The component structure of dense random subgraphs of the hypercube.

Colin McDiarmid
Department of Statistics,
University of Oxford,
24 - 29 St Giles',
Oxford, OX1 3LB, UK.
cmcd@stats.ox.ac.uk

Alex Scott*
Mathematical Institute,
University of Oxford,
Radcliffe Observatory Quarter,
Woodstock Road,
Oxford, OX2 6GG, UK.
scott@maths.ox.ac.uk

Paul Withers
Mathematical Institute,
University of Oxford,
Radcliffe Observatory Quarter,
Woodstock Road,
Oxford, OX2 6GG, UK.
paul.n.withers@gmail.com

Abstract

Given $p \in (0,1)$, we let $Q_p = Q_p^d$ be the random subgraph of the d-dimensional hypercube Q^d where edges are present independently with probability p. It is well known that, as $d \to \infty$, if $p > \frac{1}{2}$ then with high probability Q_p is connected; and if $p < \frac{1}{2}$ then with high probability Q_p consists of one giant component together with many smaller components which form the 'fragment'.

Here we fix $p \in (0, \frac{1}{2})$, and investigate the fragment, and how it sits inside the hypercube. For example, we give asymptotic estimates for the mean numbers of components in the fragment of each size, and

*Supported by a Leverhulme Trust Research Fellowship

describe their asymptotic distributions, much extending earlier work of Weber.

1 Introduction

The hypercube $Q = Q^d$ is the graph with vertex set $\{0,1\}^d$ and with two vertices adjacent when they differ in exactly one co-ordinate. Alternatively it can be considered as the graph on the power set of $[d] = \{1, 2, \ldots, d\}$ in which two sets are adjacent when their symmetric difference is a singleton. We consider the random subgraph $Q_p = Q_p^d$ where the edges appear independently with fixed probability p, and examine the component structure as $d \to \infty$. We say that Q_p has a property *with high probability (or whp)* if the property holds with probability tending to 1 as $d \to \infty$, and Q_p has a property *with very high probability (or wvhp)* if it holds with probability $1 - e^{-\Omega(d)}$.

Burtin [10] considered random subgraphs in the dense case and showed that, for fixed $p < 1/2$, whp Q_p is disconnected and, for fixed $p > 1/2$, whp Q_p is connected. Erdős and Spencer [11] showed that for $p = 1/2$, Q_p is connected with probability tending to e^{-1} (see also Bollobás [4, Theorem 14.3]). Also Weber [17] considered the dense case – we will discuss his work shortly. Ajtai, Komlós and Szemerédi [1] looked at the sparse case, and demonstrated that a phase transition occurs at $p = 1/d$: for $p = \lambda/d$ with $\lambda > 1$, whp the largest component of Q_p has size $\Omega(2^d)$ and the second largest has size $o(2^d)$, while for $\lambda < 1$ whp the largest component has size $o(2^d)$. Bollobás, Kohayakawa and Łuczak [5, 6, 7, 8] gave more detailed results around the phase transition at $p = 1/d$, and investigated the minimum degree, connectedness and the existence of a complete matching in the sequence of subgraphs of Q^d formed by adding edges randomly, one at a time. They showed that, almost surely, this graph process becomes connected exactly at the moment when the last isolated vertex disappears, and at this time a complete matching emerges. See [9, 13] for more recent work concerning behaviour around the phase transition and for further references.

This paper looks at the sizes of the components of Q_p for a fixed p with $0 < p < 1/2$. These graphs Q_p will be disconnected with a single large component whp. Note that we cannot expect some sort of elegant 'symmetry rule' as for Erdős-Rényi random graphs $G(n, p)$, where (roughly speaking), given the size of the largest component in a supercritical random graph $G(n, p)$, the rest of the graph looks like a subcritical $G(n', p')$ (see for example [14, section 5.6]): the geometry of the hypercube makes life more interesting and complicated.

We denote the number of vertices in a graph G by $v(G)$, and call this the *size* of G; and denote the number of edges by $e(G)$. In Q_p, we order the components by size (where components having the same size are ordered say by the position of the 'smallest' vertex of each component in some canonical ordering of the vertices). Denote the j-th component by $\mathcal{L}_j$ and let $L_j = v(\mathcal{L}_j)$ be the size of $\mathcal{L}_j$ (where $\mathcal{L}_j = \emptyset$ and $L_j = 0$ if G has less than j components). The *giant* component is $\mathcal{L}_1$. The *fragment* $\mathcal{Z}$ is the graph formed by all the components other than $\mathcal{L}_1$, and we let $Z = v(\mathcal{Z}) = 2^d - L_1$. Let X_t denote the number of components of Q_p of size t, and let $\mu_t = \mathbb{E}[X_t]$. Let $X = \sum_{t\geq 1} X_t$ be the total number of components of Q_p. Finally let $q = 1 - p$.

Observe that $\mu_1 = (2q)^d$; and that $\mu_1 \to \infty$ as $d \to \infty$, since $2q > 1$. The quantity m_p defined by

$$m_p = \lfloor 1/\log_2(1/q) \rfloor \tag{1}$$

is central to our results. Observe that m_p is large for small p and decreases to 1 as p increases to 1/2. For an integer t, we have $2q^t \geq 1 \Leftrightarrow t \leq m_p$. In particular, we always have $m_p \geq 1$ since $2q > 1$;

Weber [17] showed that whp the fragment size Z satisfies $Z \sim \mu_1$ (that is, $Z = (1 + o(1))\mu_1$), the second largest component size L_2 satisfies $L_2 = m_p$, and the number X_t of components of size t satisfies $X_t \sim \mu_t = \Theta(d^{t-1}(2q^t)^d)$ for each $t = 1, \ldots, m_p$; and it follows that the total number X of components satisfies $X \sim \mu_1$ whp. We much extend and sharpen these results, presenting our results in six theorems. Weber's results in [17] are contained within Theorems 1 and 4 below. (Weber later introduced also a probability for vertices to appear in the random subgraph of Q^d [19], but we do not pursue that extension here.)

Our first three theorems concern the global behaviour of components in Q_p; the next two theorems concern more local behaviour (and are needed to prove the earlier ones); and our last theorem, Theorem 6, concerns the joint distribution of random variables like the X_t.

Throughout, we fix $0 < p < 1/2$ and let $q = 1 - p$. The first theorem can be introduced now, with no further definitions. It describes the total number X of components in Q_p, the size $Z = 2^d - L_1$ of the fragment, and the size L_2 of the second largest component. Note that, as $d \to \infty$, we have $d \ll \mu_1$ and so $\sqrt{d\mu_1} \ll \mu_1$.

Theorem 1. *For fixed $0 < p < 1/2$, the random graph $Q_p = Q_p^d$ satisfies the following.*

(a) Let Y be either the number X of components of Q_p or the fragment size Z. Then $\mathbb{E}[Y] = \mu_1(1 + \Theta(dq^d))$; and for each $\varepsilon > 0$ we have $|Y - \mathbb{E}[Y]| < \varepsilon\sqrt{d\mu_1}$ wvhp.

(b) *The second largest component size L_2 in Q_p satisfies $L_2 = m_p$ wvhp, where m_p is as in (1). Also, the mean and variance satisfy $|\mathbb{E}[L_2] - m_p| = e^{-\Omega(d)}$ and $\mathrm{Var}(L_2) = e^{-\Theta(d)}$.*

Our second theorem concerns how the fragment sits in Q^d. How much do the components of the fragment cluster together? How far is it typically from a fixed vertex to the fragment $\mathcal{Z}$ of Q_p? Given a vertex u in Q^d and $r > 0$, the *r-ball $B_r(u)$ around u* is the set of vertices v at graph distance at most r from u (in Q^d). Recall that, for $0 < \eta < 1$, the *entropy $h(\eta)$* is defined to be $-\eta \log_2 \eta - (1-\eta)\log_2(1-\eta)$, and it is strictly increasing on $(0, \frac{1}{2})$ with image $(0, 1)$. Let $\eta^* = \eta^*(p)$ be the unique solution to $h(\eta) = \log_2 \frac{1}{1-p}$ with $0 < \eta < \frac{1}{2}$. For example, if $p = \frac{1}{4}$ then $\eta^* \approx 0.08$.

Theorem 2. *For fixed $0 < p < 1/2$, the random graph $Q_p = Q_p^d$ satisfies the following.*

(a) *There exists $\delta = \delta(p) > 0$ such that wvhp each δd-ball in Q^d contains at most m_p vertices of the fragment.*

(b) *For each $\varepsilon > 0$ there is $\gamma = \gamma(\varepsilon, p) > 0$ such that wvhp a proportion at most $e^{-\gamma d}$ of the vertices in Q^d are within distance $(\eta^* - \varepsilon)d$ of the fragment $\mathcal{Z}$, but all vertices are within distance $(\eta^* + \varepsilon)d$. (All distances are in Q^d.)*

In part (a) above, clearly wvhp there are δd-balls containing at least m_p vertices of the fragment – consider for example any ball with centre in a component of size m_p. Thus the statement that wvhp no δd-ball in Q^d contains strictly more than m_p vertices of the fragment is saying strongly that the components of the fragment $\mathcal{Z}$ do not cluster together in Q^d. For example, wvhp no component of $\mathcal{Z}$ of size m_p is within distance δn of any other component of $\mathcal{Z}$.

In part (b), many vertices are at a short distance in Q^d from the fragment $\mathcal{Z}$, including of course the vertices in $\mathcal{Z}$, but only a very small proportion of the total are at distance at most $(\eta^* - \varepsilon)d$. However, when $r = (\eta^* + \varepsilon)d$, wvhp every r-ball contains a vertex in $\mathcal{Z}$ (and indeed contains $2^{\Omega(d)}$ vertices in $\mathcal{Z}$). Overall, the giant gets everywhere, and indeed the fragment is heavily outnumbered everywhere.

The next theorem amplifies part (a) of Theorem 1, concerning the number X of components and the fragment size Z. Recall first that, for two random variables Y and Y' taking values in a countable set, the *total variation distance* between their distributions is given by

$$d_{TV}(Y, Y') = \frac{1}{2} \sum_k |\mathbb{P}(Y = k) - \mathbb{P}(Y' = k)|.$$

We use $d_{TV}(Y,\mathrm{Po}(\lambda))$ to denote $d_{TV}(Y,Y')$ where Y' has the Poisson distribution $\mathrm{Po}(\lambda)$ with mean λ. Several of our proofs will involve bounding $d_{TV}(Y,\mathrm{Po}(\mathbb{E}[Y]))$ for relevant random variables Y (like X_t or X), using results on Poisson approximation based on the Stein-Chen method. By a standard tail bound (see, for example, inequality (2.9) and Remark 2.6 in [14]), for any random variable Y and $\lambda > 0$, for each $t > 0$ we have

$$\mathbb{P}(|Y-\lambda| \geq t\sqrt{\lambda}) \leq 2e^{-t^2/3} + d_{TV}(Y,\mathrm{Po}(\lambda)). \tag{2}$$

Also, given a (non-trivial) random variable $Y = Y_d$ we let Y^* denote the natural centred and rescaled version $(Y-\mathbb{E}[Y])/\sqrt{\mathrm{Var}(Y)}$. It is well known (see for example [2]) that if Y_n is a sequence of random variables with mean λ_n such that $d_{TV}(Y_n,\mathrm{Po}(\lambda_n)) \to 0$ and $\lambda_n \to \infty$ as $n \to \infty$, then $(Y_n-\lambda_n)/\sqrt{\lambda_n}$ is asymptotically standard normal. Thus if also $\mathrm{Var}(Y_n) \sim \lambda_n$ then Y_n^* is asymptotically standard normal.

