ADVANCES IN APPLIED MATHEMATICS 11, 36–48 (1990)

The Distribution of Clusters in Random Graphs

RICHARD ARRATIA*

Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113

AND

ERIC S. LANDER†, ‡

Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142

Given a random graph, we investigate the occurrence of subgraphs especially rich in edges. Specifically, given $a \in [0, 1]$, a set of k points in a graph G is defined to be an a-cluster of cardinality k if the induced subgraph contains at least $a\binom{k}{2}$ edges, so that in the extreme case $a = 1$, an a-cluster is the same as a clique. We let $G = G(n, p)$ be a random graph on n vertices with edges chosen independently with probability p. Let W denote the number of a-clusters of cardinality k in G, where k and n tend to infinity so that the expected number λ of a-clusters of cardinality k does not grow or decay too rapidly. We prove that W is asymptotically distributed as Z_λ, whose distribution is Poisson with mean λ, which is the same result that Bollobás and Erdös have proved for cliques. In contrast to the situation for cliques ($a = 1$) however, for all $a < 1$ the second moment of W blows up, i.e., the expected number of neighbors of a given cluster tends to infinity. Nevertheless, the probability that there exists at least one pair of neighboring clusters tends to zero, and a Poisson approximation for W is valid.

1. INTRODUCTION

Given a random graph G, we investigate here the occurrence of subgraphs of G that are especially rich in edges. Specifically, let $G = G(n, p)$ denote a random graph with vertex set V of cardinality n and edge set E

*Supported by NIH Grant GM 36230 and NSF Grant DMS 8815106.

†Harvard University, Cambridge, MA 02138.

‡Supported by NSF Grant DCB 8611317, and a grant from the System Development Foundation.

0196-8858/90 $7.50

consisting of edges chosen independently with probability p from the set of all possible edges. For $a \in [0, 1]$, define an a-cluster of cardinality k to be a subset $\alpha \subset V$ or cardinality k such that the subgraph induced on α contains at least $a\binom{k}{2}$ edges. For $a > p$, what is the probability that G will contain at least one a-cluster of cardinality k? What is the distribution of the number W of such clusters in G?

The expected value λ of W is straightforward to compute: The probability p_α that a given k-set α is an a-cluster can be calculated from the binomial distribution, and does not vary with α. Summing over all k-sets, we have $EW = \binom{n}{k} p_\alpha$. For fixed p and a, suppose that k and n tend to infinity in such a way that $\lambda = EW$ is bounded away from zero and infinity. It seems reasonable to hope that W will tend to a Poisson random variable with mean λ. Theorem 5 below states that this is indeed the case and provides an upper bound on the rate of convergence.

For $a = 1$, an a-cluster is the same as a clique. For the special case of cliques, the size of the largest clique is found in Bollobás and Erdös [5], and the Poisson convergence was proven by Bollobás [4] by means of an approach developed by Stein [11] and Chen [6]. Define two cliques to be neighbors if they share at least two vertices. The basic method then simply requires showing that the expected number m of cliques which neighbor any given clique tends to 0 as k and $n \to \infty$.

Contrary to our initial expectations, the corresponding statement for a-clusters is false if $a \neq 1$. For all $0 < p < a < 1$, the expected number m of neighboring clusters tends to infinity, so that $EW^2 \to \infty$. Nevertheless, we show how to modify the analysis of the situation so that the Chen–Stein method can be used to prove Poisson convergence for W. Roughly speaking, this means that the probability of having at least one pair of neighboring clusters tends to zero, but in the rare event when neighboring clusters do occur, they may occur in profusion.

This work was motivated by a problem which arose in molecular biology. One of the most common and most powerful ways to infer the function of a newly discovered protein is to discover a striking similarity between its amino acid sequence (proteins can be thought of as long words written in a 20-letter alphabet of amino acids) and that of a previously sequenced protein of known function. See for example the book by Doolittle [7]. We reasoned that more distant functional and evolutionary relationships would give rise to sets of proteins in which no pair displayed a dramatic similarity, but in which many pairs showed mild similarities. In a graph whose vertices were all proteins and in which edges denoted mild similarities, such sets would be clusters. The present work arose in the attempt to evaluate the statistical significance of such clusters of proteins, before attempting to search for biological significance.

