DISMANTLING SPARSE RANDOM GRAPHS

SVANTE JANSON AND ANDREW THOMASON

Abstract. We consider the number of vertices that must be removed from a graph G in order that the remaining subgraph has no component with more than k vertices. Our principal observation is that, if G is a sparse random graph or a random regular graph on n vertices with $n \to \infty$, then the number in question is essentially the same for all values of k that satisfy both $k \to \infty$ and $k = o(n)$.

The process of removing vertices from a graph G so that the remaining subgraph has only small components is known as *fragmentation*. Typically, the aim is to remove the least possible number of vertices to achieve a given component size; this is equivalent to determining the largest induced subgraph whose components are at most that size. This process has been studied in (at least) two different lines of research, from different perspectives and with quite different component sizes. In this note we point out that, as far as sparse random graphs are concerned, these two perspectives actually arrive at the same answer.

Let Γ be a class of graphs. The classes we shall mostly be interested in are the classes $\mathcal{C}_k$, the class of graphs whose components have at most k vertices, and $\mathcal{F}$, the class of forests. Given such a class Γ, we define

$$N(G,\Gamma) := \max\{|S| \,:\, G[S] \in \Gamma\},$$

where S is a subset of the vertices of G and $G[S]$ denotes the subgraph of G induced by S. We also define

$$\nu(G,\Gamma) := N(G,\Gamma)/|G|,$$

so that $0 \le \nu(G,\Gamma) \le 1$. (To make this always defined, we set $N(G,\Gamma) = 0$ if no induced subgraph of G belongs to Γ; equivalently, we may regard the empty graph with no vertices as an element of Γ.) Thus, for example, the size of a largest independent set in G is $N(G,\mathcal{C}_1) = \nu(G,\mathcal{C}_1)|G|$. (This is known as the *independence number*.) Similarly, $n - N(G,\mathcal{F})$ is the *decycling number*, see e.g. Karp [16].

In this notation, the study of fragmentation is the study of the parameter $\nu(G,\mathcal{C}_k)$ for various values of k. From the point of view of graph theory, it is natural to consider $\nu(G,\mathcal{C}_k)$ for some large but finite value of k, for graphs G in which the number of vertices $n = |G|$ grows large. This study was initiated by Edwards and Farr [5; 7]. On the other hand, in the study

Date: September 12, 2007.
2000 *Mathematics Subject Classification.* 05C80; 05C40, 92D30.

of vaccination, see for example Britton, Janson and Martin-Löf [4] and the references therein, the vertices of the graph are individuals in some population, with edges representing the opportunity of passing on a disease. If a vertex is vaccinated it becomes unable to spread the disease; a vaccination strategy is a way to ensure that the subgraph induced by the unvaccinated vertices has only small components (relative to the total population). The vaccinator is thus interested in $\nu(G, \mathcal{C}_{\delta n})$ for small values of δ. (For further details and for variations on this theme, see [4].)

In both studies it is natural to consider the behaviour of these parameters on two standard models of sparse random graphs. Let $\mathcal{G}(n, c/n)$ denote the probability space of graphs with vertex set $\{1, \ldots, n\}$ with edges chosen independently with probability c/n, and let $\mathcal{G}_d(n)$ denote the space of d-regular graphs on the same vertex set. We shall assume that $c > 1$ and $d \geq 3$ are fixed as $n \to \infty$. (For odd d, we of course have to assume that n is even.) The main observation of this paper is that, for graphs in these spaces, the graph theoretic approach and the vaccination approach arrive at the same answer: that is, perhaps surprisingly, fragmenting into large but finite components whp costs no more than just fragmenting into components of size $o(|G|)$. (A sequence of events (A_n) is said to hold whp if $\Pr(A_n) \to 1$ as $n \to \infty$.)

Theorem 1. *Let $c > 1$, $d \geq 3$ and $\epsilon > 0$ be given. Then there exists $\delta > 0$ such that, if $G \in \mathcal{G}(n, c/n)$ or $G \in \mathcal{G}_d(n)$ then*

$$\nu(G, \mathcal{C}_{\delta n}) \leq \nu(G, \mathcal{C}_{1/\delta}) + \epsilon$$

holds whp as $n \to \infty$.