Theorem 3. *Fix $0 < p < 1/2$ and let $q = 1-p$. In $Q_p = Q_p^d$, let Y either be the number X of components or be the fragment size Z. Then the following properties hold as $d \to \infty$.*

(a) $\lambda := \mathbb{E}[Y] = (1+\Theta(dq^d))\,\mu_1$ and $\mathrm{Var}(Y) = (1+O(dq^d))\,\mu_1$.

(b) $d_{TV}(Y,\mathrm{Po}(\lambda))$ is $O(dq^d)$, and Y^* is asymptotically standard normal.

Observe that part (a) of Theorem 1 follows directly from inequality (2) and Theorem 3: for

$$\begin{aligned}\mathbb{P}(|Y-\mathbb{E}[Y]| \geq \varepsilon\sqrt{d\mu_1}) &\leq 2e^{-\frac{1}{3}\varepsilon^2 d\mu_1/\lambda} + d_{TV}(Y,\mathrm{Po}(\lambda)) \\ &\leq e^{-(\frac{1}{3}+o(1))\varepsilon^2 d} + O(dq^d).\end{aligned}$$

The remaining theorems concern more local behaviour. The first counts small components by size. It is needed in order to prove the earlier theorems. Recall that X_t is the number of components of size t in Q_p, and $\mu_t = \mathbb{E}[X_t]$. We noted earlier that $\mu_1 = (2q)^d$. It is not hard to give exact formulae also for μ_2 and μ_3 (assuming $d \geq 2$), namely

$$\mu_2 = (p/2q^2)\,d\,(2q^2)^d \quad \text{and} \quad \mu_3 = (p^2/2q^4)\,d(d-1)\,(2q^3)^d \tag{3}$$

(see also the discussion following Theorem 5).

Theorem 4. *Fix $0 < p < \frac{1}{2}$, let $q = 1-p$, and let $1 \leq t \leq m_p$. Then the following results concerning the number X_t of components of size t in $Q_p = Q_p^d$ hold, as $d \to \infty$.*

(a) $\mu_t = (1+O(\frac{1}{d}))\,\frac{t^{t-2}}{t!}\,(\frac{p}{q^2})^{t-1}\,d^{t-1}(2q^t)^d$ *and* $\mathrm{Var}(X_t) = (1+O(d^t q^{td}))\mu_t$.

(b) For each $\varepsilon > 0$*, we have* $|X_t - \mu_t| < \varepsilon\sqrt{d\mu_t}$ *wvhp, and so also* $|X_t - \mu_t| < \varepsilon\mu_t$ *wvhp.*

(c) $d_{TV}(X_t, \mathrm{Po}(\mu_t)) = O(d^t q^{td})$*, and* X_t^* *is asymptotically standard normal.*

Observe from part (a) that $\mu_t = \Omega(d)$ since $2q^t \geq 1$ (and indeed $\mu_t \gg d$ unless $p = 1 - 1/\sqrt{2}$ and $t = m_p = 2$), so the first half of part (b) above implies the second half. For a partial local limit result corresponding to part (c), see Proposition 15 at the end of Section 3.

These results help us to visualise the asymptotic disappearance of small components in Q_p as p increases from 0 to 1/2. For each fixed p, there are wvhp a giant component and many small components of every size up to a maximum size m_p. In particular $\mu_t \to \infty$ as $d \to \infty$ for each $t \leq m_p$. We noted that m_p is large for small p and decreases to 1 as p increases to 1/2. The typical number of components decreases exponentially as p increases and the maximum size L_2 of a component of the fragment drops as $1/\log_2(1/q)$ falls below each integer value. In particular, the last components of size 2 disappear as p increases past $1-1/\sqrt{2} \approx 0.29$ and the last isolated vertices disappear as p increases past 1/2. We recall that $Q_{1/2}$ is connected with probability tending to e^{-1} as $d \to \infty$. Indeed, whp $Q_{1/2}$ consists of X isolated vertices and a connected component of $2^d - X$ vertices, where X has mean value 1 and asymptotic distribution Po(1) (see [11]).

Ambient isomorphisms

We shall in fact prove a much finer and more detailed version of Theorem 4, namely Theorem 5, which uses a natural restricted version of isomorphism for subgraphs of the cube, so that we can consider also how components 'sit' in the host hypercube. We then deduce Theorem 4 from Theorem 5.

We call a graph a *cube subgraph* if it is a subgraph of the cube Q^d for some d. Let H be a connected cube subgraph. The *support* $S(H)$ is the set of indices i such that there is an edge xy in H with $x_i = 0$ and $y_i = 1$ (that is, H meets both top and bottom faces in the i-th coordinate direction). Call $|S(H)|$ the *span* of H, denoted by $\mathrm{span}(H)$. Note that if H consists of a single vertex then $\mathrm{span}(H) = 0$, and otherwise $\mathrm{span}(H) \geq 1$. Indeed, if $v(H)$ is 1, 2 or 3 then $\mathrm{span}(H) = v(H) - 1$, whereas for example if H is a 4-vertex path then $\mathrm{span}(H)$ could be 2 or 3.

The *canonical copy* H^* of H is defined as follows. If H is a single vertex then its canonical copy is the graph Q^0 (consisting of a single vertex). Suppose that H has at least one edge, so $s := \mathrm{span}(H) \geq 1$. Let ϕ be the increasing injection from $[s]$ to $[d]$ with image $S(H)$. Given $x = (x_1, x_2, \ldots, x_d) \in Q^d$

let $\phi(x) = (x_{\phi(1)}, x_{\phi(2)}, \ldots, x_{\phi(s)}) \in Q^s$. Then the vertices of the canonical copy H^* are the points $\phi(x)$ where x is a vertex of H; and the edges of H^* are the pairs $\phi(x)\phi(y)$ such that xy is an edge of H. (Note that the canonical copy is a subgraph of Q^s.) See Figure 1 for an illustration.

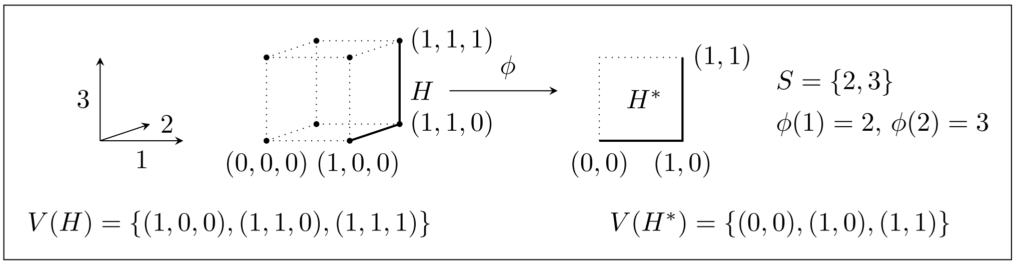

Figure 1: A subgraph H of Q^3 with canonical copy H^* in Q^2

We say that connected subgraphs H_1 of Q^{d_1} and H_2 of Q^{d_2} are *ambient isomorphic* if they have the same canonical copy. Of course, if H_1 and H_2 are ambient isomorphic then they are isomorphic, but this definition is stronger in that it requires the copies to 'sit in the cube' in the same way. For example, let O denote the zero d-vector and let e_k denote the kth unit d-vector: if $i < j$ then the three vertex path $O, e_i, e_i + e_j$ in Q^d has canonical copy the path $(0, 0), (1, 0), (1, 1)$ in Q^2 as in Figure 1, and so the original path in Q^d is not ambient isomorphic to the path $O, e_j, e_i + e_j$ which has canonical copy the path $(0, 0), (0, 1), (1, 1)$. There are four ambient isomorphism classes of three-vertex paths. Observe that if $s = \text{span}(H)$ then there is a unique subgraph of Q^s ambient isomorphic to H (namely the canonical copy of H).

Our fifth theorem concerns numbers of components ambient isomorphic to given connected cube subgraphs H_i. Note that any two subcubes of Q^d with the same dimension are ambient isomorphic. Weber [18] considered Poisson convergence of the number of subcube components of Q^d_p of a given dimension, for a range of values of p which could depend on d. Here we keep p fixed, but we consider all kinds of components. Recall that $(d)_k$ means $d(d-1)\cdots(d-k+1)$.

Theorem 5. *Let $0 < p < 1/2$ and $q = 1-p$. Let $r \geq 1$ and let $H_1, H_2, \ldots, H_r$ be pairwise non-ambient-isomorphic connected cube subgraphs each with at most m_p vertices. Let $t = \min_{i\in[r]} v(H_i)$ and $s = \max\{\text{span}(H_i) : v(H_i) = t\}$. (All these quantities are fixed, not depending on d.)*

For each i, let $Y_i = Y_i(d)$ be the (random) number of components of Q^d_p ambient-isomorphic to H_i. Let $Y = Y(d) = \sum_i Y_i$ and let $\lambda = \lambda(d) = \mathbb{E}[Y]$. Then the following hold.

(a) *There is a constant $c > 0$, given explicitly in equations (4) and (5) below, such that $\lambda = (1+O(1/d))\, c\, (d)_s (2q^t)^d$; and if t is 1, 2 or 3 then $s = t-1$, and we may replace the error bound $O(1/d)$ by $O(dq^d)$. Also* $\mathrm{Var}(Y) = (1 + O(d^t q^{td}))\, \lambda$.

(b) *For each $\varepsilon > 0$, we have $|Y - \lambda| < \varepsilon\sqrt{d\lambda}$ wvhp, and so also $|Y - \lambda| < \varepsilon\lambda$ wvhp.*

(c) $d_{TV}(Y, \mathrm{Po}(\lambda)) = O(d^t q^{td})$, *and Y^* is asymptotically standard normal.*

By part (a), λ is $\Omega(d)$ (and indeed λ is $\Omega(d^2)$ except if $p = 1 - 1/\sqrt{2}$ and $t = m_p = 2$), so the first half of part (b) implies the second half (as with Theorem 4). See Lemma 12 for a fuller version of Theorem 5, which considers more information about the components counted. That lemma, together with the estimates of μ_t from Lemma 13, will yield Theorem 4, by letting $H_1, \ldots, H_r$ list all the t-vertex connected canonical cube subgraphs, so that the random variable Y in Theorem 5 is X_t.