The paper is organized as follows. Section 2 provides necessary background for proving Poisson convergence. After pointing out that a given a-cluster can have an expected number of neighboring clusters which tends to infinity, Section 3 overcomes this difficulty to prove the main result. Section 4 discusses variations on the theme. Although unrelated to the issue of Poisson convergence, Section 5 show that for $a < 1$, the expected number of neighboring clusters always tends to infinity, and characterizes the ways in which this happens.

2. Background on Poisson Convergence

A beautiful and powerful method for proving convergence to a Poisson distribution was given by Chen [6], who adapted the differential method of Stein, developed for the normal distribution [11]. Our approach is drawn from a recent paper of Arratia, Goldstein, and Gordon [1] which describes, generalizes, and applies these methods.

Let I be an arbitrary index set. For each $\alpha \in I$, suppose we have a Bernoulli random variable X_α with $p_\alpha \equiv P(X_\alpha = 1) = 1 - P(X_\alpha = 0) > 0$, and suppose that we have a "neighborhood of dependence" $B(\alpha) \subset I$ such that X_α is independent of all of $(X_\beta)_{\beta \in I - B(\alpha)}$. When $X_\alpha = 1$, we speak of an event occurring at α. Let

$$W \equiv \sum_{\alpha \in I} X_\alpha, \qquad \lambda \equiv EW = \sum_{\alpha \in I} p_\alpha.$$

Let Z_λ denote a Poisson random variable with mean λ, i.e., let $P(Z_\lambda = k) = e^{-\lambda}\lambda^k/k!$, for $k = 0, 1, 2, \ldots$. We denote the variation distance between the distributions of W and Z_λ by

$$|L(W) - L(Z_\lambda)| \equiv 2 \max_{A \subset \{0,1,2,\ldots\}} |P(W \in A) - P(Z_\lambda \in A)|,$$

which also may be viewed as the twice the minimal value of $P(W \neq Z_\lambda)$, over all realizations of both random variables on the same probability space. Define

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B(\alpha)} p_\alpha p_\beta$$

$$b_2 = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B(\alpha)} p_\alpha p_{\beta|\alpha}, \qquad \text{where } p_{\beta|\alpha} \equiv E(X_\beta | X_\alpha = 1).$$

Arratia, Goldstein, and Gordon [1] prove the following results (in fact, they give somewhat stronger statements allowing weak dependence of X_α

on X_β for $\beta \notin B(\alpha)$):

THEOREM 1. *With notation as above,*

$$|L(W) - L(Z_\lambda)| < 2(b_1 + b_2)(1 - e^{-\lambda})/\lambda,$$
$$|P(W = 0) - e^{-\lambda}| < (b_1 + b_2)(1 - e^{-\lambda})/\lambda.$$

Furthermore, they show that when the Poisson approximation can be established by the above machinery, the entire dependent family of events can be viewed as a small perturbation of a family of independent events having the same individual probabilities:

THEOREM 2. *Consider the given dependent events as a process with values in* $\{0,1\}^I$; *i.e., consider* $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$. *Let* $\mathbf{Y} \equiv (Y_\alpha)_{\alpha \in I}$ *be a process with independent components and the same marginals:* $\forall \alpha$, $EX_\alpha = EY_\alpha$. *It is possible to realize* $\mathbf{X}$ *and* $\mathbf{Y}$ *on a single probability space so that*

$$P(\mathbf{X} \neq \mathbf{Y}) \leq 2\Big(b_1 + b_2 + \sum p_\alpha^2\Big).$$

In many applications, there is sufficient symmetry that the following three quantities do not vary with the choice of $\alpha \in I$: p_α, $|B(\alpha)|$, and $m \equiv \sum_{\alpha \neq \beta \in B(\alpha)} p_{\beta|\alpha}$. In these situations we have $b_1 = (EW)|B(\alpha)|/|I|$ and $b_2 = (EW)m$. Given an infinite series of examples in which $\lambda \equiv EW$ stays bounded away from infinity, Poisson convergence follows at once provided only that $|B(\alpha)|/|I| \to 0$ and $m \to 0$. Since the first condition is usually trivial to verify, the proof of Poisson convergence reduces to calculating m, the expected number of events neighboring a given event.

