Modular statistics for subgraph counts in sparse random graphs

Bobby DeMarco*
rvdemarco@gmail.com

Jeff Kahn†
Department of Mathematics
Rutgers University
Piscataway, NJ, U.S.A.
jkahn@math.rutgers.edu

Amanda Redlich‡
Department of Mathematics
Bowdoin College
Brunswick, ME, U.S.A.
aredlich@bowdoin.edu

Submitted: Feb 6, 2014; Accepted: Jan 28, 2015; Published: Feb 16, 2015
Mathematics Subject Classifications: 05C80, 03C13

Abstract

Answering a question of Kolaitis and Kopparty, we show that, for given integer $q > 1$ and pairwise nonisomorphic connected graphs $G_1, \ldots, G_k$, if $p = p(n)$ is such that $\Pr(G_{n,p} \supseteq G_i) \to 1$ $\forall i$, then, with ξ_i the number of copies of G_i in $G_{n,p}$, $(\xi_1, \ldots, \xi_k)$ is asymptotically uniformly distributed on $\mathbf{Z}_q^k$.

1 Introduction

For graphs G, H write $N(G, H)$ for the number of unlabeled copies of H in G (e.g. $N(K_r, K_s) = \binom{r}{s}$). We use both $G_{n,p}$ and $G(n,p)$ for the ordinary ("binomial" or "Erdős-Rényi") random graph.

We are interested here in extending to nonconstant p the following beautiful result of Kolaitis and Kopparty [4].

Theorem 1. *Fix an integer $q > 1$, $p \in (0,1)$ and pairwise nonisomorphic connected graphs $G_1, \ldots, G_k$, each with at least two vertices, and let ξ_i be $N(G_{n,p}, G_i) \pmod q$. Then the distribution of $\xi = (\xi_1, \ldots, \xi_k)$ is $e^{-\Omega(n)}$-close to uniform on $\mathbf{Z}_q^k$. In particular, for each $a \in \mathbf{Z}_q^k$, $\Pr(\xi = a) \to q^{-k}$ as $n \to \infty$.*

(Recall two distributions are *ε-close* if their statistical (a.k.a. variation) distance is at most ε.) Essentially, this theorem states that for constants p and q, subgraphs of $G(n,p)$ are uniformly distributed modulo q.

*Supported by NSF grant DMS1201337.
†Supported by NSF grant DMS1201337.
‡Supported by NSF Award No. 1004382.

Theorem 1 was motivated by an application to 0-1 laws for first order logic with a parity quantifier or, more generally, a quantifier that allows counting modulo q; see Section 3 for a little more on this.

A natural question raised in [4] (and communicated to the authors by S.K.) asks, to what extent does Theorem 1 remain true if p is allowed to tend to zero as n grows, e.g. if $p = n^{-\alpha}$ for some fixed $\alpha > 0$? Our purpose here is to answer this question.

We need a little notation. For a graph $H = (V, E)$, set $v_H = |V|$, $e_H = |H| := |E|$, $\rho(H) = e_H/v_H$ and $m(H) = \max\{\rho(H') : H' \subseteq H, v_{H'} > 0\}$. Recall (see e.g. [2]) that $n^{-1/m(H)}$ is a threshold function for containment of H; that is, the probability that $G_{n,p}$ ($p = p(n)$) contains a copy of H tends to 0 if $pn^{1/m(H)} \to 0$ and to 1 if $pn^{1/m(H)} \to \infty$. Given a collection $\mathcal{G}$ of graphs, set $m(\mathcal{G}) = \max\{m(G) : G \in \mathcal{G}\}$, $p_{\mathcal{G}}(n) = n^{-1/m(\mathcal{G})}$ and

$$\Phi_{\mathcal{G}}(n, p) = \min_{G \in \mathcal{G}} \min\{n^{v_H} p^{e_H} : H \subseteq G, v_H > 0\}.$$

Theorem 2. *Let q, $G_1, \ldots, G_k$ and $\xi = (\xi_1, \ldots, \xi_k)$ be as in Theorem 1 and $\mathcal{G} = \{G_1, \ldots, G_k\}$. If $p = \omega(p_{\mathcal{G}}(n))$, then the distribution of ξ is* $\exp[-\Omega(\Phi_{\mathcal{G}}(n, p))]$*-close to uniform on $\mathbf{Z}_q^k$.*

(Of course the constant in the exponent depends on q and $\mathcal{G}$.)

