, and (b) any duplicate edges introduced. By construction, $H/\Pi_{\tau'} \cong H^*/\Pi_{\tau'}$, and so $f'_H = f'_{H^*}$ by Equation (3). In addition, the $H^*(k+1)$ is isolated, and hence H^* is of the form $\widetilde{F}$ for some $F \in \mathsf{Conn}^b_k$. What we have shown is that for every $H \in \mathsf{Conn}^b_{k+1}$, f'_H is forced to equal f_F for some $F \in \mathsf{Conn}^b_k$. This implies that f' is specified uniquely. □

Finally, we will need to deal with random graphs $G(n,p)$ with some of the edges already exposed. The next definition captures this object.

Definition 6.11 (Conditioned Random Graph) *Let $A = (V_A, E_A)$ be a graph with $V_A \subseteq [n]$. We define the* conditioned random graph $G(n, p \mid V_A, E_A)$ *to be the graph $G = (V_G, E_G)$ with $V_G = [n]$ and $E_G = E_A \cup E'$, where each $\{i,j\} \in \binom{[n]}{2} \setminus \binom{V_A}{2}$ is included in E' independently with probability p.*

We can now state the main technical theorem that describes the distribution of labelled subgraph frequencies, and will eventually be useful for eliminating $\exists$ quantifiers.

Theorem 6.12 *Let $a \geq b$ be positive integers. Let A be a graph with $V_A \subseteq [n]$ and $|V_A| \leq n' \leq n/2$. Let $G \in G(n, p \mid V_A, E_A)$. Let $\mathbf{w} = (w_1, \ldots, w_k) \in V_A^k$, and let $u_1, \ldots, u_s \in V_A \setminus \{w_1, \ldots, w_k\}$ be distinct. Let $\tau = \mathsf{type}_G(\mathbf{w})$ and let $\tau_i = \mathsf{type}_G(\mathbf{w}, u_i)$ (note that $\tau, \tau_1, \ldots, \tau_s$ are already determined by E_A). Let f denote the random variable $\mathsf{freq}^a_G(\mathbf{w})$. Let f_i denote the random variable $\mathsf{freq}^b_G(\mathbf{w}, u_i)$.*

Then, there exists a constant $\rho = \rho(a, q, p) > 0$, such that if $s \leq \rho \cdot n$, then the distribution of $(f, f_1, \ldots, f_s)$ over $\mathsf{FFreq}_n(\tau, [k], a) \times \prod_i \mathsf{FFreq}_n(\tau_i, [k+1], b)$ is $2^{-\Omega(n)}$-close to the distribution of $(h, h_1, \ldots, h_s)$ generated as follows:

1. *h is picked uniformly at random from* $\mathsf{FFreq}_n(\tau, [k], a)$.
2. *For each i, each h_i is picked independently and uniformly from the set of all $f' \in \mathsf{FFreq}_n(\tau_i, [k+1], b)$ such that (τ_i, f') extends (τ, h).*

6.3 Proof of Theorem 5.8

We now prove Theorem 5.8, where the main quantifier elimination step is carried out.

Theorem 5.8 (restated) *For every prime q and integers $k, t > 0$, there is a constant $c = c(k, t, q)$ such that for every $\mathsf{FO}[\mathsf{Mod}_q]$ formula $\varphi(\alpha_1, \ldots, \alpha_k)$ with quantifier depth t, there is a function $\psi : \mathsf{Types}(k) \times \mathbb{Z}_q^{\mathsf{Conn}_k^c} \to \{0,1\}$ such that for all $p \in (0,1)$, the quantity*

$$\Pr_{G \in G(n,p)} \left[\begin{array}{l} \forall w_1, \ldots, w_k \in V_G, \\ \quad (G \text{ satisfies } \varphi(w_1, \ldots, w_k)) \Leftrightarrow (\psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1) \end{array} \right] \geq 1 - 2^{-\Omega(n)}.$$

Proof The proof is by induction on the size of the formula. If $\varphi(w_1, \ldots, w_k)$ is an atomic formula, then trivially there exists a $\psi : \mathsf{Types}(k) \to \{0,1\}$ such that for every graph G and every $\mathbf{w} \in V_G^k$, the statement $\varphi(w_1, \ldots, w_k)$ holds if and only if $\psi(\mathsf{type}_G(\mathbf{w})) = 1$. Thus we may take $c(k, 0, q) = 0$. We will show that one may take $c(k, t, q) = (q-1)c(k+1, t-1, q) \cdot 2^{c(k+1,t-1,q)^2}$.

Now assume the result holds for all formulae smaller than φ.

Case $\wedge$: Suppose $\varphi(\alpha_1, \ldots, \alpha_k) = \varphi_1(\alpha_1, \ldots, \alpha_k) \wedge \varphi_2(\alpha_1, \ldots, \alpha_k)$. By induction hypothesis, we have functions ψ_1, ψ_2 and a constant c such that $\Pr_G[\forall w_1, \ldots, w_k \in V_G, (\varphi_1(w_1, \ldots, w_k) \Leftrightarrow \psi_1(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)}$ and $\Pr_G[\forall w_1, \ldots, w_k \in V_G, (\varphi_2(w_1, \ldots, w_k) \Leftrightarrow \psi_2(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)}$. Setting $\psi(\tau, f) = \psi_1(\tau, f) \cdot \psi_2(\tau, f)$, it follows from the union bound that

$$\Pr_G[\forall w_1, \ldots, w_k \in V_G, (\varphi(w_1, \ldots, w_k) \Leftrightarrow \psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)}.$$

Case $\neg$: Suppose $\varphi(\alpha_1, \ldots, \alpha_k) = \neg\varphi'(\alpha_1, \ldots, \alpha_k)$. Let $\psi' : \mathsf{Types}(k) \times \mathbb{Z}_q^{\mathsf{Conn}_k^c} \to \{0,1\}$ be such that $\Pr_G[\forall w_1, \ldots, w_k \in V_G, (\varphi'(w_1, \ldots, w_k) \Leftrightarrow \psi'(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)}$. Setting $\psi(\tau, f) = 1 - \psi'(\tau, f)$, we see that

$$\Pr_G[\forall w_1, \ldots, w_k \in V_G, (\varphi(w_1, \ldots, w_k) \Leftrightarrow \psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^c(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)}.$$

Case Mod_q^i: Suppose $\varphi(\alpha_1, \ldots, \alpha_k) = \mathsf{Mod}_q^i\beta, \varphi'(\alpha_1, \ldots, \alpha_k, \beta)$. Let $c' = c(k+1, t-1, q)$ and let $\psi' : \mathsf{Types}(k+1) \times \mathbb{Z}_q^{\mathsf{Conn}_{k+1}^{c'}} \to \{0,1\}$ be given by the induction hypothesis, so that

$$\Pr_G[\forall w_1, \ldots, w_k, v \in V_G, (\varphi'(w_1, \ldots, w_k, v) \Leftrightarrow \psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G^{c'}(\mathbf{w}, v)) = 1)] \geq 1 - 2^{-\Omega(n)}.$$

Call G *good* if this event occurs, i.e., if

$$\forall w_1, \ldots, w_k, v \in V_G, (\varphi'(w_1, \ldots, w_k, v) \Leftrightarrow \psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G^{c'}(\mathbf{w}, v)) = 1).$$