Consider the number of 1-clusters (i.e., cliques) of size k in a random graph $G = G(n, p)$. The Poisson approximation here can also be established using inclusion–exclusion; see Spencer [10, pp. 21–22]. The following argument, using the Chen–Stein method, appears in Bollobás [4]. Let I be the set of all k-sets of vertices of G, let $B(\alpha)$ be the set of all k-sets intersecting α in at least two points and let X_α be the indicator of the event that the k-set α is a clique. Setting $W = \sum_{\alpha \in I} X_\alpha$, we have $EW = \binom{n}{k} p^{\binom{k}{2}}$. If k and n tend to infinity so that $\lambda \equiv EW$ satisfies $\lambda^{1/k} \to 1$, we can easily show that W converges to a Poisson random variable with mean λ by estimating the quantities $|B(\alpha)|/|I|$ and m. The condition on the growth of EW implies that

$$n \sim \frac{k}{e} p^{-(k-1)/2}$$

and hence

$$k \sim 2\log_{1/p}(n).$$

Since $|B(\alpha)|/|I|$ is the probability that two random k-sets intersect in at least two elements, it is at most $\binom{n}{k-2}\binom{k}{2}/\binom{n}{k} = O(k^4/n^2) \to 0$. The expected number m of cliques which intersect a given clique α can be decomposed as $m = \sum_{1<j<k} m_j$, where m_j is the expected number of cliques which meet α in exactly j points. Now,

$$m_j = \binom{k}{j}\binom{n-k}{k-j} p^{\binom{k}{2}-\binom{j}{2}}.$$

Thus,

$$m_{k-1} = k(n-k)p^{k-1} = O(k^3/n),$$
$$m_{k-2} \sim (1/2)k^2n^2p^{2k-3} = O(k^6/n^2),$$
$$m_2 = (\lambda/p)\binom{k}{2}\binom{n-k}{k-2}\Big/\binom{n}{k} = \lambda O(k^4/n^2).$$

By examining the ratio of successive terms in the sequence m_2, $m_3, \ldots, m_{k-2}, m_{k-1}$, we see that the m_2 and m_{k-2} dominate all the terms between them. Thus, Theorem 1 implies that $|L(W) - L(Z_\lambda)| = O((\log n)^3/n) = O(k^3/n)$.

3. Main Results

Unfortunately, the proof above does not extend automatically to a-clusters: for certain values of $p < a < 1$, the expected number m of clusters neighboring a given cluster tends to infinity. To see this, we recall the large deviation theory of the binomial distribution. See, e.g., Bahadur [3] or Arratia and Gordon [2] for a background reference.

Let $C_1, \ldots, C_t$ be independent, identically distributed Bernoulli variables with $P(C_i = 1) = p$. We have, for $p < a < 1$, with

$$r \equiv \frac{p}{a}\frac{1-a}{1-p} \in (0,1),$$

$$P(C_1 + \cdots + C_t \geq at) \sim \frac{1}{1-r}(2\pi a(1-a)t)^{-1/2}\exp(-tH(\lceil at \rceil/t, p)) \tag{1}$$

and for $j = 0, 1, 2, \ldots,$

$$P(C_1 + \cdots + C_t = \lceil at \rceil + j \mid C_1 + \cdots + C_t \geq at) \to r^j(1 - r),$$

where $H(a, p)$ is the relative entropy between a p-coin and an a-coin,

$$H(a, p) \equiv (a)\log(a/p) + (1 - a)\log((1 - a)/(1 - p)).$$

Let $G = G(n, p)$ be a random graph and fix $a \in (p, 1)$. As before, let I be the set of all k-sets of vertices of G, let $B(\alpha)$ be the set of all k-sets intersecting α in at least two points, and let X_α be the indicator of the event that the k-set α is an a-cluster. Setting $W = \sum_{\alpha \in I} X_\alpha$, we have $\lambda \equiv EW = \binom{n}{k} P(C_1 + \cdots + C_t \geq at)$ with $t = \binom{k}{2}$. If k and n tend to infinity so that $\lambda^{1/k} \to 1$, then