The constant c in part (a) may be specified as follows. Let $I^* = \{i \in [r] : v(H_i) = t,\ \mathrm{span}(H_i) = s\}$. For each $i \in I^*$, let $e'(H_i)$ be the number of edges of Q^d not in H_i but with both end vertices in H_i, and let

$$\beta_i = \frac{1}{2^s s!} \left(\frac{p}{q^2}\right)^{e(H_i)} \left(\frac{1}{q}\right)^{e'(H_i)}. \tag{4}$$

Now let

$$c = \sum_{i \in I^*} \beta_i. \tag{5}$$

If $t = 1$ then $c = 1$. If $t = 2$ then $c = p/2q^2$, so $\lambda \sim (p/2q^2)\, d(2q^2)^d$. If $t = 3$ then $1 \leq |I^*| \leq 4$ and each $\beta_i = \frac{1}{8}(p/q^2)^2$, so if $|I^*| = 4$ we have $\lambda \sim (p^2/2q^4)\, d^2 (2q^3)^d$. These results are in accord with (3).

In Theorem 4 we saw that wvhp in Q_p there are components of each size up to m_p. In Theorem 5 we see in much more detail that each connected cube subgraph of size at most m_p, with its way of sitting within the host hypercube, appears wvhp as a component of Q_p.

What we call ambient isomorphism could be called ‘ordered ambient isomorphism’, since we insist that the injection ϕ in the definition is increasing. If we drop this requirement then essentially the same results hold (mutatis mutandis), since the new isomorphism classes are unions of the old ones. When we deduce Theorem 4 from Theorem 5/Lemma 12, we may think of this as relaxing *ambient isomorphism* all the way to *isomorphism.*

Given a connected cube subgraph H, let $p_H = p_H(d)$ be the probability that Q_p has a component ambient isomorphic to H. When p is fixed with

$0 < p < \frac{1}{2}$, by Theorem 5, either p_H or $1 - p_H$ is $e^{-\Omega(d)}$. To see this, let $t = v(H)$, let Y be the number of components ambient isomorphic to H and $\lambda = \mathbb{E}[Y]$. If $t > m_p$ then $2q^t < 1$, so $\mathbb{P}(Y \geq 1) \leq \lambda = e^{-\Omega(d)}$; and if $t \leq m_p$ then $\lambda \to \infty$ (as we saw above), and by part (b) of Theorem 5 wvhp $Y \geq \lambda/2 > 0$. The situation described above is in contrast with the situation at $p = \frac{1}{2}$, when (as we noted earlier) the number of isolated vertices has asymptotic distribution Po(1).

Joint distribution of components

We saw in Theorem 4 that, for each $t = 1, \ldots, m_p$ the number X_t of components of Q_p of size t has close to the Poisson distribution $\mathrm{Po}(\mu_t)$, where $\mu_t = \mathbb{E}[X_t]$. In fact more is true: the joint distribution of $X_1, \ldots, X_{m_p}$ is close to a product of these distributions. Write $\mathcal{L}(X_1, \ldots, X_{m_p})$ for the joint law of $X_1, \ldots, X_{m_p}$; and write $\prod_{j=1}^{m_p} \mathrm{Po}(\mu_j)$ for the joint distribution of independent random variables $\mathrm{Po}(\mu_j)$. We shall see that

$$d_{TV}\big(\mathcal{L}(X_1, \ldots, X_{m_p}), \prod_{j=1}^{m_p} \mathrm{Po}(\mu_j)\big) = O(d^2 q^d). \tag{6}$$

Thus, the numbers of components in the fragment of each size t are asymptotically independent, with a Poisson distribution for $t \leq m_p$, and identically 0 for $t > m_p$. Indeed, we have the following much more detailed theorem concerning the small components, in the spirit of Theorem 5. Note that there is a finite set of canonical cube subgraphs with at most m_p vertices.

Theorem 6. *Let $H_1, \ldots, H_r$ be a list of $r \geq 1$ distinct canonical cube subgraphs each with at most m_p vertices. For each $j \in [r]$, let Y_j be the random number of components of $Q_p = Q_p^d$ ambient isomorphic to H_j, with mean λ_j. Let $t^* = \min_j v(H_j)$. Then*

$$d_{TV}\big(\mathcal{L}(Y_1, \ldots, Y_r), \prod_{j=1}^{r} \mathrm{Po}(\lambda_j)\big) = O(d^{t^*+1} q^{t^* d}). \tag{7}$$

When the H_j include all the canonical cube subgraphs of size up to m_p (so $t^* = 1$), Theorem 6 directly implies (6). We cannot quite use Theorem 6 to deduce our earlier individual bounds on d_{TV}, for example on $d_{TV}(X_t, \mathrm{Po}(\mu_t))$ in Theorem 4 part (c), since in the bound (7) there is an 'extra' factor d.

Notation

We use standard notation throughout. For non-negative functions f and g, we say that $f(d) = \Omega(g(d))$ if $\liminf_{d\to\infty} f(d)/g(d) > 0$, and $f(d) =$

$\Theta(g(d))$ if both $f(d) = \Omega(g(d))$ and $g(d) = \Omega(f(d))$. Also, we write $f \ll g$ if $f(d) = o(g(d))$.

Plan of the paper

Section 2 gives preliminary results, first concerning subgraphs in the hypercube Q^d, and then concerning the variance of counting random variables and their closeness to a Poisson distribution. In Section 3, Lemma 12 gives several results concerning numbers of components ambient-isomorphic to a given list of connected cube subgraphs. Lemma 13 gives quite precise results on the expected value of X_t for $1 \leq t \leq m_p$. These lemmas allow us to prove Theorem 5, and then Theorem 4, at the end of the section.

In order to prove Theorems 1, 2 and 3 we must show that with tiny failure probability there is just one component of size strictly greater than m_p. To do this, in Section 4 we call a vertex 'good' if its degree in Q_p is at least half the expected value dp . We show that, with tiny failure probability, all good vertices are in the same component; and then deduce that, for a suitable constant N, with tiny failure probability each component of the fragment has size at most N. From this result, we see in particular that wvhp m_p is an upper bound for the size L_2 of a second largest component. In Section 5 we complete the proofs of Theorems 1, 2 and 3. In Section 6 we consider joint distributions and prove Theorem 6. Finally, Section 7 contains some very brief concluding remarks.

These investigations arose from work on multicommodity flows in the cube Q^d when edges have independent random capacities, see [16].

2 Preliminary results

2.1 Preliminary results on the hypercube Q^d

Let us first consider span(H) for a connected cube subgraph H. We have already noted that span$(H) = v(H) - 1$ if $v(H)$ is 1, 2 or 3. It is easy to see that always span$(H) \leq v(H) - 1$, and the inequality is strict if H is not a tree (since any cycle contains at least two edges in some dimension). If we have equality we call H a *spreading tree.* Note that each edge of a spreading tree sits in a distinct dimension, and if T_1 and T_2 are ambient isomorphic trees then T_1 is spreading if and only if T_2 is spreading.

What are the subcubes in Q^d? If we are given $S \subseteq [d]$ and $z \in \{0,1\}^{[d]\setminus S}$, then clearly the vertices x such that $x_j = z_j$ for each $j \in [d]\setminus S$ form a subcube isomorphic to $Q^{|S|}$. We shall need to consider such 'cylinder' subcubes, for example in the proof of Lemma 8. As an aside, let us note that each cube subgraph H isomorphic to a hypercube Q^s is obtained in this way. Since $v(H) = 2^s$, this is easily seen to be equivalent to showing that H has span s; and it is a straightforward exercise to show the latter.

Proposition 7. *Let H be a subgraph of Q^d isomorphic to a hypercube Q^s. Then* $\operatorname{span}(H) = s$. □

Next we investigate the number $n_H = n_H(d)$ of subgraphs of Q^d ambient-isomorphic to a given subgraph H, the number of subgraphs which are spreading trees of a given size t, and the total number of connected subgraphs of size t.

Lemma 8. *(a) For each connected subgraph H of Q^d, $n_H = 2^{d-s}\binom{d}{s}$, where* $\operatorname{span}(H) = s$.

(b) For each $d \geq t-1 \geq 0$, the number of ambient-isomorphism classes of spreading trees of size t in Q^d is $2^{t-1}t^{t-3}$.

(c) For each $d \geq t-1 \geq 0$, the number of subgraphs of Q^d which are spreading trees of size t is $2^d\, t^{t-3}\binom{d}{t-1}$.

(d) For each fixed $t \geq 1$, the number of connected subgraphs of Q^d of size t is $2^d\, t^{t-3}\binom{d}{t-1}(1 + O(d^{-1}))$.

We see from parts (c) and (d) above that the population of connected subgraphs of a given size t in Q^d is asymptotically dominated by spreading trees.

Proof. We first recall that any cube subgraph of size t can be embedded in Q^{t-1} and so, for $d \geq t-1$, the number of pairwise non-ambient-isomorphic connected cube subgraphs of size t depends only on t.

(a) There is a single ambient-isomorphic copy of H in each (cylinder) subcube Q^s of Q^d, and there are $2^{d-s}\binom{d}{s}$ copies of Q^s in Q^d, so $n_H = 2^{d-s}\binom{d}{s}$, as required.

(b) By Cayley's formula there are t^{t-2} trees on the set $\{0, 1, 2, \ldots, t-1\}$ of t vertices. Given one of these trees, call vertex 0 the root and move the other vertex labels onto the edge leading towards the root. This constructs a vertex-rooted, edge-labeled tree, with edge-labels $1, 2, \ldots, t-1$. The construction is reversible, so there are exactly t^{t-2} such trees.

Given such a rooted, edge-labeled tree T, we choose a vertex in Q^{t-1} for the root, then use the labels of the edges to specify the 'dimension' in which that edge exists. This defines a t-vertex rooted spreading tree, and all the rooted trees constructed are distinct; and furthermore every t-vertex rooted spreading tree in Q^{t-1} can be constructed in this way. Thus there are $2^{t-1}t^{t-2}$ t-vertex rooted spreading trees in Q^{t-1}, and so $2^{t-1}t^{t-3}$ t-vertex unrooted spreading trees; and of these unrooted trees, no two distinct ones are ambient-isomorphic since they have span $t-1$ and so are their own canonical copies.

(c) By parts (a) and (b), the number of t-vertex spreading trees in Q^d is

$$2^{t-1}t^{t-3}\cdot 2^{d-(t-1)}\binom{d}{t-1} = 2^d t^{t-3}\binom{d}{t-1}.$$

(d) If T is a spreading tree of size t, and H is a connected cube subgraph of size t with $\mathrm{span}(H) < t-1 = \mathrm{span}(T)$, then $n_H/n_T = O(d^{-1})$ by part (a). The number of ambient-isomorphism classes of connected subgraphs of Q^d of size t does not depend on d for $d \geq t-1$; and thus the contribution to the total number of connected subgraphs of Q^d of size t by those with span less than $t-1$ is $O(d^{-1})$ of the total. □

We will need one more lemma which we will apply to the hypercube Q^d. This result is 'folk knowledge' (and indeed a more precise result is known, see equation (8)) but we give a short combinatorial proof here for completeness.