Let $\gamma(w_1, \ldots, w_k)$ be the number (mod q) of v such that $\varphi'(w_1, \ldots, w_k, v)$ is true. Then for any good G (doing arithmetic mod q),

$$\gamma(w_1, \ldots, w_k) = \sum_{v \in V_G} \psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}^{c'}_G(\mathbf{w}, v)).$$

Grouping terms, we have

$$\begin{aligned}\gamma(w_1, \ldots, w_k) &= \sum_{\tau' \in \mathsf{Types}(k+1)} \sum_{f' \in \mathbb{Z}_q^{\mathsf{Conn}^{c'}_{k+1}}} \psi'(\tau', f') \cdot |\{v \in V_G : \mathsf{type}_G(\mathbf{w}, v) = \tau') \wedge \mathsf{freq}_G(\mathbf{w}, v) = f'\}| \\ &= \sum_{\tau', f'} \psi'(\tau', f') \cdot \lambda(\tau', f', \mathsf{type}_G(\mathbf{w}), \mathsf{freq}^c_G(\mathbf{w})) \\ &\qquad \text{(applying Theorem 6.1, and taking } c = (q-1)c'2^{(c')^2})\end{aligned}$$

which is solely a function of $\mathsf{type}_G(\mathbf{w})$ and $\mathsf{freq}^c_G(\mathbf{w})$. Thus, there is a function $\psi : \mathsf{Types}(k) \times \mathbb{Z}_q^{\mathsf{Conn}^c_k} \to \{0,1\}$ such that for all good G and for all $w_1, \ldots, w_k \in V_G$, $\psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}^c_G(\mathbf{w})) = 1$ if and only if $\gamma(\mathbf{w}) \equiv i \mod q$. Thus,

$$\Pr_G[\forall w_1, \ldots, w_k, ((\mathsf{Mod}^i_q v, \varphi'(w_1, \ldots, w_k, v)) \Leftrightarrow \psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}^c_G(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)},$$

as desired.

Case $\exists$: Suppose $\varphi(\alpha_1, \ldots, \alpha_k) = \exists \beta, \varphi'(\alpha_1, \ldots, \alpha_k, \beta)$. Let $c' = c(k+1, t-1, q)$ and let $\psi' : \mathsf{Types}(k+1) \times \mathbb{Z}_q^{\mathsf{Conn}^{c'}_{k+1}} \to \{0,1\}$ be such that

$$\Pr_G[\forall w_1, \ldots, w_k, v \in V_G, (\varphi'(w_1, \ldots, w_k, v) \Leftrightarrow \psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G(\mathbf{w}, v)) = 1)] \geq 1 - 2^{-\Omega(n)}. \tag{6}$$

For this case, we may choose c to be any integer at least c'. Define $\psi : \mathsf{Types}(k) \times \mathbb{Z}_q^{\mathsf{Conn}^c_k} \to \{0,1\}$ by the rule: $\psi(\tau, f) = 1$ if there is a $(\tau', f') \in \mathsf{Types}(k+1) \times \mathsf{FFreq}_n(\tau', [k+1], c')$ extending (τ, f) such that $\psi'(\tau', f') = 1$.

Fix any $\mathbf{w} \in [n]^k$. We will show that

$$\Pr_G[(\exists v, \psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}^{c'}_G(\mathbf{w}, v)) = 1) \Leftrightarrow \psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}^c_G(\mathbf{w})) = 1] \geq 1 - 2^{-\Omega(n)}. \tag{7}$$

Taking a union bound of (7) over all $\mathbf{w} \in [n]^k$, and using Equation (6), we conclude that

$$\Pr_{G \in G(n,p)}[\forall w_1, \ldots, w_k \in V_G, (\varphi(w_1, \ldots, w_k) \Leftrightarrow \psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G(\mathbf{w})) = 1)] \geq 1 - 2^{-\Omega(n)},$$

as desired.

It remains to show Equation (7). It will help to expose the edges of the random graph G in three stages.

In the first stage, we expose all the edges between the vertices in $\{w_1, \ldots, w_k\}$.

For the second stage, let $s = \rho(c, q, p) \cdot n$ (where ρ comes from Theorem 6.12) and pick distinct vertices $u_1, \ldots, u_s \in [n] \setminus \{w_1, \ldots, w_k\}$. In the second stage, we expose all the unexposed edges between the vertices in $\{w_1, \ldots, w_k, u_1, \ldots, u_s\}$ (i.e., the edges between u_is and w_js, as well as the edges between the u_is and u_js). Denote the resulting graph induced on $\{w_1, \ldots, w_k, u_1, \ldots, u_s\}$ after the second stage by A (so that $V_A = \{w_1, \ldots w_k, u_1, \ldots, u_s\}$).

In the third stage, we expose the rest of the edges in G. Thus G is sampled from $G(n, p \mid V_A, E_A)$.

Let τ denote the random variable $\mathsf{type}_G(\mathbf{w})$. Note that τ is determined after the first stage. Let $\tau_1, \ldots, \tau_s$ denote the random variables $\mathsf{type}_G(\mathbf{w}, u_1), \ldots, \mathsf{type}_G(\mathbf{w}, u_s)$. Note that $\tau_1, \ldots, \tau_s$ are all determined after the second stage. Let f denote the random variable $\mathsf{freq}_G(\mathbf{w})$. Let $f_1, \ldots, f_s$ denote the random variables $\mathsf{freq}_G(\mathbf{w}, u_1), \ldots, \mathsf{freq}_G(\mathbf{w}, u_s)$. The variables $f, f_1, \ldots, f_s$ are all determined after the third stage. Notice that the content of Theorem 6.12 is precisely a description of the distribution of $(f, f_1, \ldots, f_s)$.

We identify two bad events B_1 and B_2.

B_1 is defined to be the event: there exists $\sigma \in \mathsf{Types}(k+1)$ extending τ, with $\{k+1\} \in \Pi_\sigma$ (ie, types σ where vertex $k+1$ is distinct from the other vertices), such that

$$|\{i \in [s] : \tau_i = \sigma\}| \leq \frac{1}{2} s \min\{p^k, (1-p)^k\}.$$

(This can be interpreted as saying that the type σ appears abnormally infrequently amongst the τ_i). Note that for any σ extending τ, the events "$\tau_i = \sigma$", for $i \in [s]$, are independent conditioned on the outcome of the first stage, since they depend on disjoint sets of edges of G. Also, for each i and each σ extending τ with $\{k+1\} \in \Pi_\sigma$, the probability that $\tau_i = \sigma$ is $\geq \min\{p^k, (1-p)^k\}$. Therefore, applying the Chernoff bound, and taking a union bound over all σ extending τ with $\{k+1\} \in \Pi_\sigma$, we see that

$$\Pr[B_1] \leq 2^k \exp(-s \min\{p^k, (1-p)^k\}) \leq 2^{-\Omega(n)}.$$

Now let

$$\begin{aligned} S = \{(\sigma, g) \in \mathsf{Types}(k+1) \times \mathsf{FFreq}_n(\sigma, [k+1], c') \mid \{k+1\} \in \Pi_\sigma \\ \text{AND } (\sigma, g) \text{ extends } (\tau, f) \text{ AND } \psi'(\sigma, g) = 1\}. \end{aligned}$$

B_2 is defined to be the event: $S \neq \emptyset$ and for each $i \in [s]$, $(\tau_i, f_i) \notin S$. We study the probability of $\neg B_1 \wedge B_2$. Let U be the set of $(d, d_1, \ldots, d_s) \in \mathsf{FFreq}_n(\tau, [k], c) \times \prod_i \mathsf{FFreq}_n(\tau_i, [k+1], c')$ such that

1. The set $S(d)$ defined by

$$\begin{aligned} S(d) = \{(\sigma, g) \in \mathsf{Types}(k+1) \times \mathsf{FFreq}_n(\sigma, [k+1], c') \mid \{k+1\} \in \Pi_\sigma \\ \text{AND } (\sigma, g) \text{ extends } (\tau, d) \text{ AND } \psi'(\sigma, g) = 1\}, \end{aligned}$$

   is nonempty.