$$n \sim \frac{k}{e} \exp\left(\frac{k - 1}{2} H(a, p)\right), \tag{2}$$

and consequently

$$\log n \sim (k/2)H(a, p). \tag{3}$$

We can compute a lower bound on the expected m of a-clusters neighboring a given a-cluster α as follows. For some subset fixed $\gamma \subset \alpha$ of cardinality j, suppose that the subgraph induced on γ is a clique. The probability that this will occur is at least $a^{\binom{j}{2}}$. Now, if $j = \lceil (\sqrt{a})k + 1 \rceil$, then γ contains $\binom{j}{2} > a\binom{k}{2}$ edges and therefore every k-set $\beta \in I$ with $\alpha \cap \beta = \gamma$ is an a-cluster. The contribution to m arising in this fashion from such subsets of cardinality j is at least $m' = \binom{k}{j}\binom{n-k}{k-j} a^{\binom{j}{2}}$. We have

$$(\log m') \Big/ \binom{k}{2} \sim H(a, p)(1 - \sqrt{a}) - a(\log a).$$

Since $H(a, p) \to \infty$ as $p \to 0$, we can choose p sufficiently small that $(\log m')/\binom{k}{2}$ tends to a positive limit and thus $m' \to \infty$. The straightforward generalization of the Poisson convergence proof therefore fails, at least for certain values of a and p. In Section 5, we will show that $m \to \infty$ whenever $a < 1$.

To summarize: subsets γ of an a-cluster α which contain a high proportion of edges make it relatively easy for a very large number of k-sets containing γ to be a-clusters; even though such subsets are exceeding rare, they contribute enough neighboring clusters when they do occur to cause

$m \to \infty$. To get around this problem, we define the notion of a "balanced" cluster, and use this in steps 1 and 2 of the proof of Theorem 3.

There is a second way in which the expected number m of a-clusters neighboring a given a-cluster α blows up: a cluster may contain a single vertex which is not rich in connections to the rest of the cluster, so that there are many ways to find a neighbor β of α in which that one vertex is replaced by some other. In Section 5, we characterize the set of (p, a) for which this contribution tends to infinity. To get around this problem, we supplement the notion of "balanced" with two more restrictions on the configuration of edges in a cluster, in the definition below of a "good" cluster. Step 3 in the proof of Theorem 3 corresponds to this problem.

Given a function $b: [0,1] \to [0,1]$, a k-set α will be called b-balanced if, for all subsets $\gamma \subset \alpha$ of cardinality j, the subgraph induced on γ contains at most $b(j/k)\binom{j}{2}$ edges. (Thus, all a-clusters are **1**-balanced, where **1** denotes the constant function with value 1.)

Let a function $b(\cdot)$, a value of $a' \in (p, a)$, and a sequence $c_1, c_2, \ldots$ with $c_k \to \infty$ be given. We will say that an a-cluster α is "good" if it satisfies the following requirements:

1. α is b-balanced.

2. For each vertex v in α, at least $a'k$ of the possible $k-1$ edges to the other vertices of α are in G.

3. The number of edges in α is less than $a\binom{k}{2} + c_k$.

Let X'_α be the indicator that α is a "good" a-cluster, and let $W' \equiv \sum_{\alpha \in I} X'_\alpha$, so that $0 \le W' \le W$. Our strategy is to show, for an appropriate choice of the parameters of "goodness," that the second moment of W' is well-behaved (Theorem 3), so that a Poisson approximation for W' can be established by the Chen–Stein method, and that $P(W \ne W') \to 0$ (Theorem 4).

THEOREM 3. *There exists a function $b: [0,1] \to [0,1]$, and for each $\varepsilon > 0$ there exists $a' \in (p, a)$, such that, if $\lambda^{1/k} \to 1$, the bounds b_1 and b_2 for the Chen–Stein method applied to W' satisfy*

$$b_1 + b_2 = O(\lambda n^{-1+\varepsilon}).$$

More precisely, we have $b_1 + b_2 = O(\lambda n^{-1} k^3 e^{k(H(a,p) - H(a',p))})$.

Proof. Suppose that we are given $b(\cdot)$, and that α is a b-balanced a-cluster of cardinality k. Let m' be the expected number of $\beta \in I$ which are also b-balanced a-clusters, so that $b_2 \le \lambda m'$, and for $j = 2, \ldots, k-1$, let m'_j be the contribution to m' arising from those β such that $|\alpha \cap \beta| = j$.