Suppose e.g. that $q = k = 2$, $G_1 = K_3$, and $G_2 = K_4$. Then $m(\mathcal{G}) = m(G_2) = 3/2$ ($m(G_1) = 1$) and $p_{\mathcal{G}}(n) = n^{-2/3}$, so the theorem says that, asymptotically speaking, the parities of the numbers of copies of K_3 and K_4 are independent with each equally likely to be even or odd, provided $p = \omega(n^{-2/3})$.

For the special case $\mathcal{G} = \{K_3\}$, a somewhat weaker version of Theorem 2—with $\exp[-\Omega(\Phi_{\mathcal{G}}(n, p))]$ replaced by something polynomial in n and p—has been shown by Noga Alon [3].

We should also note here an immediate consequence of Theorem 2, which again answers a question from [4].

Corollary 3. *Let q, $\mathcal{G}$ be as in Theorem 1, fix a positive irrational α, and let $I = \{i \in [k] : m(G_i) < \alpha^{-1}\}$ and $J = [k] \setminus I$. Then for $p = n^{-\alpha}$ and $a \in \mathbf{Z}_q^k$ (and ξ as in Theorem 1),*

$$\Pr(\xi = a) \to \begin{cases} q^{-|I|} & \text{if } a_j = 0 \ \forall j \in J, \\ 0 & \text{otherwise.} \end{cases}$$

This is of interest partly for its possible relevance to proving a modular convergence law (again see Section 3) for $p = n^{-\alpha}$ with α irrational (*cf.* [5, Theorem 6], which says that for such p a 0-1 law holds for any first order property); but we also have, again from [4]: "Even the behavior of subgraph frequencies mod 2 in this setting [i.e. with p as in Corollary 3] seems quite intriguing."

The proof of Theorem 2, given in the next section, is similar to that of Theorem 1 in [4]. In truth, we just add one little idea to the machinery of [4]; nonetheless, as the proof answers a rather basic question, and was apparently not quite trivial to find, it seems worth recording.

2 Proof

We will need the following two facts, the first of which, from [4], generalizes a result of Babai, Nisan and Szegedy [1].

Lemma 4. *Let $q > 1$ and $d > 0$ be integers and $p \in (0,1)$. Let $\mathcal{F} \subseteq 2^{[m]}$ and let $Q(z_1, \dots z_m) \in \mathbf{Z}_q[z_1, \dots z_m]$ be a polynomial of the form*

$$\sum_{S \in \mathcal{F}} a_S \prod_{i \in S} z_i + Q'(z_1, \dots z_m),$$

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

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

Let $\mathbf{z} = (\mathbf{z}_1, \dots \mathbf{z}_m) \in \mathbf{Z}_q^m$ be the random variable where, independently for each i, $\Pr(\mathbf{z}_i = 1) = p$ and $\Pr(\mathbf{z}_i = 0) = 1 - p$. Then for $\omega \in \mathbf{C}$ a primitive q^{th}-root of unity,

$$|\mathsf{E}[\omega^{Q(\mathbf{z})}]| \leqslant 2^{-\Omega(r)}. \tag{1}$$

(We again observe that the implied constant in the $\Omega(r)$ term depends on q, p and d.)

Lemma 5 ("Vazirani XOR Lemma")**.** *Let $q > 1$ be an integer and $\omega \in \mathbf{C}$ a primitive q^{th}-root of unity. Let $\xi = (\xi_1, \dots, \xi_l)$ be a random variable taking values in $\mathbf{Z}_q^l$. Suppose that for every nonzero $c \in \mathbf{Z}_q^l$,*

$$|\mathsf{E}[\omega^{\sum c_i \xi_i}]| \leqslant \epsilon.$$

Then the distribution of ξ is $(q^l \epsilon)$-close to uniform on $\mathbf{Z}_q^l$.