2. For each $i \in [s]$, $(\tau_i, d_i) \notin S(d)$.

By definition, the event B_2 occurs precisely when $(f, f_1, \ldots, f_s) \in U$.

By Theorem 6.12, for any fixing of E_A, the probability that $(f, f_1, \ldots, f_s) \in U$ is at most $2^{-\Omega(n)}$ more than the probability that $(h, h_1, \ldots, h_s) \in U$. As the event B_1 is solely a function of E_A, we conclude that $\Pr[\neg B_1 \wedge (f, f_1, \ldots, f_s) \in U)] \leq \Pr[\neg B_1 \wedge (h, h_1, \ldots, h_s) \in U] + 2^{-\Omega(n)}$.

It remains to bound $\Pr[\neg B_1 \wedge (h, h_1, \ldots, h_s) \in U]$. If $S(h) \neq \emptyset$, take a $(\sigma, g) \in S(h)$. In the absence of B_1, the number of $i \in [s]$ with $\tau_i = \sigma$ is at least $\frac{1}{2} s \min\{p^k, (1-p)^k\}$. For all these i, it must hold that $h_i \neq g$ in order for $(h, h_1, \ldots, h_s)$ to lie in U. Therefore,

$$\Pr[\neg B_1 \wedge (h, h_1, \ldots, h_s) \in U] \leq \left(1 - \frac{1}{|\mathsf{FFreq}_n(\tau, k+1, c')|}\right)^{\frac{1}{2} s \min\{p^k, (1-p)^k\}}.$$

Notice that this last quantity is of the form $2^{-\Omega_{p,q,k,d}(s)}$.

Putting everything together,

$$\Pr[\neg B_1 \wedge B_2] \leq \Pr[\neg B_1 \wedge (h, h_1, \ldots, h_s) \in U] + 2^{-\Omega(n)} \leq 2^{-\Omega(s)} + 2^{-\Omega(n)} \leq 2^{-\Omega(n)}.$$

Therefore, with probability at least $1 - 2^{-\Omega(n)}$, the event B_2 does not occur. The next claim finishes the proof of Equation (7), and with that the proof of Theorem 5.8.

Claim 6.13 *If B_2 does not occur, then*

$$(\exists v, \psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G^{c'}(\mathbf{w}, v)) = 1) \Leftrightarrow (\psi(\mathsf{type}_G(\mathbf{w}), \mathsf{freq}_G^{c}(\mathbf{w})) = 1).$$

Proof Let $\tau = \mathsf{type}_G(\mathbf{w})$ and $f = \mathsf{freq}_G^c(\mathbf{w})$.

If $\psi(\tau, f) = 0$, then we know that for all $(\tau', f') \in \mathsf{Types}(k+1) \times \mathsf{FFreq}_n(\tau', k+1, c')$ extending (τ, f), we have $\psi'(\tau', f') = 0$. Thus by Lemma 6.8, for all $v \in V_G$, $\psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G^{c'}(\mathbf{w}, v)) = 0$, as required.

If $\psi(\tau, f) = 1$, then we consider two situations.

- **The self-fulfilling situation:** If there is a $(\tau', f') \in \mathsf{Types}(k+1) \times \mathsf{FFreq}_n(\tau', k+1, c')$ extending (τ, f) with $\{k+1\} \notin \Pi_{\tau'}$ and $\psi'(\tau', f') = 1$. In this case, take any $j \in [k]$ with $\Pi_{\tau'}(j) = \Pi_{\tau'}(k+1)$, and let $v = w_j$. Thus $\mathsf{type}_G(\mathbf{w}, v) = \tau'$. By Lemma 6.10, since (τ', f') extends (τ, f) with $\{k+1\} \notin \Pi_{\tau'}$, it follows that $\mathsf{freq}_G^{c'}(\mathbf{w}, v) = f'$. Therefore, with this choice of v, we have $\psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G^{c'}(\mathbf{w}, v)) = 1$, as required.
- **The default situation:** In this case, there is a $(\tau', f') \in \mathsf{Types}(k+1) \times \mathsf{FFreq}_n(\tau', k+1, c')$ extending (τ, f) with $\{k+1\} \in \Pi_{\tau'}$ and $\psi'(\tau', f') = 1$. This is precisely the statement that $S \neq \emptyset$. Therefore, by the absence of the event B_2, there must be an $i \in [r]$ such that $(\tau_i, f_i) \in S$. Taking $v = u_i$, we see that $\psi'(\mathsf{type}_G(\mathbf{w}, v), \mathsf{freq}_G^{c'}(\mathbf{w}, v)) = 1$, as required.

This completes the proof of the claim. □ □

7 Counting Extensions

In this section we prove Theorem 6.1.

7.1 Subgraph frequency arithmetic

We begin with a definition. A *partial matching* between two I-labelled graphs F_1, F_2 is a subset $\eta \subseteq (V_{F_1} \setminus \mathcal{L}(F_1)) \times (V_{F_2} \setminus \mathcal{L}(F_2))$ that is one-to-one. For two graphs F_1, F_2, let $\mathsf{PMatch}(F_1, F_2)$ be the set of all partial matchings between them.

Definition 7.1 (Gluing along a partial matching) *Let F_1 and F_2 be two I-labelled graphs, and let $\eta \in \mathsf{PMatch}(F_1, F_2)$. Define* the gluing of F_1 and F_2 along η, *denoted $F_1 \vee_\eta F_2$, to be the graph obtained by first taking the disjoint union of F_1 and F_2, identifying pairs of vertices with the same label, and then identifying the vertices in each pair of η (and removing duplicate edges). We omit the subscript when $\eta = \emptyset$.*

We have the following simple identity.