For any positive ν_1 and ν_2, we first show how to define the function $b(x)$ on the interval $[\nu_1, 1-\nu_2] \subset (0,1)$. We then show how to choose ν_1 such that we may take $b(x) = 1$ on the domain $(0, \nu_1)$. In the third step we show how to choose ν_2 and a' to take care of the contributions from $j \in [\nu_2 k, k-1]$.

Step 1. Let $[\nu_1, 1-\nu_2] \subset (0,1)$ be given. Let j be an integer such that $x \equiv j/k \in [\nu_1, 1-\nu_2]$. Suppose that all subsets of α having cardinality j induce at most $b(x)\binom{j}{2}$ edges. Let $\beta \in I$ with $\gamma \equiv \alpha \cap \beta$ of cardinality j. Since the subgraph induced on γ contains at most $b(x)\binom{j}{2}$ edges, the induced subgraph on β must contribute an additional $a\binom{k}{2} - b(x)\binom{j}{2}$ edges out of a potential $t \equiv \binom{k}{2} - \binom{j}{2}$ edges in order for β to be an a-cluster. Since each of these potential edges actually occurs with probability p, we have, with C_i denoting p-coins,

$$m'_j \le \binom{k}{j}\binom{n-k}{k-j} P\left[C_1 + \cdots + C_t \ge a\binom{k}{2} - b(x)\binom{j}{2}\right].$$

Taking logarithms and dividing by $\binom{k}{2}$, we have an upper bound u_j satisfying

$$(\log m'_j)\Big/\binom{k}{2} \le u_j \sim (1-x)H(a,p) - (1-x^2)H\big((a - b(x)x^2)/(1-x^2), p\big).$$

If we had $b(x) = a$, the right-hand side above would be strictly negative. We can therefore choose a constant $a_0 > a$ to be the value of $b(x)$ for all $x \in [\nu_1, 1-\nu_2]$, so that the right-hand side is less than $-\delta < 0$. Thus, $m'_j = O(e^{-c\binom{k}{2}}) = O(n^{-c' \log n})$, for some constants $c, c' > 0$. There are at most k such values of j to consider, for a total contribution to m' that is $o(n^{-1})$.

Step 2. Again, let $\beta \in I$ with $\gamma \equiv \alpha \cap \beta$ of cardinality $j = xk$. Even if the subgraph induced on γ is a clique (i.e., allowing $b(x) = 1$), the induced subgraph on β must contribute an additional $a\binom{k}{2} - \binom{j}{2}$ edges out of a potential $\binom{k}{2} - \binom{j}{2}$ edges in order for β to be an a-cluster. Counting as before, we have an upper bound v_j satisfying

$$(\log m'_j)/(k/2) \le v_j \sim k\big[(1-x)H(a,p) - (1-x^2)H\big((a-x^2)/(1-x^2), p\big)\big].$$

Let $h(x)$ denote the expression in square brackets, considered as a function of x. Then $h(0) = 0$ and $h'(0) = -H(a, p)$. For any $\varepsilon > 0$, we can choose ν_1 sufficiently small that for $j = 2, \ldots, \nu_1 k$,

$$\log m'_j/(k/2) \leq v_j \sim h(j/k)k < -\tfrac{2}{3}jH(a, p) \leq -\tfrac{4}{3}H(a, p).$$

Since $(\log n^{-1})/(k/2) \sim -H(a, p)$, we have $m'_2 + \cdots + m'_{\nu_1 k} = o(n^{-1})$.

Step 3. Consider m_{k-1}, which involves replacing a single vertex of α. The new vertex, in order to form a "good" cluster with the rest of α, must contribute at least $a'k$ out of a possible $k - 1$ edges to the other vertices of α. The probability of this is at most $e^{-kH(a', p)}$. Thus

$$m_{k-1} \leq k(n-k)e^{-kH(a', p)} \sim e^{-2}k^3 n^{-1} e^{(k-1)H(a, p)} e^{-kH(a', p)}.$$

Given $\varepsilon > 0$, we can choose a' close enough to a so that this upper bound is $O(n^{-1+\varepsilon})$.

The same argument shows that for fixed $d = 1, 2, \ldots$, we have, with the same choice of a', that $m_{j-d} = O(n^{-d+\varepsilon})$, and there is a value $\nu_2 < 1$ such that $m_j = O(n^{-1})$, uniformly in $j = \nu_2 k, \ldots, k - 2$. Thus m_{k-1} is the term which makes the dominant contribution to our upper bound on b_2. Q.E.D.