Lemma 9. *Let the graph G be rooted at vertex r and have maximum degree at most d. Then for each non-negative integer t, the number of subtrees containing r and exactly t other vertices is at most $(ed)^t$.*

Proof. We first show (a) that the number of $(t+1)$-vertex subtrees in G containing r is at most the number $f(d,t+1)$ of $(t+1)$-vertex subtrees containing the root in an infinite d-ary tree T^∞; and then show (b) that $f(d,t+1)$ is at most the number of points $x \in \{0,1\}^{td}$ with t 1's. The number of such points is $\binom{td}{t} \leq (ed)^t$. Clearly we may assume that $t \geq 1$.

The *path tree* $T(G,r)$ [12] has a vertex for each path P in G from r, adjacent to each vertex corresponding to a path extending P by one edge; and as the root has the vertex corresponding to the path with a single vertex r. It is easy to see that, for each tree in G containing r, there is a corresponding tree in $T(G,r)$ containing the root. Thus the the number of $(t+1)$-vertex subtrees in G containing r is at most the number of $(t+1)$-vertex subtrees containing the root in $T(G,r)$; and since $T(G,r)$ embeds in T^∞, part (a) of the proof follows.

For part (b), let T be a $(t+1)$-vertex subtree in T^∞ containing the root. We may suppose that T^∞ is embedded in the plane, with the root at the top and children listed in order from left to right. We construct $x(T) \in \{0,1\}^{td}$

with t 1's as follows. Initially the vector x is null and the list L contains just the root. We repeat the following t times. Remove the first vertex v in L, and let $y \in \{0,1\}^d$ indicate its children (with a 1 for each child): append y to x and append the children to L (listed in order). The output $x(T)$ is the final value of x. Clearly we can reconstruct T from $x(T)$, so the number of possible trees T is at most the number of possible vectors $x(T)$, which completes the proof. □

We shall not use this result here, but the precise value of $f(d,t)$ is given by

$$f(d,t) = \frac{1}{(d-1)t+1}\binom{dt}{t} \quad \text{for each } d,t \geq 1\,, \tag{8}$$

see exercise 11 in [15, section 2.3.4.4] (pages 397 and 589).

2.2 Preliminary results on variance and approximation to Poisson distribution

Let $(A_i : i \in I)$ be a family of events with a dependency graph L (so that A_i and A_j are independent if i and j are not adjacent in L and $i \neq j$). Write $i \sim j$ if i and j are adjacent in L. For each i, let $\pi_i = \mathbb{P}(A_i)$ and let $\mathbb{I}_i$ be the indicator function of A_i. Let $X = \sum_i \mathbb{I}_i$ (in this subsection we do not use X as the number of components in Q_p). Then

$$\begin{aligned}\mathrm{Var}(X) &= \sum_i \sum_j (\mathbb{P}(A_i \wedge A_j) - \pi_i \pi_j) \\ &= \sum_i (\pi_i - \pi_i^2) + \sum_i \sum_{j \sim i} (\mathbb{P}(A_i \wedge A_j) - \pi_i \pi_j) \\ &= \mathbb{E}[X] + \Delta^+ - \Delta^-, \end{aligned} \tag{9}$$

where

$$\Delta^+ = \sum_i \sum_{j \sim i} \mathbb{P}(A_i \wedge A_j) \tag{10}$$

and

$$\Delta^- = \sum_i \pi_i^2 + \sum_i \sum_{j \sim i} \pi_i \pi_j. \tag{11}$$

The following lemma is essentially Theorem 6.23 of [14], proved by the Stein-Chen method, which shows that a sum X as above has close to a Poisson distribution, provided Δ^+ and Δ^- are small.

Lemma 10. *With notation as above, and letting $\lambda = \mathbb{E}[X]$, we have*

$$d_{TV}(X, \mathrm{Po}(\lambda)) \;\; \leq \;\; \min\{\lambda^{-1}, 1\} \left(\Delta^+ + \Delta^-\right).$$

We shall also need a minor extension of the above. Suppose that we are given a family $(t_i : i \in I)$ of positive integers, and let $\tilde{X} = \sum_i t_i \mathbb{I}_i$. Then much as above, we have

$$\begin{aligned}\mathrm{Var}(\tilde{X}) &= \sum_i \sum_j t_i t_j \left(\mathbb{P}(A_i \wedge A_j) - \pi_i \pi_j\right) \\ &= \sum_i t_i^2(\pi_i - \pi_i^2) + \sum_i \sum_{j \sim i} t_i t_j (\mathbb{P}(A_i \wedge A_j) - \pi_i \pi_j) \\ &= \mathbb{E}[\tilde{X}] + \tilde{\Delta}^+ - \tilde{\Delta}^- \qquad (12)\end{aligned}$$

where

$$\tilde{\Delta}^+ = \sum_i t_i(t_i - 1)\pi_i + \sum_i \sum_{j \sim i} t_i t_j \, \mathbb{P}(A_i \wedge A_j) \qquad (13)$$

and

$$\tilde{\Delta}^- = \sum_i t_i^2 \pi_i^2 + \sum_i \sum_{j \sim i} t_i t_j \, \pi_i \pi_j. \qquad (14)$$

Lemma 11. *With notation as above, and letting $\lambda = \mathbb{E}[\tilde{X}]$, we have*

$$d_{TV}(\tilde{X}, \mathrm{Po}(\lambda)) \;\; \leq \;\; \min\{\lambda^{-1}, 1\}\big(\tilde{\Delta}^+ + \tilde{\Delta}^-\big).$$

Proof. Replace each event A_i by t_i identical (not independent) copies. Note that, for each i, the t_i copies of A_i are dependent, and so they are adjacent to each other in the natural extended dependency graph. Now apply Lemma 10. □

3 The numbers of small components

The first lemma in this section, Lemma 12, gives expected values and variances for the numbers of small components in certain ambient-isomorphism classes, and for the number of vertices in such components; and gives some results on approximation by a Poisson distribution. The second lemma uses Lemma 12, together with counting results from Subsection 2.1, to deduce results corresponding to those in Lemma 12 when we consider *all* components of a given size. Using these lemmas we prove Theorem 5 and then Theorem 4.

In Lemma 12, we consider both the numbers of components in Q_p ambient isomorphic to given graphs, and the total numbers of vertices in such components.

Lemma 12. *Let $0 < p < \frac{1}{2}$ and let $q = 1 - p$. Let r be a positive integer and let $H_1, H_2, \ldots, H_r$ be pairwise non-ambient-isomorphic connected cube subgraphs. For each $i \in [r]$, let $s_i = \mathrm{span}(H_i)$, and recall that $e'(H_i)$ is the number of cube edges not in H_i but with both end vertices in H_i. (All these quantities are fixed, not depending on d.)*

For each $i \in [r]$, let Y_i be the number of components of Q_p ambient-isomorphic to H_i. Let $t = \min_i v(H_i)$, and let $s = \max\{s_i : v(H_i) = t\}$. Let $I^* = \{i \in [r] : v(H_i) = t, s_i = s\}$, and let

$$c = \frac{1}{2^s s!} \sum_{i \in I^*} (p/q^2)^{e(H_i)} q^{-e'(H_i)}.$$

Then the following hold.

(a) For each $i \in [r]$, once $d \geq s_i$ we have

$$\mathbb{E}[Y_i] = (p/q^2)^{e(H_i)} q^{-e'(H_i)} \, 2^{d-s_i} \binom{d}{s_i} q^{v(H_i)d}.$$

(b) The sum $Y = \sum_{i=1}^r Y_i$ satisfies (i) $\mathbb{E}[Y] = (1 + O(1/d))\, c\, (d)_s (2q^t)^d$, (ii) $\mathrm{Var}(Y) = (1 + O(d^t q^{td}))\, \mathbb{E}[Y]$, and (iii) $d_{TV}(Y, \mathrm{Po}(\mathbb{E}[Y])) = O(d^t q^{td})$. Furthermore, if t is 1, 2 or 3 then $s = t - 1$ and in the expression for $\mathbb{E}[Y]$ we can improve the error term, so $\mathbb{E}[Y] = (1 + O(dq^d))\, c\, (d)_{t-1} (2q^t)^d$.

(c) The weighted sum $\tilde{Y} = \sum_{i=1}^r v(H_i) Y_i$ satisfies (i) $\mathbb{E}[\tilde{Y}] = (1 + O(dq^d))\, t\, \mathbb{E}[Y]$. Furthermore, if $t = 1$ then (ii) $\mathrm{Var}(\tilde{Y}) = (1 + O(dq^d))\, \mathbb{E}[\tilde{Y}]$ and (iii) $d_{TV}(\tilde{Y}, \mathrm{Po}(\mathbb{E}[\tilde{Y}])) = O(dq^d)$.

Proof. (a) Consider a fixed graph H_i. Let G be a subgraph of Q^d which is ambient-isomorphic to H_i, and let A be the event that the subgraph of Q_p induced by the vertices of G is exactly G, and it is also a component of Q_p. Then

$$\mathbb{P}(A) = p^{e(H_i)} q^{e'(H_i)} q^{v(H_i)d - 2e(H_i) - 2e'(H_i)} = (p/q^2)^{e(H_i)} q^{-e'(H_i)} q^{v(H_i)d}. \tag{15}$$

Hence, by Lemma 8 part (a)

$$\mathbb{E}[Y_i] = 2^{d-s_i} \binom{d}{s_i} \, (p/q^2)^{e(H_i)} q^{-e'(H_i)} \, q^{v(H_i)d},$$

completing the proof of part (a).

(b) Observe from part (a) that $\mathbb{E}[Y_i] = \Theta\big(d^{s_i} (2q^{v(H_i)})^d\big)$. Thus the dominant contribution to $\mathbb{E}[Y]$ is from graphs H_i with $i \in I^*$ (for if $i \in I^*$ and $j \in I \backslash I^*$, then $\mathbb{E}[Y_j] = O(1/d)\, \mathbb{E}[Y_i]$). Using part (a) we now see that

$$\mathbb{E}[Y] = (1 + O(1/d))\, c (d)_s (2q^t)^d = (1 + O(1/d))\, c d^s (2q^t)^d.$$

Now suppose that t is 1, 2 or 3. If $i \in I^*$ and $j \in I \backslash I^*$, then $v(H_j) > t$ so $\mathbb{E}[Y_j] = O(dq^d)\,\mathbb{E}[Y_i]$ (note that if $v(H_j) = t+1$ then $s_j \leq s+1$). Hence $\mathbb{E}[Y] = (1+O(dq^d))\,c(d)_s(2q^t)^d$. (If $t \geq 4$ then there could be t-vertex graphs H_i with different spans, and if one has span $s-1$ then $\mathbb{E}[Y] = (1+\Theta(1/d))\,c(d)_s(2q^t)^d$.)