Lemma 7.2 *For any I-labelled graphs F_1, F_2, any graph G and any $\mathbf{w} \in V_G^I$:*

$$[F_1](G, \mathbf{w}) \cdot [F_2](G, \mathbf{w}) = \sum_{\eta \in \mathsf{PMatch}(F_1, F_2)} [F_1 \vee_\eta F_2](G, \mathbf{w}). \tag{8}$$

Proof We give a bijection

$$\alpha : \mathsf{Inj}(F_1, (G, \mathbf{w})) \times \mathsf{Inj}(F_2, (G, \mathbf{w})) \to \coprod_{\eta \in \mathsf{PMatch}(F_1, F_2)} \mathsf{Inj}(F_1 \vee_\eta F_2, (G, \mathbf{w})).$$

Define $\alpha(\chi_1, \chi_2)$ as follows. Let $\eta = \{(v_1, v_2) \in (V_{F_1} \setminus \mathcal{L}(F_1)) \times (V_{F_2} \setminus \mathcal{L}(F_2)) \mid \chi_1(v_1) = \chi_2(v_2)\}$. Let $\iota_1 \in \mathsf{Inj}(F_1, F_1 \vee_\eta F_2)$ and $\iota_2 \in \mathsf{Inj}(F_2, F_1 \vee_\eta F_2)$ be the natural inclusions. Let $\chi \in \mathsf{Inj}(F_1 \vee_\eta F_2, (G, \mathbf{w}))$ be the unique homomorphism such that for all $v \in V_{F_1}$, $\chi \circ \iota_1(v) = \chi_1(v)$, and for all $v \in V_{F_2}$, $\chi \circ \iota_2(v) = \chi_2(v)$. We define $\alpha(\chi_1, \chi_2) := \chi$.

To see that α is a bijection, we give its inverse β. Let $\eta \in \mathsf{PMatch}(F_1, F_2)$ and $\chi \in \mathsf{Inj}(F_1 \vee_\eta F_2, (G, \mathbf{w}))$. Let $\iota_1 \in \mathsf{Inj}(F_1, F_1 \vee_\eta F_2)$ and $\iota_2 \in \mathsf{Inj}(F_2, F_1 \vee_\eta F_2)$ be the natural inclusions. Define $\beta(\chi) := (\chi \circ \iota_1, \chi \circ \iota_2)$.

Then β is the inverse of α. □

We can now prove Lemma 6.2.

Lemma 6.2 (Label-connected subgraph frequencies determine all subgraph frequencies, restated) *For every k-labelled graph F' with $|V_{F'} \setminus \mathcal{L}(F')| \leq t$, there is a polynomial $\delta_{F'} \in \mathbb{Z}[(X_F)_{F \in \mathsf{Conn}_k^t}]$ such that for all graphs G and $\mathbf{w} \in V_G^k$,*

$$[F'](G, \mathbf{w}) = \delta_{F'}(x),$$

where $x \in \mathbb{Z}^{\mathsf{Conn}^t_k}$ *is given by* $x_F = [F](G, \mathbf{w})$.

Proof By induction on the number of connected components of $F' \setminus \mathcal{L}(F')$. If F' is label-connected , then we take $\delta_{F'}(\mathbf{X}) = X_{F'}$.

Now suppose F' is label-disconnected. Write $F' = F_1 \vee F_2$ where F_1 and F_2 are both k-labelled graphs, and $F_1 \setminus \mathcal{L}(F_1)$ and $F_2 \setminus \mathcal{L}(F_2)$ have fewer connected components.

By equation (8), for all G and $\mathbf{w}$,

$$[F_1 \vee F_2](G, \mathbf{w}) = [F_1](G, \mathbf{w}) \cdot [F_2](G, \mathbf{w}) - \sum_{\emptyset \neq \eta \in \mathsf{PMatch}(F_1, F_2)} [F_1 \vee_\eta F_2](G, \mathbf{w}).$$

Observe that for any $\eta \neq \emptyset$, each graph $F_1 \vee_\eta F_2$ has at least one fewer label-connected component than $F_1 \vee F_2 = F'$. Thus, by induction hypothesis, we may take

$$\delta_{F'}(\mathbf{X}) = \delta_{F_1}(\mathbf{X}) \cdot \delta_{F_2}(\mathbf{X}) - \sum_{\emptyset \neq \eta \in \mathsf{PMatch}(F_1, F_2)} \delta_{F_1 \vee_\eta F_2}(\mathbf{X}).$$

This completes the proof of the lemma. □

7.2 Proof of Theorem 6.1

Theorem 6.1 (restated) *Let* q *be a prime, let* $k, b > 0$ *be integers and let* $a \geq (q-1) \cdot b \cdot |\mathsf{Conn}^b_{k+1}|$. *There is a function*

$$\lambda : \mathsf{Types}(k+1) \times \mathbb{Z}_q^{\mathsf{Conn}^b_{k+1}} \times \mathsf{Types}(k) \times \mathbb{Z}_q^{\mathsf{Conn}^a_k} \to \mathbb{Z}_q$$

such that for all $\tau' \in \mathsf{Types}(k+1)$, $f' \in \mathbb{Z}_q^{\mathsf{Conn}^b_{k+1}}$, $\tau \in \mathsf{Types}(k)$, $f \in \mathbb{Z}_q^{\mathsf{Conn}^a_k}$, *it holds that for every graph* G, *and every* $w_1, \ldots, w_k \in V_G$ *with* $\mathsf{type}_G(\mathbf{w}) = \tau$ *and* $\mathsf{freq}^a_G(\mathbf{w}) = f$, *the cardinality of the set*

$$\{v \in V_G : \mathsf{type}_G(\mathbf{w}, v) = \tau' \wedge \mathsf{freq}^b_G(\mathbf{w}, v) = f'\}$$

is congruent to $\lambda(\tau', f', \tau, f) \mod q$.

Proof We describe the function $\lambda(\tau', f', \tau, f)$ explicitly. If τ' does not extend τ, then we set $\lambda(\tau', f', \tau, f) = 0$.

Now assume τ' extends τ. We take cases on whether $k+1$ is a singleton in $\Pi_{\tau'}$ or not.

Case 1: $\{k+1\} \in \Pi_{\tau'}$. In this case, there is an $I \subseteq [k]$ such that $\mathsf{type}_G(w_1, \ldots, w_k, v) = \tau'$ if and only if $v \notin \{w_1, \ldots, w_k\}$ and $(v, w_i) \in E_G \Leftrightarrow i \in I$ (explicitly, $I = \{i \in [k] \mid \{\{k+1\}, \Pi_{\tau'}(i)\} \in E_{\tau'}\}$).

Let for each $u, v \in V_G$, let $x_{uv} \in \{0, 1\}$, where $x_{uv} = 1$ if and only if u is adjacent to v in G.

Then, using the fact that q is prime, the number (mod q) of v with $\mathsf{type}_G(\mathbf{w}, v) = \tau'$ and $\mathsf{freq}_G(\mathbf{w}, v) = f'$ can be compactly expressed as (doing arithmetic mod q):

$$\sum_{v \in V_G \setminus \{w_1, \ldots, w_k\}} \prod_{i \in I} x_{vw_i} \prod_{j \in [k] \setminus I} (1 - x_{vw_j}) \prod_{F \in \mathsf{Conn}^b_{k+1}} \left(1 - \left([F]_q(G, \mathbf{w}, v) - f'_F\right)^{q-1}\right)$$

Expanding, the expression $\prod_{i\in I} x_{vw_i} \prod_{j\in[k]\setminus I}(1-x_{vw_j})$ may be expressed in the form $\sum_{S\subseteq[k]} b_S \prod_{i\in S} x_{vw_i}$. Using Lemma 7.2, the expression $\prod_{F\in\mathsf{Conn}^b_{k+1}} \left(1-([F]_q(G,\mathbf{w},v)-f'_F)^{q-1}\right)$ may be expressed in the form $\sum_j c_j[F_j]_q(G,\mathbf{w},v)$, where each F_j is a $k+1$-labelled graph with at most $|\mathsf{Conn}^b_{k+1}|\cdot b\cdot(q-1)\le a$ vertices.