THEOREM 4. *Suppose that* $b(x) : [0, 1] \to [0, 1]$ *with* $b(\cdot) = 1$ *on* $[0, \nu_1) \cup (1 - \nu_2, 1]$ *and* $b(\cdot) = a_0 > a$ *on* $[\nu_1, \nu_2]$, *that* $a' \in (p, a)$, *and that* $c_k/k \to \infty$, $c_k/k^2 \to 0$. *For some constant* $\delta > 0$ (*depending only on* a, p, *and* b),

$$P(W \neq W') = O(n^{-\delta}).$$

More precisely, $P(W \neq W') = O(\lambda k e^{-kH(a', a)})$.

Proof. Since $\lambda \equiv EW$ is assumed to satisfy $\lambda^{1/k} \to 1$, and $0 \leq W' \leq W$ with $E(W - W') = \lambda P(X'_\alpha = 0 | X_\alpha = 1)$, it suffices to show that $P(X'_\alpha = 0 | X_\alpha = 1) \to 0$ exponentially fast in k. The contribution to this conditional expectation from unbalanced but otherwise good subsets of size $j = xk$ is less than $\binom{k}{j} P(C_1 + \cdots + C_t > a_0 t - c_k)$, where the C_i are a-coins, and $t = \binom{j}{2}$. Taking logarithms and dividing by $\binom{k}{2}$, and using $c_k/k^2 \to 0$, we get the limit $-x^2 H(a_0, a) < 0$, so the entire contribution from unbalanced subsets decays faster than exponentially in k. The contribution to the conditional probability that $X'_\alpha = 0$ from single vertices which have fewer than $a'k$ edges to the rest of α is less than $kP(C_1 + \cdots + C_{k-1} < a'k) < ke^{-kH(a', a)}$. The contribution to the conditional probability due to the total number of edges exceeding $a\binom{k}{2}$ by c_k or more decays faster than exponentially in k since $c_k/k \to \infty$. Q.E.D.

We have now proven the main results.

THEOREM 5. *Let $0 < p < a \leq 1$. Suppose that $G = G(n, p)$ is a random graph on n vertices with edges chosen independently with probability p. Let W denote the number of a-clusters of cardinality k in G, where k and n tend to infinity in such a way that the expected number $\lambda \equiv EW$ of a-clusters of cardinality k satisfies $\lambda^{1/k} \to 1$. Then W has asymptotically a Poisson distribution, and the total variation distance, between the entire collection of dependent events, $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$ and a process $\mathbf{X}'$ with the same marginal but independent coordinates, tends to zero.*

Proof. Combining theorems 1, 3, and 4, we have that $|L(W) - L(Z_\lambda)| \leq b_1 + b_2 + P(W \neq W') = O(n^{-1}k^3 e^{kH(a,p)-H(a',p)} + ke^{-H(a',a)})$. Using relation (2), we see that this upper bound has the form $|L(W) - L(Z_\lambda)| = O(k^2 n^{-\delta})$, where the optimal choice of a' relative to a and p leads to an explicit value $\delta \equiv \delta(p, a) \in (0, 1)$. Using Theorem 2 instead of Theorem 1 leads to a similar upper bound on the total variation distance between the original and decoupled families of events, $\mathbf{X}$ and $\mathbf{X}'$. Q.E.D.

Just as in the study of cliques in random graphs, corresponding to $a = 1$, when k and n satisfy (2), changing k by 1 causes λ to change by a factor which is of the same order as k. Thus, as $n \to \infty$ along the integers, it is not possible to pick $k \equiv k(n)$ such that λ stays bounded away from zero and infinity. These considerations motivate the following theorem, which is an easy corollary of the theorem above.

THEOREM 6. *Suppose further that $n \to \infty$, and $k \equiv k(n)$ is such that the expected number of a-clusters of cardinality k stays bounded away from infinity, while the expected number of a-clusters of cardinality $k - 1$ tends to infinity. If $\mathrm{cl}_a(G)$ denotes the cardinality of the largest a-cluster in G, then*

$$|P(\mathrm{cl}_a(G(n, p)) = k - 1) - e^{-\lambda}| \to 0,$$

$$|P(\mathrm{cl}_a(G(n, p)) = k) - (1 - e^{-\lambda})| \to 0.$$

4. VARIATIONS ON THE THEME

Certain variations on the basic theme arise naturally in the context of applications to the problem of evaluating the significance of patterns of protein similarities in molecular biology. We omit the proofs, which are relatively straightforward modifications of the arguments given above.