Now we prove parts (b)(ii) and (b)(iii). Given d, let $\mathcal{S} = \mathcal{S}(d)$ be the set of subgraphs of Q^d ambient isomorphic to one of the graphs $H_1, \ldots, H_r$. List the members of $\mathcal{S}$ as $G_1, \ldots, G_N$ (where $N = N(d)$); and let A_i be the event that G_i is a component of Q_p. For distinct $i, j \in [N]$ let $i \sim j$ if either the vertex sets $V(G_i)$ and $V(G_j)$ intersect or there is an edge of Q^d between them. Observe that if $i \neq j$ and $i \not\sim j$ then the events A_i and A_j are independent, so we have a dependency graph. Now by (9) $\mathrm{Var}(Y) = \mathbb{E}[Y] + \Delta^+ - \Delta^-$, where Δ^+ and Δ^- are defined in (10) and (11) respectively. We next bound Δ^+ then Δ^-.

If $i \neq j$ and the vertex sets $V(G_i)$ and $V(G_j)$ intersect, then $\mathbb{P}(A_i \wedge A_j) = 0$, so in the sum for Δ^+ in (10) we need consider only the case where the two vertex sets $V(G_i)$ and $V(G_j)$ are disjoint but have connecting edges in Q^d (of which there can be at most $v(G_i)v(G_j)$). By (15), there is a constant α such that

$$\mathbb{P}(A_i) \leq \alpha\, q^{v(G_i)d} \quad \text{for each } i. \tag{16}$$

Thus, if $i \neq j$ then

$$\mathbb{P}(A_i \wedge A_j) \leq \mathbb{P}(A_i)\,\mathbb{P}(A_j)\,q^{-v(G_i)v(G_j)} \leq \mathbb{P}(A_i)\,\alpha q^{v(G_j)d} q^{-v(G_i)v(G_j)}. \tag{17}$$

For each integer k let $h(k)$ be the number of graphs H_i in the list with $v(H_i) = k$. Observe that for each set W of k vertices of Q^d, there are at most $h(k)$ graphs G_j with vertex set W, and there are no such graphs G_j if the induced subgraph $Q^d[W]$ of Q^d on W is not connected. For a given graph G_i of size t_1, the number of vertices v in Q^d adjacent to vertices in G_i is at most $t_1 d$. By Lemma 9 each vertex v is in at most $(ed)^{t_2-1}$ sets W of t_2 vertices such that the induced subgraph $Q^d[W]$ is connected. But each such vertex set W is the vertex set of at most $h(t_2)$ graphs G_j. Thus each vertex v could be in at most $(ed)^{t_2-1}h(t_2)$ graphs G_j of size t_2. In the sums below, t_1 and t_2 run over the possible sizes of the graphs G_i and G_j. From the definition (10), and using (17) and the last observation, we have

$$\begin{aligned}
\Delta^+ &= \sum_{t_1}\sum_{t_2}\sum_{i:v(G_i)=t_1}\sum_{j:j\sim i,v(G_j)=t_2} \mathbb{P}(A_i\wedge A_j)\\
&\leq \sum_{t_1}\sum_{t_2}\sum_{i:v(G_i)=t_1} \mathbb{P}(A_i)(t_1d)(ed)^{t_2-1}h(t_2)\,\alpha q^{t_2d}q^{-t_1t_2}\\
&\leq (1+o(1))\sum_{t_1}\sum_{i:v(G_i)=t_1} \mathbb{P}(A_i)(t_1d)(ed)^{t-1}h(t)\,\alpha q^{td}q^{-t_1t}\\
&= \mathbb{E}[Y]\,O(d^tq^{td}),
\end{aligned}$$

that is

$$\Delta^+ = \mathbb{E}[Y]\,O(d^tq^{td}). \tag{18}$$

Now consider Δ^-. By (16)

$$\sum_i \mathbb{P}(A_i)^2 \leq \sum_i \mathbb{P}(A_i)\cdot \alpha q^{td} = \mathbb{E}[Y]\cdot \alpha q^{td},$$

and, as for Δ^+ except without the factor $q^{-t_1t_2}$ (also including pairs i, j with $V(G_i)\cap V(G_j)\neq\emptyset$), we have

$$\sum_i\sum_{j\sim i}\mathbb{P}(A_i)\mathbb{P}(A_j) = \mathbb{E}[Y]\,O(d^tq^{td});$$

thus

$$\Delta^- = \mathbb{E}[Y]\,O(d^tq^{td}). \tag{19}$$

Now that we have (18) and (19), from (9) we have $\mathrm{Var}(Y) = \mathbb{E}[Y](1 + O(q^{td}d^t))$, and by Lemma 10 we have $d_{TV}(Y, \mathrm{Po}(\mathbb{E}[Y])) = O(d^tq^{td})$, as required.

(c) The contribution to $\mathbb{E}[Y]$ from graphs H_i with $v(H_i) > t$ is $O(dq^d)\cdot\mathbb{E}[Y]$, and similarly for $\mathbb{E}[\tilde{Y}]$. This gives equation (c)(i).

For parts (c) (ii) and (iii), we may argue as for parts (b) (ii) and (iii), but using Lemma 11 instead of Lemma 10. Assume that $t = 1$. Let G_i and A_i be as before, and let $t_i = v(G_i)$. Then $\tilde{Y} = \sum_{i=1}^r t_i\mathbb{I}_{A_i}$. Since the t_i are uniformly bounded, the quantity $\tilde{\Delta}^-$ (as in (14)) is at most a constant times the unweighted version Δ^-, and similarly for the second term in $\tilde{\Delta}^+$ (as in (13)). For the first term in $\tilde{\Delta}^+$, there is no contribution from the isolated vertices (graphs G_i with $t_i = 1$), so the term is $O(d(2q^2)^d)$: but $\mathbb{E}[Y]\geq \mu_1 = (2q)^d$, so the term is $O(\mathbb{E}[Y]\,dq^d)$. Hence by (18) and (19), both $\tilde{\Delta}^+$ and $\tilde{\Delta}^-$ are $O(\mathbb{E}[Y]\,dq^d)$. Equation (12) and Lemma 11 now complete the proof. □

Recall that X_t denotes the number of components of size t in Q_p, and that $\mu_t = \mathbb{E}[X_t]$. We noted earlier (more than once) that $\mu_1 = (2q)^d$, and the precise values of μ_2 and μ_3 are given in (3).

Lemma 13. *Let $0 < p < \frac{1}{2}$ and let $q = 1 - p$. Let $t \geq 1$ be fixed. Then*

$$\mu_t = (1 + O(\tfrac{1}{d}))\, \tfrac{t^{t-2}}{t!}\, (\tfrac{p}{q^2})^{t-1} d^{t-1} (2q^t)^d = \Theta(d^{t-1}(2q^t)^d).$$

Proof. If H_j is a spreading tree of size t, then $\operatorname{span}(H_j) = t-1$ and $e'(H_j) = 0$, and so by Lemma 12 (a),

$$\mathbb{E}[Y_j] = (p/q^2)^{t-1} 2^{d-t+1} \binom{d}{t-1} q^{td}, \tag{20}$$

where Y_j is the number of components of Q_p ambient-isomorphic to H_j. To calculate μ_t we need to sum $\mathbb{E}[Y_j]$ over all the ambient-isomorphism classes of t-vertex connected cube subgraphs H_j. We see from Lemma 12 (a) (and equation (20)) that if H_j is a spreading tree and $H_{j'}$ is not (so $\operatorname{span}(H_{j'}) \leq t - 2$) then $\mathbb{E}[Y_{j'}] = O(d^{-1})\, \mathbb{E}[Y_j]$. Thus the only significant terms are those corresponding to ambient-isomorphism classes of spreading trees, and by Lemma 8 (b) there are $2^{t-1}t^{t-3}$ such classes. Hence

$$\begin{aligned} \mu_t &= (1 + O(\tfrac{1}{d}))\, 2^{t-1} t^{t-3}\, 2^{d-t+1} \binom{d}{t-1} q^{td} (p/q^2)^{t-1} \\ &= (1 + O(\tfrac{1}{d}))\, d^{t-1} (2q^t)^d (t^{t-3}/(t-1)!) (p/q^2)^{t-1}, \end{aligned}$$

as required. □

Let us now complete the proof of Theorem 5 and then of Theorem 4.

Proof of Theorem 5. In part (a), the expected value is from Lemma 12 part (b)(i), and the variance is from Lemma 12 part (b)(ii); and the first half of part (c) (on Poisson approximation) is from Lemma 12 part (b)(iii).

Consider part (b). By a Chernoff bound (see for example inequality (2.9) and Remark 2.6 of [14]),

$$\begin{aligned} \mathbb{P}(|Y - \lambda| \geq \varepsilon (d\lambda)^{\frac{1}{2}}) &\leq \mathbb{P}(|\mathrm{Po}(\lambda) - \lambda| \geq \varepsilon (d\lambda)^{\frac{1}{2}}) + d_{TV}(Y, \mathrm{Po}(\lambda)) \\ &\leq 2e^{-\varepsilon^2 d/3} + O(d^t q^{td}), \end{aligned}$$

by the Poisson approximation bound. Thus $\mathbb{P}(|Y - \lambda| \geq \varepsilon (d\lambda)^{\frac{1}{2}}) = e^{-\Omega(d)}$, as required.

Finally, consider the second half of part (c). Since as $d \to \infty$ we have $\lambda \to \infty$, $d_{TV}(Y, \mathrm{Po}(\lambda)) \to 0$ and $\mathrm{Var}(Y) \sim \lambda$, it follows that Y^* is asymptotically standard normal – see the discussion before Theorem 3. This concludes the proof of Theorem 5. □

Proof of Theorem 4. The expression for the mean μ_t in part (a) is from Lemma 13. The rest follows directly from Theorem 5, with $H_1, \ldots, H_r$ listing a representative of each ambient-isomorphism class of t-vertex connected cube subgraphs. □

Remark 14. *In Theorem 5 it was natural to restrict our attention to connected graphs H_i with at most m_p vertices, and similarly in Theorem 4 it was natural to restrict our attention to components with at most m_p vertices. However, both these theorems are based on Lemma 12 in which there are no such restrictions. Thus in fact both these theorems hold without any such restrictions on the numbers of vertices, apart from in the two places in each theorem where we need the expected value λ to be large, namely the second half of part (b) and the second half of part (c) (in each of Theorems 4 and 5). We shall use this remark in the proof of Theorem 1.*

We have now proved Theorem 4, which says in particular that the distribution of the number X_t of components in Q_p of size t is close to the Poisson distribution $\mathrm{Po}(\mu_t)$. From what we have already proved, we can quickly give a first corresponding local limit result, showing that for suitable t we have $\mathbb{P}(X_t = \nu) \sim \mathbb{P}(\mathrm{Po}(\mu_t) = \nu)$ uniformly over the 'central range' of integers ν. Recall from Theorem 4 that $\mu_t = \Theta(d^{t-1}(2q^t)^d)$.