Thus we may rewrite the expression for $\lambda(\tau',f',\tau,f)$ as:

$$\begin{aligned}
&\sum_{v\in[n]\setminus\{w_1,\ldots,w_k\}} \left(\sum_S b_S \prod_{i\in S} x_{vw_i}\right)\left(\sum_j c_j[F_j]_q(G,\mathbf{w},v)\right)\\
&=\sum_{S,j} b_S c_j \sum_{v\in[n]\setminus\{w_1,\ldots,w_k\}} \left(\left(\prod_{i\in S} x_{vw_i}\right)[F_j]_q(G,\mathbf{w},v)\right)\\
&=\sum_{S,j} b_S c_j [F'_{S,j}]_q(G,\mathbf{w}),
\end{aligned}$$

where $F'_{S,j}$ is the k-labelled graph obtained from F_j by

(a) For each $i\in S$, adding an edge between the vertex labelled $k+1$ and the vertex labelled i, and

(b) Removing the label from the vertex labelled $k+1$.

Finally, note that by Lemma 6.2, $[F'_{S,j}]_q(G,\mathbf{w})$ is determined by $\mathsf{freq}^a_G(\mathbf{w})$.

Case 2: $\{k+1\}\notin\Pi_{\tau'}$. This case is much easier to handle. Pick any $j\in[k]$ such that $\Pi_{\tau'}(j)=\Pi_{\tau'}(k+1)$. Then there is only one $v\in V_G$ such that $\mathsf{type}_G(\mathbf{w},v)=\tau'$ (namely, w_j).

Then $\lambda(\tau',f',\tau,f)=1$ if and only if for all $F'\in\mathsf{Conn}^b_{k+1}$, $f'_{F'}=f_F$, where $F\in\mathsf{Conn}^b_k$ is the graph obtained by identifying the vertex labelled $k+1$ with the vertex labelled j, and labelling this new vertex j. Otherwise $\lambda(\tau',f',\tau,f)=0$.

This completes the definition of our desired function λ. □

8 The Distribution of Labelled Subgraph Frequencies mod q

In this section, we prove Theorem 6.12. As in Section 3, the proof will be via an intermediate theorem (Theorem 8.2) that proves the equidistribution of the number of *copies* of labelled subgraphs in $G(n,p)$.

8.1 Equidistribution of labelled subgraph copies

First, we gather some simple observations about injective homomorphisms from label-connected graphs for later use (the proofs are simple and are omitted).

Proposition 8.1 (Simple but delicate observations about label-connected graphs) *Let $F, F' \in \mathsf{Conn}_I^t$. Let G be a graph and let $\mathbf{w} \in V_G^I$ with all $(w_i)_{i\in I}$ distinct.*

1. *If $E \in \mathsf{Cop}(F,(G,\mathbf{w}))$, the $|E| = |E_F|$.*
2. *If $F \ncong F'$, we have $\mathsf{Cop}(F,(G,\mathbf{w})) \cap \mathsf{Cop}(F',(G,\mathbf{w})) = \emptyset$.*
3. *Let $\chi_1,\dots,\chi_r \in \mathsf{Inj}(F,(G,\mathbf{w}))$ be such that for any distinct $j,j' \in [r]$, $\chi_j(V_F \setminus \mathcal{L}(F)) \cap \chi_{j'}(V_F \setminus \mathcal{L}(F)) = \emptyset$. Let $\chi \in \mathsf{Inj}(F',(G,\mathbf{w}))$. Suppose $\chi(E_{F'}) \subseteq (\cup_j \chi_j(E_F))$. Then there is a $j \in [r]$ such that $\chi(E_{F'}) \subseteq \chi_j(E_F)$.*

We can now state and prove an equidistribution theorem for the number of copies of labelled subgraphs in a conditioned random graph. Theorem 6.12 will follow from this.

Theorem 8.2 *Let A be a graph with $V_A \subseteq [n]$ and $|V_A| \le n'$. Let $\mathbf{w} = (w_1,\dots,w_k) \in V_A^k$ with $w_1,\dots,w_k$ distinct. Let $u_1,\dots u_s \in V_A \setminus \{w_i : i \in I\}$ be distinct. Let $F_1,\dots,F_\ell$ be distinct k-labelled label-connected graphs, with $1 \le |E_{F_i}| \le d$. Let $H_1,\dots,H_{\ell'}$ be distinct $k+1$-labelled label-connected graphs dependent on label $k+1$, with $1 \le |E_{H_i}| \le d$.*

Let $G \in G(n,p \mid V_A, E_A)$. Then the distribution of

$$((\langle F_i\rangle_q(G,\mathbf{w}))_{i\in[\ell]}, (\langle H_{i'}\rangle_q(G,\mathbf{w},u_{j'}))_{i'\in[\ell'],j'\in[s]})$$

on $\mathbb{Z}_q^{\ell+s\ell'}$ is $2^{-\Omega_{q,p,d}(n-n')+(\ell+\ell' s)\log q}$-close to uniform in statistical distance.

Proof By the Vazirani XOR lemma (Lemma 3.4), it suffices to show that for any nonzero $(c,c') \in \mathbb{Z}_q^\ell \times \mathbb{Z}_q^{\ell'\times s}$, we have $\left|\mathbb{E}\left[\omega^R\right]\right| \le 2^{-\Omega_{q,p,d}(n-n')}$, where

$$R := \sum_{i\in[\ell]} c_i \langle F_i\rangle_q(G,\mathbf{w}) + \sum_{i'\in[\ell']}\sum_{j'\in[s]} c'_{i'j'}\langle H_{i'}\rangle_q(G,\mathbf{w},u_{j'})$$

and $\omega \in \mathbb{C}$ is a primitive q^{th}-root of unity.

We will show this by appealing to Lemma 3.3. Let $m = \binom{n}{2} - \binom{a}{2}$. Let $\mathbf{z} \in \{0,1\}^{\binom{[n]}{2}}$ be the random variable where, for each $e \in \binom{[n]}{2}$, $z_e = 1$ if and only if edge e is present in G. Thus, independently for each $e \in \binom{[n]}{2} \setminus \binom{V_A}{2}$, $\Pr[z_e = 1] = p$, while for $e \in \binom{V_A}{2}$, the value of z_e is either identically 1 or identically 0 (depending on whether $e \in E_A$ or not).