THEOREM 7. *Let $G = G(n, p)$ be a random graph, let $p < a \leq 1$ and let c be a fixed positive integer. Define a flag to be any pair (v, e), consisting of a vertex v and an edge e from v. Suppose that every flag (v, e) in G is randomly assigned one of c colors. A subgraph H is said to be nicely colored if, for all vertices $v \in H$, the flags in H containing v are all the same color. Let W denote the number of nicely colored subgraphs of cardinality k containing at least $a\binom{k}{2}$ edges, where $k, n \to \infty$ so that with $t = \binom{k}{2}$ and $C_1, C_2, \ldots$ independent, $\{0,1\}$-valued with mean p/c^2, $EW = c^k\binom{n}{k}P(C_1 + \cdots + C_t \geq at)$ satisfies $(EW)^{1/k} \to 1$. Then W has asymptotically a Poisson distribution.*

THEOREM 8. *The results of Theorem* 5, *Corollary* 6, *and Theorem* 7 *remain true if G is chosen*: (i) *at random from the set of all graphs on n vertices with constant valence $[pn]$*; (ii) *at random from the set of all graphs on n vertices with $[p\binom{n}{2}]$ edges*; (iii) *by randomly choosing, for each vertex, a set of $[xn]$ edges incident with the vertex, with $x \in (0,1)$ satisfying $2x - x^2 = p$.*

5. THE EXPECTED NUMBER OF NEIGHBORS TENDS TO INFINITY

As above, let m denote the expected number of a-clusters neighboring a given a-cluster. What necessitated the additional work above was the recognition that $m \to \infty$ for certain values of a and p. We close by briefly characterizing the ways in which this occurs, in particular showing that $m \to \infty$ for all $a \neq 1$.

There are two different effects that can cause $m \to \infty$. The first, corresponding to step 1 in the proof of Theorem 3, involves a contribution to m that grows like $\exp(ck^2)$, due to unbalanced subsets of a given cluster. The second, corresponding to Step 3 in the proof of Theorem 3, is a contribution that grows like e^{ck}, due to single vertices in a given cluster which are not rich enough in connections to the rest of the cluster and thus are easily replaced.

First we identify the constant c in the contribution to m that grows like $\exp(ck^2)$, due to unbalanced subsets of a given cluster. Let α be an a-cluster and let m_j denote the contribution to m arising from $\beta \in I$ such that $\gamma \equiv \alpha \cap \beta$ has cardinality j. Step 2 in the proof of Theorem 4 shows that we may neglect those m_j with $j/k \in (0, \nu_1)$.

We can compute m_j, for cases where $x \equiv j/k$ is bounded away from 1, as follows. For all choices of γ and $\beta - \gamma$, consider all choices for the number U of edges in the subgraph induced on γ and the number V of additional edges contributed by the subgraph induced on β subject to the

condition that $U + V \geq a\binom{k}{2}$. Thus,

$$m_j = \binom{k}{j}\binom{n-k}{k-j} \sum_{u=0}^{\binom{j}{2}} P\left(U = u, V \geq a\binom{k}{2} - u\right).$$

Since j/k is bounded away from 0 and 1, we can use the large deviation estimates for the binomial. Writing $x = j/k$, $y = u/\binom{j}{2}$, and $z = v/\left(\binom{k}{2} - \binom{j}{2}\right)$, we have

$$(\log m)\Big/\binom{k}{2} \to \max F_{a,p}(x, y, z),$$

where

$$F_{a,p}(x, y, z) = H(a, p)(1 - x) - H(y, a)x^2 - H(z, p)(1 - x^2),$$

and the maximum is taken over the set $\{(x, y, z): yx^2 + z(1 - x^2) \geq a, 0 \leq x \leq 1, a \leq y \leq 1, p \leq z \leq a\}$. Let $f_{a,p}$ denote this maximum. Thus, $f_{a,p} > 0$ implies $m \to \infty$, and the coefficient c of k^2 in the exponential growth rate of m is $\frac{1}{2}f_{a,p}$.