Proof of Theorem 2. Letting e run over edges of K_n, the argument of [4] expresses each $\sum c_i \xi_i$ in the natural way as a polynomial in the indicators $\mathbf{z}_e := \mathbf{1}_{\{e \in G(n,p)\}}$ ($e \in E(K_n)$)—namely,

$$\sum_i c_i \xi_i = \sum_i c_i \sum \{ \prod_{e \in H} \mathbf{z}_e : G_i \cong H \subseteq K_n \}$$

—and for the $\mathcal{E}$ of Lemma 4 uses $\Omega(n)$ vertex-disjoint copies of some largest G_i among those with $c_i \neq 0$. The problem with this in the present situation is the (hidden) dependence of the bound in (1) on p.

We get around this difficulty by choosing our random graph in two steps, so that when we come to apply Lemma 4 we are back to constant p. For simplicity we now write Φ for $\Phi_{\mathcal{G}}(n,p)$, $\mathbf{G}'$ for $G(n, 2p)$ and $\mathbf{G}$ for the random subgraph of $\mathbf{G}'$ in which each edge is present, independently of other choices, with probability 1/2; in particular, our ξ_i's are functions of $\mathbf{G}$ $(= G(n,p))$.

Given $\mathbf{G}'$, we will apply Lemma 4 with variables $\mathbf{z}_e = \mathbf{1}_{\{e\in\mathbf{G}\}}$ $(e \in \mathbf{G}')$, $\mathcal{F}$ the collection of copies of $G_1,\dots,G_k$ in $\mathbf{G}'$, and $\mathcal{E} \subseteq \mathcal{F}$ a large collection of vertex-disjoint copies of an appropriate G_i; so first of all we need existence of such an $\mathcal{E}$. For a given ε, let $\mathcal{D} = \mathcal{D}_\varepsilon$ be the event that $\mathbf{G}'$ contains, for each i, a collection of $r := \varepsilon\Phi$ vertex-disjoint copies of G_i.

Proposition 6. *There is a fixed $\varepsilon > 0$ (depending on $\mathcal{G}$) for which*

$$\Pr(\overline{\mathcal{D}}) < \exp[-\Omega(\Phi)]. \tag{2}$$

Proof.
Though we don't know a reference, this is presumably not new and the ideas needed to prove it may all be found in [2];
so we just indicate what's involved.
Fix $i \in [k]$ and write H for G_i. Let Y be the maximum size of a collection of disjoint copies of H in $\mathbf{G}'$. It is enough to show that the (more properly, "a") *median* of Y is $\Omega(\Phi)$; (2) then follows *via* an inequality of Talagrand ([7] or [2, Theorem 2.29]) as in the argument for the edge-disjoint analogue of Proposition 6 given on page 77 of [2]. (In our case Talagrand's inequality says that for a median m of Y and $t > 0$, $\Pr(Y \leqslant m-t) \leqslant 2\exp[-t^2/(4\psi(m))]$, where $\psi(r) = r|H|$.)
For a lower bound on the median of Y, write X for the number of copies of H (in $\mathbf{G}'$) and Z for the number of (unordered) pairs of non-disjoint copies. Then:

(i) $\mathsf{E}(X) = \Omega(\Phi)$ (this is immediate from the definitions);

(ii) w.h.p. $X > (1-o(1))\mathsf{E}X$ (a basic application of the 2nd moment method; see [2, Remark 3.7]);

(iii) $\mathsf{E}Z < c\mathsf{E}^2X/\Phi$ for a suitable fixed c (a straightforward calculation using the definition of Φ), so with probability at least 3/4, $Z < 4c\mathsf{E}^2X/\Phi$;

(iv) by Turán's Theorem (applied to the graph with vertices the copies of H, edges the non-disjoint pairs and (therefore) independence number Y; *cf.* [2, Eq. (3.21)]), $Y \geqslant X^2/(X+2Z)$; and thus

(v) with probability at least $3/4 - o(1)$,

$$Y > \frac{(1-o(1))\mathsf{E}^2X}{\mathsf{E}X + 8c\mathsf{E}^2X/\Phi} = \Omega(\Phi)$$

(where the first inequality uses the fact that $x^2/(x+2z)$ is increasing in x for $x, z > 0$). □