We may now express R in terms of the z_e. We have,

$$\begin{aligned}
R &= \sum_{i\in[\ell]} c_i\langle F_i\rangle_q(G,\mathbf{w}) + \sum_{i'\in[\ell']}\sum_{j'\in[s]} c'_{i'j'}\langle H_{i'}\rangle_q(G,\mathbf{w},u_{j'}) \\
&= \sum_{i\in[\ell]} c_i \sum_{E\in\mathsf{Cop}(F_i,(\mathrm{K}_n,\mathbf{w}))}\prod_{e\in E} z_e + \sum_{i'\in[\ell']}\sum_{j'\in[s]} c'_{i'j'}\sum_{E\in\mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'}))}\prod_{e\in E} z_e \\
&= \sum_{E\in\mathcal{F}_1} c_E \prod_{e\in E} z_e + \sum_{E\in\mathcal{F}_2} c'_E \prod_{e\in E} z_e,
\end{aligned}$$

where $\mathcal{F}_1 \subseteq 2^{\binom{[n]}{2}}$ is the set $\bigcup_{i\in[\ell]:c_i\neq 0} \mathsf{Cop}(F_i,(\mathrm{K}_n,\mathbf{w}))$, $\mathcal{F}_2$ is the set $\bigcup_{i'\in[\ell'],j'\in[s]:c'_{i'j'}\neq 0} \mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'}))$, for each $E\in\mathcal{F}_1$, $c_E = c_i$ where $i\in[\ell]$ is such that $E\in\mathsf{Cop}(F_i,(\mathrm{K}_n,\mathbf{w}))$ (note that by Proposition 8.1 there is exactly one such i), and similarly, for $E\in\mathcal{F}_2$, $c'_E = \sum_{i'\in[\ell'],j'\in[s]:E\in\mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'})} c'_{i'j'}$. Thus if E is such that there is a unique $(i',j')\in[\ell']\times[s]$ for which $E\in\mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'}))$ and $c'_{i',j'}\neq 0$, then $c'_E\neq 0$.

Let $Q(\mathbf{Z})\in\mathbb{Z}_q[\mathbf{Z}]$, where $\mathbf{Z}=(Z_e)_{e\in\binom{[n]}{2}\setminus\binom{V_A}{2}}$, be the polynomial

$$\sum_{E\in\mathcal{F}_1} c_E \prod_{e\in E\cap\binom{V_A}{2}} z_e \prod_{e\in E\setminus\binom{V_A}{2}} Z_e + \sum_{E\in\mathcal{F}_2} c'_E \prod_{e\in E\cap\binom{V_A}{2}} z_e \prod_{e\in E\setminus\binom{V_A}{2}} Z_e.$$

Let $\widehat{\mathbf{z}}\in\{0,1\}^{\binom{[n]}{2}\setminus\binom{V_A}{2}}$ be the random variable $\mathbf{z}$ restricted to the coordinates indexed by $\binom{[n]}{2}\setminus\binom{V_A}{2}$ (thus each coordinate of $\widehat{\mathbf{z}}$ independently equals 1 with probability p). Then $R=Q(\widehat{\mathbf{z}})$. We wish to show that

$$\left|\mathbb{E}\left[\omega^{Q(\widehat{\mathbf{z}})}\right]\right| \leq 2^{-\Omega_{q,p,d}(n-n')}. \tag{9}$$

We do this by demonstrating that the polynomial $Q(\mathbf{Z})$ satisfies the hypotheses of Lemma 3.3.

Let $d_1^* = \max_{i:c_i\neq 0}|E_{F_i}|$. Let $d_2^* = \max_{i',j':c'_{i'j'}\neq 0}|E_{H_{i'}}|$. We take cases depending on whether $d_1^* < d_2^*$ or $d_1^*\geq d_2^*$.

Case 1: Suppose $d_1^* < d_2^*$. Let i'_0,j'_0 be such that $c'_{i'_0j'_0}\neq 0$ and $|E_{H_{i'_0}}| = d_2^*$. Then $Q(\mathbf{Z})$ may be written as $\sum_{E\in\mathcal{F}} c'_E\prod_{e\in E} Z_e + Q'(\mathbf{Z})$, where $\mathcal{F}=\{E\in\mathcal{F}_2 : E\cap\binom{V_A}{2}=\emptyset\}$ and $\deg(Q')<d_2^*$.

Let $\chi_1,\chi_2,\ldots,\chi_r\in\mathsf{Inj}(H_{i'_0},(\mathrm{K}_n,\mathbf{w},u_{j'_0}))$ be a collection of homomorphisms such that:

1. For all $j\in[r]$, we have $\chi_j(V_{H_{i'_0}}\setminus\mathcal{L}(H_{i'_0}))\subseteq[n]\setminus V_A$.
2. For all distinct $j,j'\in[r]$, we have $\chi_j(V_{H_{i'_0}}\setminus\mathcal{L}(H_{i'_0}))\cap\chi_{j'}(V_{H_{i'_0}}\setminus\mathcal{L}(H_{i'_0}))=\emptyset$.

Such a collection can be chosen greedily so that $r=\Omega(\frac{n-n'}{d})$. Let $E_j\in\mathsf{Cop}(H_{i'_0},(\mathrm{K}_n,\mathbf{w},u_{j'_0}))$ be given by $\chi_j(E_{H_{i'_0}})$. Let $\mathcal{E}$ be the family of sets $\{E_1,\ldots,E_r\}\subseteq\mathcal{F}$. We observe the following properties of the E_j:

1. For each $j\in[r]$, $|E_j|=d_2^*$ (since χ_j is injective and $w_1,\ldots,w_k,u_{j'_0}$ are distinct).
2. For each $j\in[r]$, $c'_{E_j}\neq 0$. This is because there is a unique (i',j') (namely (i'_0,j'_0)) for which $c_{i'j'}\neq 0$ and $E_j\in\mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'}))$. Indeed, if $j'\neq j'_0$, then each $E^*\in\mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'}))$ has some element incident on $u_{j'}$ (while E_j does not). On the other hand, if $j'=j'_0$ and $i'\neq i'_0$, then Proposition 8.1 implies that $\mathsf{Cop}(H_{i'},(\mathrm{K}_n,\mathbf{w},u_{j'}))\cap\mathsf{Cop}(H_{i'_0},(\mathrm{K}_n,\mathbf{w},u_{j'_0}))=\emptyset$.
3. For distinct $j,j'\in[r]$, $E_j\cap E_{j'}=\emptyset$ (by choice of the χ_j).

4. For any $S \in \mathcal{F} \setminus \mathcal{E}$, $|S \cap (\cup_j E_j)| < d_2^*$. To see this, take any $S \in \mathcal{F} \setminus \mathcal{E}$ and suppose $|S \cap (\cup_j E_j)| \geq d_2^*$. Let $i' \in [\ell'], j' \in [s]$ be such that $S \in \mathsf{Cop}(H_{i'}, (\mathrm{K}_n, \mathbf{w}, u_{j'}))$. Let $\chi \in \mathsf{Inj}(H_{i'}, (\mathrm{K}_n, \mathbf{w}, u_{j'}))$ with $\chi(E_{H_{i'}}) = S$. By choice of d_2^*, we know that $|S| \leq d_2^*$. Therefore, the only way that $|S \cap (\cup_j E_j)|$ can be $\geq d_2^*$ is if (a) $|S| = d_2^*$, and (b) $S \cap (\cup_j E_j) = S$, or in other words, $S \subseteq (\cup_j E_j)$. Since $H_{i'}$ is dependent on label $k+1$, we know that S has some element incident on vertex $u_{j'}$, and thus (b) forces $j' = j'_0$ (otherwise no E_j is incident on $u_{j'}$). Now by Proposition 8.1, this implies that $S \subseteq E_j$ for some j. But since $|E_j| = |S|$, we have $S = E_j$, contradicting our choice of S. Therefore, $|S \cap (\cup_j E_j)| < d_2^*$ for any $S \in \mathcal{F} \setminus \mathcal{E}$.