Proposition 15. *Let $0 < p < 1/2$ and let t be an integer with $m_p/3 < t \leq m_p$. Then for any fixed $c > 0$*

$$\sup_\nu \left|\mathbb{P}(X_t = \nu)/\mathbb{P}(\mathrm{Po}(\mu_t) = \nu) \ -1\right| = e^{-\Omega(d)}$$

where the sup *is over integers ν with $|\nu - \mu_t| \leq c\sqrt{\mu_t}$.*

Proof. Note first that $\mathbb{P}(\mathrm{Po}(\mu_t) = \nu) = \Theta(\mu_t^{-\frac{1}{2}})$, uniformly over integers ν with $|\nu - \mu_t| \leq c\sqrt{\mu_t}$. By Theorem 4 part (c), $d_{TV}(X_t, \mathrm{Po}(\mu_t)) = O(d^t q^{td})$, so $|\mathbb{P}(X_t = \nu) - \mathbb{P}(\mathrm{Po}(\mu_t) = \nu)| = O(d^t q^{td})$ uniformly over integers ν; and hence

$$|\mathbb{P}(X_t = \nu)/\mathbb{P}(\mathrm{Po}(\mu_t) = \nu) \ - 1| = O(d^t q^{td} \mu_t^{1/2}),$$

uniformly over integers ν with $|\nu - \mu_t| \leq c\sqrt{\mu_t}$. But $d^t q^{td} \mu_t^{1/2} = O(d^{3t/2}(2q^{3t})^{d/2}) = o(1)$ provided $2q^{3t} < 1$. Finally, we have $2q^{3t} < 1$ if $t > m_p/3$ (and indeed if $t = m_p/3$ unless $(2q)^{m_p} = 1$). □

4 The fragment $\mathcal{Z}$ has no large components

It will be straightforward to handle components of any fixed size $t > m_p$. We need to show also that wvhp there are no components in $\mathcal{Z}$ larger than some constant size (see Lemma 18 below). We use two preliminary lemmas. Given a spanning subgraph Q' of Q, call a vertex Q'-*good* if its degree in Q' is at least $dp/2$ and *bad* otherwise.

Lemma 16. *The probability that there is a pair of Q_p-good vertices at distance at most 3 in Q which are not joined by a path of length at most 7 in Q_p is $2^{-\Omega(d^2)}$.*

Proof. For a vertex v we let $\Gamma(v)$ denote its neighbourhood in Q_p. Fix vertices $u \neq v$ in Q at distance at most 3. Consider the case when $d_Q(u,v) = 3$ (the other cases are similar). For convenience, we consider Q^d as a graph on the power set of $[d]$. We may then suppose wlog that $u = \emptyset$ and $v = \{1,2,3\}$. Let A and B be sets of at least $dp/2$ neighbours in Q of u and v respectively.

For each $i \neq j$ in $\{4,\ldots,d\}$ with $\{i\} \in A$ and $v \cup \{j\} \in B$, there is a path

$$\{i\}, \{i,j\}, \{i,j,1\}, \{i,j,1,2\}, \{i,j,1,2,3\}, \{j,1,2,3\}$$

in Q, not using any edges incident with u or v. These form at least $(|A|-3)(|B|-4) \geq (pd/2-3)(pd/2-4)$ paths in Q of length 5 between A and B; and the paths are pairwise edge-disjoint since each edge identifies the pair (i,j). But the number of paths is at least $p^2d^2/5$ for d sufficiently large, and then

$$\begin{aligned}&\mathbb{P}(\text{no } u{-}v \text{ path of length 7 in } Q_p \mid \Gamma(u) = A, \Gamma(v) = B)\\ &\quad \leq (1-p^5)^{p^2d^2/5} \quad \leq \quad e^{-p^7d^2/5}.\end{aligned}$$

But $\mathbb{P}(\text{no } u{-}v \text{ path of length 7 in } Q_p \mid u,\ v \text{ are } Q_p\text{-good})$ is a weighted average of such probabilities, so

$$\begin{aligned}&\mathbb{P}((\text{no } u{-}v \text{ path of length 7 in } Q_p) \wedge (u,\ v \text{ are } Q_p\text{-good}))\\ &\quad \leq \mathbb{P}(\text{no } u{-}v \text{ path of length 7 in } Q_p \mid u,\ v \text{ are } Q_p\text{-good}) \quad \leq \quad e^{-p^7d^2/5}.\end{aligned}$$

Now, by a union bound, the probability that there is a pair of Q_p-good vertices at distance 3 in Q which are not joined by a path of length 7 in Q_p is at most

$$2^d d^3 e^{-p^7d^2/5} = 2^{-\Omega(d^2)}.$$

Similarly, with failure probability $2^{-\Omega(d^2)}$, if $d_Q(u,v) = 2$ then wvhp there is a $u{-}v$ path of length 6, and if $d_Q(u,v) = 1$ then wvhp there is a $u{-}v$ path of length 1 or 5. □

The second preliminary lemma is deterministic.

Lemma 17. *Let Q' be a (fixed) spanning subgraph of Q. Suppose that each vertex has a Q'-good neighbour in Q, and that for each pair u, v of Q'-good vertices at distance at most 3 in Q there is a $u-v$ path in Q'. Then for each pair u, v of Q'-good vertices there is a $u - v$ path in Q', and so all Q'-good vertices are in the same component of Q'.*

Proof. Let u, v be Q'-good vertices at distance $t > 3$ in Q. We must show that there is a $u - v$ path in Q'. Let $u = x_0, x_1, \ldots, x_{t-1}, x_t = v$ be a $u - v$ path in Q of length t. For each $i = 1, \ldots, t-1$, let y_i be a Q'-good neighbour in Q of x_i, where we choose $y_1 = u$ and $y_{t-1} = v$. Then since $d_Q(y_i, y_{i+1}) \leq 3$ for each $i = 1, \ldots, t-2$ there is a $y_i - y_{i+1}$ path in Q'. Hence there is a $u - v$ path in Q'. □

We may now deduce an upper bound for L_2 as required. When applying this upper bound, we shall later typically set $\gamma = 3$, so that failure probabilities will be negligibly small.

Lemma 18. *Let $0 < p < 1/2$ and let $\gamma > 0$. Then there is a constant N such that $\mathbb{P}(L_2 > N) = o(2^{-\gamma d})$.*

Proof. By a Chernoff bound and a union bound,

$$\begin{aligned} &\mathbb{P}(\text{some vertex has no } Q_p\text{-good neighbour in } Q) \\ \leq\ & 2^d\, \mathbb{P}(\text{Bin}(d,p) < pd/2)^d \ \leq\ 2^d\, e^{-(pd/8)\,d} \ =\ 2^{-\Omega(d^2)}. \end{aligned}$$

Let A be the event that all Q_p-good vertices in Q_p are in the same component. From the above bound and the last two lemmas

$$\mathbb{P}(\bar{A}) = 2^{-\Omega(d^2)}. \tag{21}$$

Now let $N = \lfloor \frac{16(1+\gamma)}{p} \rfloor$. If some component of the fragment has size at least $N + 1$, then also the giant component has size at least $N + 1$. Hence, if $L_2 > N$ and the event A holds then there is a component with size at least $N+1$ consisting entirely of bad vertices, and so in Q_p there is a subtree with $N + 1$ vertices each of which is bad. But consider any subtree T of Q with $N + 1$ vertices. Since Q is bipartite there is a set W of at least $(N + 1)/2$ vertices of T which forms a stable set in Q; and the probability that each vertex in such a set W is bad is

$$\mathbb{P}(\text{Bin}(d,p) < pd/2)^{|W|} \ \leq\ e^{-\frac{pd}{8}\frac{N+1}{2}} \ \leq\ e^{-(1+\gamma)d}$$

by a Chernoff bound and the inequality $(N + 1)pd/16 \geq (1 + \gamma)d$. Hence by Lemma 9 and a union bound, the probability that there is a subtree of Q_p with $N + 1$ vertices each of which is bad is at most

$$2^d(ed)^N e^{-(1+\gamma)d} = (ed)^N (2/e)^{(1+\gamma)d}\, 2^{-\gamma d} = o(2^{-\gamma d}).$$

Finally, using also (21), we have

$$\mathbb{P}(L_2 > N) \leq \mathbb{P}((L_2 > N) \wedge A) + \mathbb{P}(\bar{A}) = o(2^{-\gamma d}),$$

which completes the proof. □

5 Proofs of Theorems 1, 2 and 3

In this section, we complete the proofs of Theorems 1, 2 and 3.

5.1 Proof of Theorem 1

We have already noted that part (a) of Theorem 1 will follow directly from Theorem 3 and inequality (2).

Proof of Theorem 1 part (b). Let N be as in Lemma 18 for $\gamma = 3$, so that $\mathbb{P}(L_2 > N) = o(2^{-3d})$. Consider an integer t with $m_p < t \leq N$. By Markov's inequality and Lemma 13,

$$\mathbb{P}(X_t \geq 1) \leq \mathbb{E}[X_t] = O(d^{t-1}(2q^t)^d) = e^{-\Omega(d)},$$

where the last step follows since $2q^t < 1$. Hence wvhp the fragment $\mathcal{Z}$ has no component containing exactly t vertices. Putting these results together, we see that $L_2 \leq m_p$ wvhp; and that

$$\mathbb{E}[L_2] \leq m_p + N\,\mathbb{P}(m_p < L_2 \leq N) + 2^d\,\mathbb{P}(L_2 > N) = m_p + e^{-\Omega(d)}.$$

But $L_2 \geq m_p$ wvhp by Theorem 4 part (b) with $t = m_p$ (since $X_t \geq \mu_t/2$ wvhp). Hence $L_2 = m_p$ wvhp. It follows that $\mathbb{E}[L_2] \geq m_p - e^{-\Omega(d)}$, and thus $|\mathbb{E}[L_2] - m_p| = e^{-\Omega(d)}$.

Now consider $\mathrm{Var}(L_2)$, starting with an upper bound. We have

$$\mathbb{E}[(L_2 - m_p)^2 \mathbb{I}_{L_2 \leq N}] \leq N^2\,\mathbb{P}(L_2 \neq m_p) = e^{-\Omega(d)},$$

and

$$\mathbb{E}[(L_2 - m_p)^2 \mathbb{I}_{L_2 > N}] \leq 2^{2d}\,\mathbb{P}(L_2 > N) = e^{-\Omega(d)},$$

where $\mathbb{I}$ denotes an indicator variable (as earlier). Hence

$$\mathrm{Var}(L_2) \leq \mathbb{E}[(L_2 - m_p)^2] = e^{-\Omega(d)},$$

which is an upper bound as required. Finally we show that

$$\mathrm{Var}(L_2) \gg q^d. \tag{22}$$

We start by noting a simple general lower bound on variance. Let the random variable L be integer-valued; let k be an integer and let $x > 0$; and suppose that both $\mathbb{P}(L \leq k)$ and $\mathbb{P}(L \geq k+1)$ are at least x. Then $\mathrm{Var}(L) \geq x(1-x)$.