Before giving the proof of this theorem, we make a few more remarks. Edwards and Farr [5; 7] considered general graphs of bounded maximum degree; in particular they studied the parameter

$$\beta_d := \sup_k \min\{\nu(G, \mathcal{C}_k) : G \text{ has maximum degree } d\}$$

(note that the α_d of Edwards and Farr equals $1 - \beta_d$). One way to think of this parameter is that, if $\beta < \beta_d$, then there is some finite k for which every graph G of maximum degree d has an induced subgraph $G[S]$ with at least $\beta|G|$ vertices but with no component larger than k. Trivially $\beta_1 = \beta_2 = 1$. and it is shown in [5] that $\beta_3 = \frac{3}{4}$. In general they showed that $\beta_d \geq \frac{3}{d+1}$; a complementary inequality $\beta_{2d} \leq \frac{2}{d+1}$ was proved by Haxell, Pikhurko and Thomason [13] (so answering affirmatively the question posed by Edwards and Farr [6] as to whether $\beta_d \to 0$ as $d \to \infty$).

The parameter $\nu(G, \mathcal{F})$, describing the largest induced forest in G, is significant in the study of fragmentation, because any forest F is easily fragmented by removing a few vertices. The following simple lemma is given only because it is best possible, as exemplified by a path.

Lemma 2. *If F is a forest then $\nu(F, \mathcal{C}_k) \geq 1 - (k + 1)^{-1}$.*

Proof. We may assume that F is a tree, and proceed by induction on $n = |F|$, the case $n \leq k+1$ being trivial. For larger n, note that the removal of any edge leaves two components. Orient the edge towards the larger of these (break ties arbitrarily) and colour the edge red if both components have more than k vertices. If there are no red edges, remove a sink vertex (there must be one since F is acyclic) and observe that this leaves only components with at most k vertices, and thus $\nu(F, \mathcal{C}_k) = 1 - 1/n \geq 1 - 1/(k+1)$. If there are red edges, it is easy to see that they form a connected subgraph and so a tree; the removal of a leaf vertex of the red tree breaks F into a tree with at most $n-k-1$ vertices plus some components of size at most k, and the proof follows by induction. □

The parameter $\nu(G, \mathcal{P})$, where $\mathcal{P}$ is the class of planar graphs, is likewise significant for fragmentation, because a planar graph P can be fragmented quite efficiently by means of the separator theorem of Lipton and Tarjan [17], and in fact $\nu(P, \mathcal{C}_k) \geq 1 - 24k^{-1/2}$ holds (see [5] or [8]). From this and from Lemma 2 it follows that, for any graph G, both $\nu(G, \mathcal{C}_k) \geq \nu(G, \mathcal{F}) - (k+1)^{-1}$ and $\nu(G, \mathcal{C}_k) \geq \nu(G, \mathcal{P}) - 24k^{-1/2}$ hold. On the other hand, if $n_k(G)$ denotes the number of cyles in G of length at most k, then (by removing a vertex from each cycle) we have $\nu(G, \mathcal{P}) \geq \nu(G, \mathcal{F}) \geq \nu(G, \mathcal{C}_k) - n_k(G)/|G|$. Thus if we restrict our attention to large graphs G for which $n_k(G) = o(|G|)$ for each fixed k, then the three parameters $\nu(G, \mathcal{F})$, $\nu(G, \mathcal{P})$ and $\nu(G, \mathcal{C}_k)$ are asymptotically the same for large k. Graphs in $\mathcal{G}(n, c/n)$ or $\mathcal{G}_d(n)$ enjoy this property whp (see [2] or [15]).

The parameters $\nu(G, \mathcal{C}_1)$ and $\nu(G, \mathcal{F})$ for $G \in \mathcal{G}(n, c/n)$ and $G \in \mathcal{G}_d(n)$ have already received considerable attention. The first of these (the independence number) was studied by Frieze [10] for $\mathcal{G}(n, c/n)$, see also [15, Section 7.1], and Frieze and Łuczak [11] for $G \in \mathcal{G}_d(n)$, and information on the second is given by Bau, Wormald and Zhou [1]. In fact it is shown in [11] that, for $G \in \mathcal{G}_d(n)$, $\nu(G, \mathcal{C}_1) \sim 2 \log d/d$ holds whp for large d, which is already enough to answer the above-mentioned question about the limit of β_d, and in [13] it is verified that $\nu(G, \mathcal{F}) \sim 2 \log d/d$ whp. (These statements involve double limits, as first $n \to \infty$ and then $d \to \infty$. More precisely, by [11; 13], for every $\varepsilon > 0$, there exists d_ε such that for $\mathcal{G}_d(n)$ with any fixed $d \geq d_\varepsilon$, whp holds $(2-\varepsilon) \log d/d \leq \nu(G, \mathcal{C}_1) \leq \nu(G, \mathcal{F}) \leq (2+\varepsilon) \log d/d$ as $n \to \infty$. A similar result holds for $\mathcal{G}(n, c/n)$, by [10] and a first moment argument as in [13].)