It now follows that $Q(\mathbf{Z}), \mathcal{F}$ and $\mathcal{E}$ satisfy the hypothesis of Lemma 3.3. Consequently, (noting that $d_2^* \leq d$) Equation (9) follows, completing the proof in Case 1.

Case 2: Suppose $d_1^* \geq d_2^*$. Let i_0 be such that $c_{i_0} \neq 0$ and $|E_{F_{i_0}}| = d_1^*$. Then $Q(\mathbf{Z})$ may be written as $\sum_{E \in \mathcal{F}} (c_E + c'_E) \prod_{e \in E} Z_e + Q'(\mathbf{Z})$, where $\mathcal{F} = \{E \in \mathcal{F}_1 \cup F_2 : E \cap \binom{V_A}{2} = \emptyset\}$ and $\deg(Q') < d_1^*$.

Let $\chi_1, \chi_2, \ldots, \chi_r \in \mathsf{Inj}(F_{i_0}, (\mathrm{K}_n, \mathbf{w}))$ be a collection of homomorphisms such that:

1. For all $j \in [r]$, we have $\chi_j(V_{F_{i_0}} \setminus \mathcal{L}(F_{i_0})) \subseteq [n] \setminus V_A$.
2. For all distinct $j, j' \in [r]$, we have $\chi_j(V_{F_{i_0}} \setminus \mathcal{L}(F_{i_0})) \cap \chi_{j'}(V_{F_{i_0}} \setminus \mathcal{L}(F_{i_0})) = \emptyset$.

Such a collection can be chosen greedily so that $r = \Omega(\frac{n-n'}{d})$. Let $E_j \in \mathsf{Cop}(F_{i_0}, (\mathrm{K}_n, \mathbf{w}))$ be given by $\chi_j(E_{F_{i_0}})$. Let $\mathcal{E}$ be the family of sets $\{E_1, \ldots, E_r\} \subseteq \mathcal{F}$. We observe the following properties of the E_j:

1. For each $j \in [r]$, $|E_j| = d_1^*$ (since χ_j is injective and $w_1, \ldots, w_k$ are distinct).
2. For each $j \in [r]$, $c_{E_j} + c'_{E_j} \neq 0$. This is because $c_{E_j} = c_{i_0} \neq 0$ and for any (i', j'), $E_j \notin \mathsf{Cop}(H_{i'}, (\mathrm{K}_n, \mathbf{w}, u_{j'}))$ (and so $c'_{E_j} = 0$). To see the latter claim, note that each $E^* \in \mathsf{Cop}(H_{i'}, (\mathrm{K}_n, \mathbf{w}, u_{j'}))$ has an element incident on $u_{j'}$ (which E_j does not).
3. For distinct $j, j' \in [r]$, $E_j \cap E_{j'} = \emptyset$ (by choice of the χ_j).
4. For any $S \in \mathcal{F} \setminus \mathcal{E}$, $|S \cap (\cup_j E_j)| < d_1^*$. To see this, take any $S \in \mathcal{F} \setminus \mathcal{E}$ and suppose $|S \cap (\cup_j E_j)| \geq d_1^*$.
   (a) If $S \in \mathcal{F}_1$, then let $i \in [\ell]$ be such that $S \in \mathsf{Cop}(F_i, (\mathrm{K}_n, \mathbf{w}))$. Let $\chi \in \mathsf{Inj}(F_i, (\mathrm{K}_n, \mathbf{w}))$ with $\chi(E_{F_i}) = S$. We know that $|S| \leq d_1^*$. Therefore, the only way that $|S \cap (\cup_j E_j)|$ can be $\geq d_1^*$ is if (1) $|S| = d_1^*$, and (2) $S \cap (\cup_j E_j) = S$, or in other words, $S \subseteq (\cup_j E_j)$. However, by Proposition 8.1, this implies that $S \subseteq E_j$ for some j. But since $|E_j| = |S|$, we have $S = E_j$, contradicting our choice of S.
   (b) If $S \in \mathcal{F}_2$, then let $i' \in [\ell'], j' \in [s]$ be such that $S \in \mathsf{Cop}(H_{i'}, (\mathrm{K}_n, \mathbf{w}, u_{j'}))$. Let $\chi \in \mathsf{Inj}(H_{i'}, (\mathrm{K}_n, \mathbf{w}, u_{j'}))$ with $\chi(E_{H_{i'}}) = S$. We know that $|S| \leq d_2^* \leq d_1^*$. Now S has an element incident on $u_{j'}$. On the other hand none of the E_j have any edges incident on $u_{j'}$. Therefore $|S \cap (\cup_j E_j)| < |S| \leq d_1^*$.

Therefore, $|S \cap (\cup_j E_j)| < d_1^*$ for any $S \in \mathcal{F} \setminus \mathcal{E}$.

It now follows that $Q(\mathbf{Z}), \mathcal{F}$ and $\mathcal{E}$ satisfy the hypothesis of Lemma 3.3. Consequently, (noting that $d_1^* \leq d$) Equation (9) follows, completing the proof in Case 2. □

8.2 Proof of Theorem 6.12

Theorem 6.12 (restated) *Let $a \geq b$ be positive integers. Let A be a graph with $V_A \subseteq [n]$ and $|V_A| \leq n' \leq n/2$. Let $G \in G(n, p \mid V_A, E_A)$. Let $\mathbf{w} = (w_1, \ldots, w_k) \in V_A^k$, and let $u_1, \ldots, u_s \in V_A \setminus \{w_1, \ldots, w_k\}$ be distinct. Let $\tau = \mathsf{type}_G(\mathbf{w})$ and let $\tau_i = \mathsf{type}_G(\mathbf{w}, u_i)$ (note that $\tau, \tau_1, \ldots, \tau_s$ are already determined by E_A). Let f denote the random variable $\mathsf{freq}_G^a(\mathbf{w})$. Let f_i denote the random variable $\mathsf{freq}_G^b(\mathbf{w}, u_i)$.*

Then, there exists a constant $\rho = \rho(a, q, p) > 0$, such that if $s \leq \rho \cdot n$, then the distribution of $(f, f_1, \ldots, f_s)$ over $\mathsf{FFreq}_n(\tau, [k], a) \times \prod_i \mathsf{FFreq}_n(\tau_i, [k+1], b)$ is $2^{-\Omega(n)}$-close to the distribution of $(h, h_1, \ldots, h_s)$ generated as follows:

1. *h is picked uniformly at random from $\mathsf{FFreq}_n(\tau, [k], a)$.*
2. *For each i, each h_i is picked independently and uniformly from the set of all $f' \in \mathsf{FFreq}_n(\tau_i, [k+1], b)$ such that (τ_i, f') extends (τ, h).*

Proof Let $\mathbf{v} = \mathbf{w}/\Pi_\tau$. Let $F_1, \ldots, F_\ell$ be an enumeration of the elements of $\mathsf{Conn}_{\Pi_\tau}^a$.

Let $\Pi' \in \mathsf{Partitions}([k+1])$ equal $\Pi_\tau \cup \{\{k+1\}\}$. Notice that for each $i \in [s]$, $\Pi_{\tau_i} = \Pi'$. Let $H_1, \ldots, H_{\ell'}$ be an enumeration of those elements of $\mathsf{Conn}_{\Pi'}^b$ that are dependent on label i^*.