It is easy to verify that for all $0 < p < a < 1$, we have $f_{a,p} > 0$. First, fix $p < z < a$ so that $H(z, p) < H(a, p)/3$. Take $y = a + \varepsilon(a - z)$ and $x^2 = 1/(1 + \varepsilon)$. As $\varepsilon \to 0$ we have $F_{a,p}(x, y, z) \sim H(a, p)\varepsilon/2 + 0 - H(z, p)\varepsilon$, so that for sufficiently small positive ε, $F_{a,p}(x, y, z) > 0$ for a point (x, y, z) in the desired region.

We analyze the growth of m_{k-1} as follows. Consider the contribution that arises when the number U of excess edges in the cluster above the minimum number $a\binom{k}{2}$ is about ck, and the number D of edges between a particular vertex and the rest of the cluster is about bk. With C_i denoting p-coins, the contribution just described is approximately

$$k(n - k)P(U \geq ck)P(D \leq bk)P(C_1 + \cdots + C_{k-1} \geq (b - c)k).$$

Taking the limit of $1/k$ times the logarithm of this contribution yields

$$H(a, p)/2 - [cH'(a, p) + H(b, a) + H(b - c, p)].$$

In the region $p \leq b - c \leq b \leq a$, the expression in brackets is minimized with $c = 0$. This indicates that excess edges in the cluster as a whole do not play the dominant role in this contribution. With

$$g_{a,p} \equiv H(a, p)/2 - \min_{p \leq b \leq a} [H(b, a) + H(b, p)],$$

we have that if $g_{a,p} > 0$ then $m \to \infty$, and if $g_{a,p} < 0$ then m_{k-1} goes to zero. A similar analysis applies to m_{k-j} whenever $j/k \to 0$.

Thus we understand why the expected number of clusters which neighbor a given cluster tends to infinity. Of course, even when the expected number of neighbors is large, the Poisson convergence result above implies that the actual number of neighbors is 0 with probability tending to 1.

ACKNOWLEDGMENTS

E.S.L. wishes to thank the Fondation les Treilles and Mme. Gruner-Schlumberger for organizing the unusual conference from which the questions addressed in this work arose. The two authors are grateful to our mutual colleague who brought us together, namely Michael S. Waterman.

REFERENCES

1. R. ARRATIA, L. GOLDSTEIN, AND L. GORDON, Two moments suffice for Poisson approximations: The Chen–Stein method, *Ann. Probab.* **17**, 9–25.
2. R. ARRATIA AND L. GORDON, Tutorial on large deviations for the binomial distribution, *Bull. Math. Biol.* **51** (1989), 125–131.
3. R. H. BAHADUR, "Some Limit Theorems in Statistics," SIAM Regional Conference Series in Applied Mathematics, Vol. 4, SIAM, Philadelphia, 1971.
4. B. BOLLABÁS, "Random Graphs," Academic Press, New York/London, 1985.
5. B. BOLLOBÁS AND P. ERDÖS, Cliques in random graphs, *Math. Proc. Cambridge Philos. Soc.* **80** (1976), 419–427.
6. L. H. Y. CHEN, Poisson approximation for dependent trials, *Ann. Probab.* **3** (1975), 534–545.
7. R. F. DOOLITTLE, "Of Urfs and Orfs," University Science Books, Mill Valley, CA, 1986.
8. E. S. LANDER, J. P. MESIROV, AND W. TAYLOR, Protein sequence comparison on a data parallel computer, *in* "Proceedings of the First International Conference on Parallel Processing, Vol. 3, pp. 257–263, 1988."
9. E. S. LANDER, J. P. MESIROV, AND W. TAYLOR, Study of protein sequence comparison metrics on the connection machine CM-2, Proceedings of the Supercomputing 1988 Conference, Vol. 2, IEEE Computer Society Press, 1988.
10. J. SPENCER, "Ten Lectures on the Probabilistic Method," SIAM CBMS-NSF Regional Conferences Series in Applied Mathematics, SIAM Vol. 52, SIAM, Philadelphia, 1987.
11. C. M. STEIN, "Approximate Computation of Expectations," IMS, Hayward, CA, 1986.