We are now ready for the proof of Theorem 1.

Proof of Theorem 1. We claim that the following holds whp, if δ is small enough: *Each set T of at most δn vertices spans at most $(1+\epsilon/3)|T|$ edges.*

The theorem follows from this claim; for let S be a set such that $G[S] \in \mathcal{C}_{\delta n}$ and $|S| = \nu(G, \mathcal{C}_{\delta n})$. By the claim, from each component $G[T]$ of $G[S]$ we may remove at most $\epsilon|T|/3$ edges so that it becomes acyclic or unicyclic; thus, by removing $\epsilon|S|/3$ edges we can make all components of $G[S]$ acyclic or unicylic. There are at most $\epsilon|S|/3$ components of size larger than $3/\epsilon$, and so by removing a further $\epsilon|S|/3$ edges we can make all these large components acyclic. Hence, removing vertices instead of edges, there exists $S' \subset S$, $|S'| \geq |S| - 2\epsilon n/3$, such that $G[S']$ consists of a forest F plus components of size at most $3/\epsilon$. By Lemma 2, the removal of a further $\epsilon|S'|/3$ vertices from $G[S']$ leaves only components of size at most $3/\epsilon$. Therefore $\nu(G, \mathcal{C}_{3/\epsilon}) \geq \nu(G, \mathcal{C}_{\delta n}) - \epsilon$. We can of course assume that $\delta < \epsilon/3$, and this proves the theorem.

To prove the claim, consider first the case $G \in \mathcal{G}(n, c/n)$. Let T be a set of $t \leq n\delta$ vertices, and let $\tau = n/t \geq 1/\delta$; we can make τ large by making δ small. Let X be the random variable counting the number of edges in $G[T]$. Then X is binomially distributed with mean $\lambda = \frac{c}{n}\binom{t}{2} \leq \frac{ct}{2\tau}$. By the version of the Chernoff bound in [15, Corollary 2.4], if $x = m\lambda > \lambda$, then $\mathbb{P}(X \geq x) \leq \exp(-lx)$, where $l = \log m - 1 + 1/m$. Taking $x = (1+\epsilon/4)t$ we have $m = x/\lambda \geq (1+\epsilon/4)2\tau/c > 1$ if $\delta < 2/c$, and $l > \log m - 1 \geq \log\tau - 1 - \log c$. The number of sets of size t is $\binom{n}{t} \leq (e\tau)^t$. Therefore the probability P_t that the claim fails for some set T of size t satisfies

$$P_t \;\leq\; \exp\{t(1+\log\tau) - t(1+\epsilon/4)(\log\tau - 1 - \log c)\} \;\leq\; \exp\{-\epsilon t(\log\tau)/8\}$$

if δ is small enough. For $t \geq \log n$ we have $P_t \leq n^{-2}$ if δ is small, and for $t \leq \log n$ we have $\log\tau \geq (\log n)/2$ so $P_t \leq n^{-\epsilon/16}$. Thus $\sum_{1 \leq t \leq \delta n} P_t = o(1)$, which proves the claim.

The case $G \in \mathcal{G}_d(n)$ is similar. The calculation is messier but fortunately we need not give it here, because it is essentially that of the proof of Lemma 5.1 of Janson and Luczak [14]. They prove (see Remark 5.2) that each set T with $|T| \leq \delta n$ has average degree less than k, where $k \geq 3$. However their interest was in integer values of k, and the proof, in which k appears everywhere as a variable, works perfectly well for any fixed $k > 2$, which is exactly our claim. ☐