In view of Proposition 6 it is enough to show that for any G' satisfying $\mathcal{D}$, the conditional distribution of ξ given $\{\mathbf{G}' = G'\}$ is $\exp[-\Omega(\Phi)]$-close to uniform on $\mathbf{Z}_q^k$. Given such a G' and $\underline{0} \neq c \in \mathbf{Z}_q^k$, take $\mathcal{F}_i$ to consist of all copies of G_i in G' $(i \in [k])$ and $\mathcal{F} = \cup\{\mathcal{F}_i :$

$c_i \neq 0\}$. Fix, in addition, some $i_0 \in [k]$ with $c_{i_0} \neq 0$ and $|G_{i_0}| = \max\{|G_i| : c_i \neq 0\} =: d$, and some $\mathcal{E} = \{E_1, \ldots, E_r\} \subseteq \mathcal{F}_{i_0}$, with the E_i's vertex-disjoint.

We have

$$\sum_{i\in[k]} c_i\xi_i = \sum_{i\in[k]} c_i \sum_{H\in\mathcal{F}_i} \prod_{e\in H} \mathbf{z}_e =: Q(\mathbf{z}),$$

where $\mathbf{z}_e = \mathbf{1}_{\{e\in\mathbf{G}\}}$ for $e \in G'$. We then need to say that Q, $\mathcal{F}$ and $\mathcal{E}$ (with q, d and $p = 1/2$) satisfy the requirements of Lemma 4. But the first three of these are immediate and the fourth follows from the connectivity of the G_i's: for $H \in \mathcal{F} \setminus \mathcal{E}$, if $V(H) \not\subseteq V(E_i)$ $\forall i$, then (since H is connected and the E_i's are vertex-disjoint) $H \not\subseteq \cup E_i$, whence $|H \cap (\cup E_i)| < |H| \leqslant d$; otherwise we have $V(H) \subseteq V(E_j)$ for some j and, since $H \neq E_j$, $|H \cap (\cup E_i)| = |H \cap E_j| < |E_i| = d$. Thus Lemma 4 applies, yielding

$$|\mathsf{E}\ \omega^{Q(z)}| \leqslant \exp[-\Omega(\Phi)], \tag{3}$$

and then (since this was for any $c \neq \underline{0}$) Lemma 5 says that, as desired, the conditional distribution of ξ given $\{\mathbf{G}' = G'\}$ is $\exp[-\Omega(\Phi)]$-close to uniform on $\mathbf{Z}_q^k$.

□

3 Discussion

As mentioned earlier, Theorem 1 is a key ingredient in the proof of the Kolaitis-Kopparty "modular convergence law" for first order logic with a parity quantifier, or, more generally, a quantifier that allows counting mod q. This law says, briefly, that, for fixed p and $n \to \infty$, the probability of a given sentence in the system under consideration tends to a limit that depends only on the congruence class of n mod q. (See also [6] for an in-depth discussion of 0-1 laws for random graphs.)

As suggested in [4], it would be interesting to understand to what extent such a law holds in the sparse setting. Theorem 2 gets about half way to this goal (for p in its range); but the other half—an assertion like Theorem 2.3 of [4] to the effect that all relevant information is contained in the subgraph frequencies—seems to require something new, since the quantifier elimination process underlying that step depends critically on properties of $G(n, p)$ that hold for constant p but fail when p tends to zero.

In closing we just mention that it would be interesting to find a proof of Theorem 2 that proceeds from first principles and does not depend on the "generalized inner product" polynomials underlying Lemma 4.

Acknowledgement

We would like to thank Swastik Kopparty for telling us the problem and for helpful conversations on the material of Section 3.

References

[1] L. Babai, N. Nisan and M. Szegedy, Multiparty protocols and logspace-hard pseudorandom sequences, in *Proc. 21st ACM Symposium on the Theory of Computing*, 1989.

[2] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, Wiley-Interscience, 2000.

[3] S. Kopparty, personal communication.

[4] P. G. Kolaitis and S. Kopparty, *Random graphs and the parity quantifier*, J. ACM, vol. 60 (5) Article 37, 2013.

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

[6] J. Spencer, *The Strange Logic of Random Graphs*, Springer, 2001

[7] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, *Inst. Hautes Études Sci. Publ. Math.* **81** (1995), 73-205.