We know that $L_2 = m_p$ wvhp. Recall from Remark 14 that in Theorem 4 both part (a) and the first half of part (c) hold for any given positive integer t (not just for $t \leq m_p$). Let $t = m_p + 1$ (≥ 2). By the first half of part (c) of Theorem 4

$$\mathbb{P}(L_2 \geq t) \geq \mathbb{P}(X_t \geq 1) = \mathbb{P}(\mathrm{Po}(\mu_t) \geq 1) + O(d^t q^{td}).$$

But since $\mu_t = o(1)$ and $2q^{m_p} \geq 1$, by part (a) of Theorem 4

$$\mathbb{P}(\mathrm{Po}(\mu_t) \geq 1) = (1 + o(1))\, \mu_t = \Theta(d^{t-1}(2q^t)^d) \gg d^t q^{td}.$$

Thus

$$\mathbb{P}(L_2 \geq m_p + 1) \geq (1 + o(1))\, \mu_{m_p+1} \gg q^d.$$

Now (22) follows from the general lower bound on variance given above, and this completes the proof of the theorem. □

5.2 Proof of Theorem 2

We prove the two parts of the theorem separately. We denote the r-ball $B_r(\mathbf{0})$ centred on the vertex $\mathbf{0}$ by B_r for short.

Proof of Theorem 2 part (a). Let $s = m_p + 1$ and let $V = V(Q)$. Recall from Theorem 1(b) that $L_2 \leq m_p$ wvhp. We use $\deg(v)$ for the degree of a vertex v in Q_p. Also, for $v \in V$ and $W \subseteq V$, let $e(v, W)$ be the number of edges in Q_p between v and W. For each subset $S \subseteq V$ with $|S| = s$ we have

$$\begin{aligned}
\mathbb{P}((S \subseteq V(\mathcal{Z})) \wedge (L_2 \leq m_p)) &\leq \mathbb{P}(\deg(v) \leq m_p - 1 \ \ \forall v \in S) \\
&\leq \mathbb{P}(e(v, V \setminus S) \leq m_p - 1 \ \ \forall v \in S) \\
&= (\mathbb{P}(\mathrm{Bin}(d-s, p) \leq m_p - 1))^s \\
&\leq \left(\binom{d-s}{m_p - 1} q^{d-s-(m_p-1)} \right)^s \\
&\leq \left(d^{m_p-1} q^{d-2m_p} \right)^s \ \leq \ (d/q^2)^{m_p s} q^{sd}.
\end{aligned}$$

Hence, for any $r > 0$,

$$\begin{aligned}
\mathbb{P}(|V(\mathcal{Z}) &\cap B_r(u)| \geq s \text{ for some } u \in V) \\
&= \mathbb{P}\Big(\bigcup_{u\in V}\bigcup_{S\subseteq B_r(u),|S|=s}(S \subseteq V(\mathcal{Z}))\Big) \\
&\leq \mathbb{P}\Big(\bigcup_{u\in V}\bigcup_{S\subseteq B_r(u),|S|=s}(S \subseteq V(\mathcal{Z})) \wedge (L_2 \leq m_p)\Big) + \mathbb{P}(L_2 > m_p) \\
&\leq 2^d\binom{|B_r|}{s}(d/q^2)^{m_p s}q^{sd} + \mathbb{P}(L_2 > m_p) \\
&\leq (d/q^2)^{m_p s}\,|B_r|^s\,(2q^s)^d + \mathbb{P}(L_2 > m_p). \qquad (23)
\end{aligned}$$

Since $s > m_p$ and $q < 1/2$, we have $2q^s < 1$ and $1 > \log_2(1/q) - 1/s > 0$. Let η_1 be the unique $x \in (0, \frac{1}{2})$ such that $h(x) = \log_2(1/q) - 1/s$. Let $0 < \eta < \eta_1$. Then $h(\eta) < \log_2(1/q) - 1/s$, and so

$$2\,(2^{h(\eta)}q)^s < 1.$$

Set $r = \eta d$. Then $|B_r| = 2^{h(\eta)d+o(d)}$ by standard estimates. Thus, by the last inequality,

$$|B_r|^s\,(2q^s)^d = (2\,(2^{h(\eta)}q)^s)^d 2^{o(d)} = 2^{-\Omega(d)}.$$

Hence, by (23) and using $\mathbb{P}(L_2 > m_p) = 2^{-\Omega(d)}$, we have

$$\mathbb{P}(|V(\mathcal{Z}) \cap B_r(u)| \geq s \text{ for some } u \in V) = 2^{-\Omega(d)}$$

as required. □

Consider η_1 in the above proof: it can be shown that if $\eta > \eta_1$ then the expected number of ηd-balls containing more than m_p vertices in $\mathcal{Z}$ tends to ∞ as $d \to \infty$.

Proof of Theorem 2 part (b). Recall that η^* is defined immediately before Theorem 2. We may assume that $\varepsilon > 0$ is sufficiently small that $\eta^* - \varepsilon > 0$ and $\eta^* + \varepsilon < \frac{1}{2}$. Given $0 < \eta \leq \frac{1}{2}$, we have $|B_{\eta d}| = 2^{h(\eta)d+o(d)}$, as we noted above. Also, $2^{-h(\eta^*)} = q$. Hence, by Theorem 1 (a), wvhp

$$\begin{aligned}
|B_{(\eta^*-\varepsilon)d}| \cdot Z &\leq 2^{h(\eta^*-\varepsilon)d+o(d)} \cdot 2\mu_1 \\
&= 2^{(h(\eta^*-\varepsilon)-h(\eta^*)+o(1))d} \cdot 2^d \\
&= 2^{-\Omega(d)} \cdot 2^d,
\end{aligned}$$

As the number of vertices within distance at most $(\eta^* - \varepsilon)d$ of $\mathcal{Z}$ is at most $|B_{(\eta^*-\varepsilon)d}| \cdot Z$, this proves the first half of part (b).

For the second half, let B' denote $B_{(\eta^*+\varepsilon)d}$. By the definition of η^*, and recalling that $h(\eta)$ is strictly increasing on $(0, \frac{1}{2})$, we have $q^d|B'| = e^{\Omega(d)}$. Since Q^d is bipartite, there is a stable subset B'' of B' with $|B''| \geq \frac{1}{2}|B'|$; and the probability that no vertex of $\mathcal{Z}$ is in B' is at most the probability that no vertex in B'' is isolated, which equals

$$(1-q^d)^{|B''|} \leq \exp(-\tfrac{1}{2}q^d|B'|) = \exp(-e^{\Omega(d)}).$$

This bound refers to the ball B' centred at $\mathbf{0}$, and indeed to any fixed centre vertex. Taking a union bound over all 2^d possible centre vertices shows that the probability that some vertex is not within distance $(\eta^* + \varepsilon)d$ of $\mathcal{Z}$ is $\exp(-e^{\Omega(d)})$, and thus completes the proof. □

In the last part of the proof above, the number of isolated vertices in B'' has distribution $\mathrm{Bin}(|B''|, q^d)$, with mean at least $\frac{1}{2}|B'|q^d = e^{\Omega(d)}$. Hence, by a Chernoff bound, the probability that there are at most $\frac{1}{4}|B'|q^d$ isolated vertices in the ball B' is at most $e^{-e^{\Omega(d)}}$; and so, by a union bound, wvhp each $(\eta^* + \varepsilon)d$-ball contains exponentially many isolated vertices.

5.3 Proof of Theorem 3

By Lemma 18 we may choose a fixed integer $N \geq 2$ such that $\mathbb{P}(L_2 > N) \leq 2^{-3d}$.

Proof of Theorem 3 part (a). Note that $Z \leq 2^d$ and so

$$Z \leq \sum_{t=1}^{N} X_t + 2^d \mathbb{I}_{L_2 > N}.$$

By Lemma 13, for each $2 \leq t \leq N$, $\mu_t = \mathbb{E}[X_t] = \Theta(d^{t-1}(2q^t)^d)$, so μ_t is $O(d(2q^2)^d)$. Hence,

$$\begin{aligned} \mathbb{E}[Z] &\leq \sum_{t=1}^{N} \mu_t + 2^d\, \mathbb{P}(L_2 > N) \\ &\leq \mu_1 + O(d(2q^2)^d) + 2^{-2d} \;=\; (1 + O(dq^d))\mu_1. \end{aligned}$$

Also, of course, $\mu_1 + \mu_2 \leq \mathbb{E}[X] \leq \mathbb{E}[Z]$, which completes the proof for the expected values.

Now consider variances. Let $X_{\leq N} = \sum_{t=1}^{N} X_t$ be the total number of components in Q_p of size at most N; and similarly let $Z_{\leq N} = \sum_{t=1}^{N} tX_t$ be the total size of the components of size at most N. Then

$$\mathrm{Var}(Y) - \mathrm{Var}(Y_{\leq N}) \leq \mathbb{E}[Y^2 - Y_{\leq N}^2] \leq 2^{2d}\mathbb{P}(L_2 > N) \leq 2^{-d},$$

and

$$\mathrm{Var}(Y_{\le N}) - \mathrm{Var}(Y) \le \mathbb{E}[Y + Y_{\le N}]\, \mathbb{E}[Y - Y_{\le N}] \le 2\mathbb{E}[Y] 2^d \mathbb{P}(L_2 > N) = o(2^{-d}),$$

and so

$$|\mathrm{Var}(Y) - \mathrm{Var}(Y_{\le N})| = O(2^{-d}).$$

Hence by Lemma 12(b) and (c), with $H_1, \ldots, H_r$ listing a representative of each ambient-isomorphism class of connected cube subgraphs with at most N vertices, we see that $\mathrm{Var}(Y) = (1 + O(dq^d))\mu_1$, as required. □

Proof of Theorem 3 part (b). Let us show first that

$$d_{TV}(Y, \mathrm{Po}(\lambda)) = O(dq^d). \tag{24}$$

Write $\lambda_{\le N}$ for $\mathbb{E}[Y_{\le N}]$. Now $d_{TV}(Y, \mathrm{Po}(\lambda))$ is at most

$$d_{TV}(Y, Y_{\le N}) + d_{TV}(Y_{\le N}, \mathrm{Po}(\lambda_{\le N})) + d_{TV}(\mathrm{Po}(\lambda_{\le N}), \mathrm{Po}(\lambda)).$$

We consider the three terms in the sum in order. Firstly, we have

$$d_{TV}(Y, Y_{\le N}) \le \mathbb{P}(Y \ne Y_{\le N}) = \mathbb{P}(L_2 > N) \le 2^{-3d} = o(q^d).$$

Secondly, by Lemma 12(b) and (c) (with H_i as above)

$$d_{TV}(Y_{\le N}, \mathrm{Po}(\lambda_{\le N})) = O(dq^d).$$

Thirdly, for $\mu, \delta > 0$ the sum of independent $\mathrm{Po}(\mu)$ and $\mathrm{Po}(\delta)$ random variables has distribution $\mathrm{Po}(\mu + \delta)$; and so

$$d_{TV}(\mathrm{Po}(\mu), \mathrm{Po}(\mu + \delta)) \le \mathbb{P}(\mathrm{Po}(\delta) \ne 0) = 1 - e^{-\delta} \le \delta.$$

Thus

$$d_{TV}(\mathrm{Po}(\lambda_{\le N}), \mathrm{Po}(\lambda)) \le \lambda - \lambda_{\le N} \le 2^d\, \mathbb{P}(L_2 > N) \le 2^{-2d} = o(q^d).$$

Putting these inequalities together we obtain (24).