In the light of Theorem 1 it is interesting to consider the fragmentation of random graphs G into components of size proportional to $|G|$. Now, given $c > 0$ and $0 < x \leq 1$, the inequalities of Azuma–Hoeffding [15, Corollary 2.27] or of Talagrand [15, Theorem 2.29 and Remark 2.36] show that the value of $\nu(G, \mathcal{C}_{xn})$ for $G \in \mathcal{G}(n, c/n)$ is highly concentrated. It seems very likely that there is a real number $f_c(x)$ such that $\nu(G, \mathcal{C}_{xn})$ tends to $f_c(x)$ in probability; that is, for all $\epsilon > 0$, $\Pr\{|\nu(G, \mathcal{C}_{xn}) - f_c(x)| > \epsilon\} \to 0$ as $n \to \infty$. But it is not actually known whether $f_c(x)$ exists; this is an unfortunate state of affairs shared with many standard parameters such as $\nu(G, \mathcal{C}_1)$ and $\nu(G, \mathcal{F})$. (An exception here is $\nu(G, \mathcal{C}_1)$ when $c \leq e$; see [12, Corollary 1].) Nevertheless, our final comments can be stated more cleanly by assuming both that $f_c(x)$ exists, and also that, for each fixed k, $\nu(G, \mathcal{C}_k)$ tends to a limit $\phi_c(k)$ in probability. Corresponding limits will be assumed for $G \in \mathcal{G}_d(n)$ too, and we denote these by $g_d(x)$ and $\gamma_d(k)$. (Note that, at least, one can show that $\nu(G, \mathcal{C}_{xn})$ and $\nu(G, \mathcal{C}_k)$ are highly concentrated for $G \in \mathcal{G}_d(n)$ too, using a version for random permutations of the inequality by Talagrand [20, Theorem 5.1], see also McDiarmid [19, Theorem 1.1], and arguing as in Haxell, Pikhurko and Thomason [13].)

We interpret $f_c(x)$ as meaning that the largest induced subgraph of a random graph $G \in \mathcal{G}(n, c/n)$ having no component larger than xn has (about) $f_c(x)n$ vertices, and the largest induced subgraph having no component larger than k has $\phi_c(k)n$ vertices. The numbers $g_d(x)$ and $\gamma_d(k)$ have similar interpretations. All these functions are increasing in their arguments. So the limits $\lim_{x\to 0} f_c(x)$ and $\lim_{x\to 0} g_d(x)$ exist; we define $f_c(0)$ and $g_d(0)$ to be the values of these limits, thus making f_c and g_d continuous at zero. The content of Theorem 1 is then that

$$\lim_{k\to\infty} \phi_c(k) = f_c(0) := \lim_{x\to 0} f_c(x) \quad \text{and} \quad \lim_{k\to\infty} \gamma_d(k) = g_d(0) := \lim_{x\to 0} g_d(x)\,.$$

[Note that $\lim \gamma_d(k)$ corresponds very closely to the parameter β_d defined earlier, the only difference being that β_d takes account of all d-regular graphs, whereas here we consider only almost all.]

The function f_c satisfies the Lipschitz condition $f_c(x) - f_c(y) < \frac{1}{y}(x - y)$ for $y < x$, since, if $G[S]$ has component sizes at most xn, then at most $1/y$ of these are larger than yn, and each of these can be reduced to size yn by removing at most $(x - y)n$ vertices. This, together with the fact that f_c is increasing, shows that f_c is a continuous function on the unit interval, and similarly so is g_d.

The famous theorem of Erdős and Rényi [9], that $G \in \mathcal{G}(n, c/n)$ almost certainly has a unique giant component of size $(1+o(1))\rho(c)n$, where $\rho(c) = 1-e^{-c\rho(c)}$, means that $f_c(x) = 1$ for $\rho(c) \le x \le 1$. There is no corresponding fact for random regular graphs, of course; we just have $g_d(1) = 1$.

It seems likely that the function f_c is *strictly* increasing on the interval $[0, \rho(c)]$ and that g_d is strictly increasing on $[0, 1]$. This would mean that continuous inverse functions f_c^{-1} and g_d^{-1} exist. For the vaccinator, the function f_c^{-1} would be more natural than f_c itself; $f_c^{-1}(z)n$ is the smallest component size achievable by vaccinating $(1 - z)n$ people. But we cannot show strict monotonicity except at the right-hand end of the range. It was proved by Bollobás, Janson and Riordan [3, Theorem 3.9] that for every $\epsilon > 0$ there exists $\delta > 0$ such that whp after removal of any δn vertices from $\mathcal{G}(n, c/n)$, there is still a giant component of order at least $(\rho(c) - \varepsilon/2)n$, and thus $f_c(\rho(c) - \epsilon) < 1 - \delta = f(\rho(c)) - \delta$; this is an easy consequence of the corresponding result for edge deletions by Luczak and McDiarmid [18, Lemma 2], who proved that for every $\epsilon > 0$ there exists $\delta > 0$ such that the giant component whp has no two sets, each of size at least ϵn, having at most δn edges between the sets. A similar argument can be given to show that $g_d(1 - \epsilon) < g_d(1) - \delta$.