By Theorem 8.2 and the hypothesis on s for a suitable constant ρ, the distribution of

$$(g, g^1, \ldots, g^s) = ((\langle F_i \rangle_q(G, \mathbf{v}))_{i \in [\ell]}, (\langle H_{i'} \rangle(G, \mathbf{v}, u_{j'}))_{i' \in [\ell'], j' \in [s]})$$

is $2^{-\Omega(n)}$ close to uniform over $\mathbb{Z}_q^{\ell + \ell' s}$. Given the vector $(g, g^1, \ldots, g^s)$, we may compute the vector $(f, f_1, \ldots, f_s)$ as follows:

1. For $F = \mathrm{K}_1([k])$, we have $f_F = n - |\Pi_\tau|$.
2. For all other $F \in \mathsf{Conn}_k^a$, let $i \in [\ell]$ be such that $F/\Pi_\tau \cong F_i$. Then $f_F = g_i \cdot \mathsf{aut}(F_i)$.
3. For $H \in \mathsf{Conn}_{k+1}^b$ dependent on label $k+1$, let $i' \in [\ell']$ be such that $H/(\Pi') \cong H_i'$. Then for each $j' \in [s]$, $(f_{j'})_H = g_{i'}^{j'} \cdot \mathsf{aut}(H_{i'})$.
4. For $H \in \mathsf{Conn}_{k+1}^b$ not dependent on label $k+1$ and for any $j' \in [s]$, there is a *unique* setting of $(f_{j'})_H$ (given the settings above) that is consistent with the fact that (τ_j, f_j) extends (τ, f). This follows from Lemma 6.9.

This implies the desired claim about the distribution of $(f, f_1, \ldots, f_s)$. □

9 Concluding Remarks

The results presented here constitute the first systematic investigation of the asymptotic probabilities of properties expressible in first-order logic with counting quantifiers. Moreover, these results have been established by combining, for the first time, algebraic methods related to multivariate polynomials over finite fields with the method of quantifier elimination from mathematical logic.

We conclude with two open problems:

1. What is the complexity of computing the numbers $a_0, \ldots, a_{q-1}$ in Theorem 2.1? We know that it is PSPACE-hard to compute these numbers (it is already PSPACE-hard to tell if the asymptotic probability of a FO sentence is 0 or 1). Our proof shows that they may be computed in time $2^{2^{2^{\cdots}}}$ of height proportional to the quantifier depth of the formula. It is likely that a more careful analysis of our approximation of $\mathsf{FO}[\mathsf{Mod}_q]$ by polynomials can yield better upper bounds.

2. Is there a modular convergence law for $\mathsf{FO}[\mathsf{Mod}_m]$ for arbitrary m? The same obstacles that prevent the Razborov-Smolensky approach from generalizing to $\mathsf{AC0}[\mathsf{Mod}_6]$ impede us. Perhaps an answer to the above question will give some hints for $\mathsf{AC0}[\mathsf{Mod}_6]$?

Acknowledgements

Swastik Kopparty is very grateful to Eli Ben-Sasson, Danny Gutfreund and Alex Samorodnitsky for encouragement and stimulating discussions. We would also like to thank Miki Ajtai, Ron Fagin, Prasad Raghavendra, Ben Rossman, Shubhangi Saraf and Madhu Sudan for valuable discussions.

References

[BEH81] Andreas Blass, Geoffrey Exoo, and Frank Harary. Paley graphs satisfy all first-order adjacency axioms. *J. Graph Theory*, 5(4):435–439, 1981.

[BGK85] A. Blass, Y. Gurevich, and D. Kozen. A zero–one law for logic with a fixed point operator. *Information and Control*, 67:70–90, 1985.

[BNS89] L. Babai, N. Nisan, and M. Szegedy. Multiparty protocols and logspace-hard pseudorandom sequences. In *STOC: ACM Symposium on Theory of Computing (STOC)*, 1989.

[BR05] A. Blass and B. Rossman. Explicit graphs with extension properties. *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS*, (86):166–175, 2005.

[BV07] A. Bogdanov and E. Viola. Pseudorandom bits for polynomials. In *FOCS*, pages 41–51, 2007.

[Fag74] R. Fagin. Generalized first–order spectra and polynomial–time recognizable sets. In R. M. Karp, editor, *Complexity of Computation, SIAM-AMS Proceedings, Vol. 7*, pages 43–73, 1974.

[Fag76] R. Fagin. Probabilities on finite models. *Journal of Symbolic Logic*, 41:50–58, 1976.

[GKLT69] Y. V. Glebskii, D. I. Kogan, M. I. Liogonki, and V. A. Talanov. Range and degree of realizability of formulas in the restricted predicate calculus. *Cybernetics*, 5:142–154, 1969.

[Gow01] W. T. Gowers. A new proof of Szemerédi's theorem. *Geom. Funct. Anal.*, 11(3):465–588, 2001.

[GS71] R. L. Graham and J. H. Spencer. A constructive solution to a tournament problem. *Canad. Math. Bull.*, 14:45–48, 1971.

[GT08] B. Green and T. Tao. The primes contain arbitrarily long arithmetic progressions. *Ann. of Math. (2)*, 167(2):481–547, 2008.

[HKL96] L. Hella, Ph.G. Kolaitis, and K. Luosto. Almost everywhere equivalence of logics in finite model theory. *Bulletin of Symbolic Logic*, 2(4):422–443, 1996.

[KV87] Ph. G. Kolaitis and M. Y. Vardi. The decision problem for the probabilities of higher-order properties. In *Proc. 19th ACM Symp. on Theory of Computing*, pages 425–435, 1987.

[KV90] Ph. G. Kolaitis and M. Y. Vardi. 0-1 laws and decision problems for fragments of second-order logic. *Information and Computation*, 87:302–338, 1990.

[Lov08] S. Lovett. Unconditional pseudorandom generators for low degree polynomials. In *STOC*, pages 557–562, 2008.

[NNT05] M. Naor, A. Nussboim, and E. Tromer. Efficiently constructible huge graphs that preserve first order properties of random graphs. In *TCC*, pages 66–85, 2005.

[PS89] L. Pacholski and W. Szwast. The 0-1 law fails for the class of existential second-order Gödel sentences with equality. In *Proc. 30th IEEE Symp. on Foundations of Computer Science*, pages 280–285, 1989.

[Raz87] A. Razborov. Lower bounds on the size of bounded depth circuits over a complete basis with logical addition. *MATHNASUSSR: Mathematical Notes of the Academy of Sciences of the USSR*, 41, 1987.

[Smo87] R. Smolensky. Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *STOC*, pages 77–82, 1987.

[SS87] J. Spencer and S. Shelah. Threshold spectra for random graphs. In *Proc. 19th ACM Symp. on Theory of Computing*, pages 421–424, 1987.

[SS88] S. Shelah and J. Spencer. Zero-one laws for sparse random graphs. *J. Amer. Math. Soc.*, 1:97–115, 1988.

[Vio08] E. Viola. The sum of d small-bias generators fools polynomials of degree d. In *IEEE Conference on Computational Complexity*, pages 124–127, 2008.

[VW07] E. Viola and A. Wigderson. Norms, xor lemmas, and lower bounds for gf(2) polynomials and multiparty protocols. In *22th IEEE Conference on Computational Complexity (CCC)*, 2007.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/