Finally, since also $\mathrm{Var}(Z) \sim \lambda \to \infty$ as $d \to \infty$, it follows from (24) that Z^* is asymptotically standard normal (see the discussion immediately before Theorem 3). This completes the proof of part (b), and thus of Theorem 3 (and thus also of Theorem 1). □

6 Joint distributions: proof of Theorem 6

In this section we prove Theorem 6 on the joint distribution of the numbers of components of different types in the fragment. We start by presenting a general lemma on approximating a joint distribution by a product of Poisson distributions. As in Subsection 2.2, let $(A_i : i \in I)$ be a family of events with a dependency graph L, and write $i \sim j$ if i and j are adjacent in L. For each i, let $\pi_i = \mathbb{P}(A_i)$ and let $\mathbb{I}_i$ be the indicator function of A_i. Now we let I be partitioned into $I_1 \cup \cdots \cup I_r$ for some $r \geq 1$. For each $j \in [r]$, let $X_j = \sum_{i \in I_j} \mathbb{I}_{A_i}$ and let $\lambda_j = \mathbb{E}[X_j]$. The following lemma is essentially a special case of Theorem 10.K of Barbour, Holst and Janson [2] when all means $\lambda_j \to \infty$. Sums and products over j or j' always mean over j or j' in $[r]$.

Lemma 19. *With notation as above, assume that each $\lambda_j \to \infty$ as $d \to \infty$. Then for d sufficiently large*

$$
\begin{aligned}
& d_{TV}(\mathcal{L}(X_1, \ldots, X_r), \prod_j \mathrm{Po}(\lambda_j)) \\
\leq \; & \sum_j \frac{\ln(\lambda_j)}{\lambda_j} \sum_{i \in I_j} \pi_i^2 + \sum_j \sum_{j'} \frac{\ln(\lambda_j \lambda_{j'})}{\sqrt{\lambda_j \lambda_{j'}}} \sum_{i \in I_j} \sum_{i' \in I_{j'}} \mathbb{I}_{i \sim i'} (\mathbb{P}(A_i \wedge A_{i'}) + \pi_i \pi_{i'}).
\end{aligned}
$$

Proof of Theorem 6. As earlier, given d let $\mathcal{S} = \mathcal{S}(d)$ be the set of subgraphs of Q^d ambient isomorphic to one of the graphs $H_1, \ldots, H_r$. List the members of $\mathcal{S}$ as $G_1, \ldots, G_N$; and let A_i be the event that G_i is a component of Q_p. We let i, i' run over $[N]$ and j, j' run over $[r]$. For distinct i, i' let $i \sim i'$ if either the vertex sets $V(G_i)$ and $V(G_{i'})$ intersect or there is an edge of Q^d between them; and note that this gives a dependency graph L. For each j, let $I_j = \{i : G_i \text{ is ambient isomorphic to } H_j\}$.

Now we can apply Lemma 19. We must bound the two terms in the lemma. First, by (16), there is a constant α such that, for each j,

$$\sum_{i \in I_j} \pi_i^2 \leq \sum_{i \in I_j} \pi_i \cdot \alpha q^{v(G_i)d} = \lambda_j \cdot \alpha q^{v(H_j)d}.$$

Hence

$$\sum_j \frac{\ln(\lambda_j)}{\lambda_j} \sum_{i \in I_j} \pi_i^2 \leq \alpha \sum_j \ln(\lambda_j)\, q^{v(H_j)d} = O(dq^{t^* d}) \tag{25}$$

since $\ln(\lambda_j) = O(d)$ uniformly over j.

For the second term, let $j, j' \in [r]$ (not necessarily distinct). For $i \in I_j$ and $i' \in I_{j'}$, as in (17) we have

$$\mathbb{P}(A_i \wedge A_{i'}) \leq \pi_i \pi_{i'}\, q^{-v(H_j)v(H_{j'})} \leq \pi_i\, \alpha q^{v(H_{j'})d} q^{-v(H_j)v(H_{j'})}.$$

Hence, arguing as in the proof of (18),

$$\begin{aligned}\beta(j,j') &:= \sum_{i\in I_j}\sum_{i'\in I_{j'}} \mathbb{I}_{i\sim i'}\big(\mathbb{P}(A_i \wedge A_{i'}) + \pi_i\pi_{i'}\big)\\ &\le \sum_{i\in I_j} \pi_i \cdot \alpha q^{v(H_{j'})d}\big(q^{-v(H_j)v(H_{j'})}+1\big)v(H_j)d\,(ed)^{v(H_{j'})-1}\\ &= \lambda_j \cdot O\big(d^{v(H_{j'})}q^{v(H_{j'})d}\big) \;=\; \lambda_j \cdot O\big((dq^d)^{v(H_{j'})}\big).\end{aligned}$$

Similarly, swapping j and j', we have

$$\beta(j,j') \le \lambda_{j'} \cdot O\big((dq^d)^{v(H_j)}\big);$$

and so

$$\beta(j,j') \le \sqrt{\lambda_j\lambda_{j'}} \cdot O\big((dq^d)^{t^*_{jj'}}\big),$$

where $t^*_{jj'} = \min\{v(H_j), v(H_{j'})\}$. Hence,

$$\frac{\ln(\lambda_j\lambda_{j'})}{\sqrt{\lambda_j\lambda_{j'}}}\,\beta(j,j') = O(d)\,O((dq^d)^{t^*_{jj'}}) = O\big(d^{t^*+1}q^{t^*d}\big).$$

So, summing over the bounded number of choices of j and j', we obtain

$$\begin{aligned}&\sum_j\sum_{j'}\frac{\ln(\lambda_j\lambda_{j'})}{\sqrt{\lambda_j\lambda_{j'}}}\sum_{i\in I_j}\sum_{i'\in I_{j'}} \mathbb{I}_{i\sim i'}(\mathbb{P}(A_i\wedge A_{i'}) + \pi_i\pi_{i'})\\ &= \sum_j\sum_{j'}\frac{\ln(\lambda_j\lambda_{j'})}{\sqrt{\lambda_j\lambda_{j'}}}\,\beta(j,j') \;=\; O\big(d^{t^*+1}q^{t^*d}\big).\end{aligned}$$

This result, together with (25) lets us use Lemma 19 to complete the proof of Theorem 6. □

7 Concluding remarks

In Theorems 1 to 6 we have seen quite a full picture of the rich component structure of the random graph $Q_p = Q_p^d$, for fixed p with $0 < p < \frac{1}{2}$. In particular, given an integer t with $1 \le t \le m_p$, by Theorem 4 the number X_t of components in Q_p of size t, with mean μ_t, has close to the Poisson distribution $\mathrm{Po}(\mu_t)$, and thus the standardised version X_t^* has close to the standard normal distribution. In Proposition 15 we gave a partial corresponding local limit result for convergence to the Poisson distribution: it would be interesting to learn more on such local behaviour.

It would also be interesting to consider the component structure in the case when p is not fixed in $(0, \frac{1}{2})$, but $p = p(d)$ decreases suitably slowly to 0 as $d \to \infty$. (Thanks to Remco van der Hofstadt for asking about this case.)

Acknowledgement: We would like to thank the referees for their careful reading and very helpful comments.

References

[1] M. Ajtai, J. Komlós and E. Szemerédi, Largest random component of a k-cube, *Combinatorica* **2** (1982), 1-7.

[2] A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Clarendon Press, Oxford 1992.

[3] A.J. Bernstein, Maximally connected arrays on the n-cube, *SIAM J.Appl. Math.* **15** (1967), 1485-1489.

[4] B. Bollobás, *Random Graphs,* 2nd ed., Cambridge University Press, 2001.

[5] B. Bollobás, Complete matchings in random subgraphs of the cube, *Random Structures and Algorithms* **1** (1990), 95-104.

[6] B. Bollobás, Y. Kohayakawa and T. Łuczak, The evolution of random subgraphs of the cube, *Random Structures and Algorithms* **3** (1992), 55-90.

[7] B. Bollobás, Y. Kohayakawa and T. Łuczak, On the diameter and radius of random subgraphs of the cube, *Random Structures and Algorithms* **5** (1994), 627-648.

[8] B. Bollobás, Y. Kohayakawa and T. Łuczak, Connectivity properties of random subgraphs of the cube, *Random Structures and Algorithms* **6** (1995), 221-230.

[9] C. Borgs, J. Chayes, R. van der Hofstadt, G. Slade and J. Spencer, Random subgraphs of finite graphs: III the phase transition for the n-cube. *Combinatorica* **26** (2006), 395 – 410.

[10] Yu. D. Burtin, On the probability of connectedness of a random subgraph of the n-cube, *Problemy Pered. Inf.* (Problems of Information Transmission) **13** (1977), 90 – 95 (in Russian).

[11] P. Erdős and J. Spencer, Evolution of the n-cube, *Comp. and Math with Appl.* **5** (1979), 33-39.

[12] C. D. Godsil, Matchings and walks in graphs, *J. Graph Th.* **5** (1981) 285 – 297.

[13] T. Hulshof and A. Nachmias, Slightly subcritical hypercube percolation, *Random Structures and Algorithms* **56** (2020) 557 – 593.

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

[15] D. E. Knuth, *The Art of Computer Programming, Vol. 1 Fundamental Algorithms,* 3rd ed., Addison-Wesley, 1997.

[16] C. McDiarmid, A. Scott and P. Withers, Uniform multicommodity flow through the hypercube with random edge-capacities. *Random Structures and Algorithms* **50** (2017) 437 – 463 (published online 7 November 2016 in Wiley Online Library).

[17] K. Weber, On components of random graphs in the n-cube. *Elektron. Inf. verarb. Kybern. EIK* **22** (12) (1986), 601 – 613.

[18] K. Weber, Poisson Convergence in the n-Cube, *Math. Nachr.* **131** (1987) 49 – 57.

[19] K. Weber, On components of random subgraphs of the n-cube. In *Random Graphs, Volume 2*, A. Frieze and T. Łuczak eds, pages 263 – 278, Wiley, 1992.