In conclusion, our main open question is whether f_c and g_d are strictly increasing. We would also like to know more about the subgraphs $G[S]$ of order $f_c(x)n$ or $g_d(x)n$ that have no component larger than xn: how many components do they have? The reader who is interested in these questions can readily formulate them in a way that does not involve the uncertain existence of f_c and g_d.

We finally remark that corresponding questions can be formulated for removal of edges instead of vertices. In that case, the central parameter is the largest number of edges in a (not necessarily induced) subgraph of G that belongs to Γ. For $\mathcal{F}$, the class of forests, this is easy (unlike the vertex case treated above, see Karp [16]), but we see no easy answers for e.g. $\mathcal{C}_{\delta n}$ and leave these versions as problems for the interested reader.

Acknowledgement

This research was mainly done during a visit by SJ to the University of Cambridge, partly funded by Trinity College.

References

[1] S. Bau, N. C. Wormald & S. Zhou, Decycling numbers of random regular graphs. *Random Struct. Alg.* **21** (2002), 397–413.
[2] B. Bollobás, *Random Graphs*, 2nd edn, Cambridge Univ. Press, Cambridge, 2001.
[3] B. Bollobás, S. Janson & O. Riordan, The phase transition in inhomogeneous random graphs. *Random Struct. Alg.* **31** (2007), 3–122.
[4] T. Britton, S. Janson & A. Martin-Löf, Graphs with specified degree distributions, simple epidemics and local vaccination strategies. Preprint 2007. `arXiv:math/0702021v1`
[5] K. Edwards & G. Farr, Fragmentability of graphs. *J. Combin. Theory B* **82** (2001), 30–37.
[6] K. Edwards & G. Farr, Research problems from the 18th British Combinatorial Conference, *Discrete Math.* **266** (2003), 441–451.
[7] K. Edwards & G. Farr, Planarization and fragmentability of some classes of graphs. *Discrete Math.* (2007) (to appear).
[8] K. Edwards & C. McDiarmid, New upper bounds on harmonious colorings. *J. Graph Theory*, **18** (1994), 257–267.
[9] P. Erdős & A. Rényi, On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* **5** (1960), 17–61.
[10] A. M. Frieze, On the independence number of random graphs. *Discrete Math.* **81** (1990), 171–175.
[11] A. M. Frieze & T. Łuczak, On the independence and chromatic numbers of random regular graphs. *J. Combin. Theory* (B) **54** (1992), 123–132.
[12] D. Garmanik, T. Nowicki & G. Swirszcz, Maximum weight independent sets and matchings in sparse random graphs. Exact results using the local weak convergence method. *Random Struct. Alg.* **28** (2006), 76–106.
[13] P. Haxell, O. Pikhurko & A. Thomason, Maximum acyclic and fragmented sets in regular graphs. *J. Graph Th.*, to appear.

[14] S. Janson & M. Luczak, A simple solution to the k-core problem. *Random Struct. Alg.* **30** (2007), 50–62.

[15] S. Janson, T. Łuczak & A. Ruciński, *Random Graphs.* Wiley, New York, 2000.

[16] R. M. Karp, Reducibility among combinatorial problems. *Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*, pp. 85–103. Plenum, New York, 1972.

[17] R.J. Lipton & R.E. Tarjan, A separator theorem for planar graphs. *SIAM J. Appl. Math.* **36** (1979), 177–189.

[18] M. Luczak & C. McDiarmid, Bisecting sparse random graphs. *Random Struct. Alg.* **18** (2001), 31–38.

[19] C. McDiarmid, Concentration for independent permutations. *Combin. Prob. Computing*, **11** (2002), 163–178.

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

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden

E-mail address: `svante.janson@math.uu.se`

URL: `http://www.math.uu.se/~svante/`

DPMMS, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, United Kingdom

E-mail address: `a.g.thomason@dpmms.cam.ac.uk`