CONTAGIOUS SETS IN RANDOM GRAPHS

By Uriel Feige[*,§], Michael Krivelevich[†,¶] and Daniel Reichman[*,‖]

The Weizmann Institute of Science[§] Tel Aviv University[¶] and University of California, Berkeley[‖]

We consider the following activation process in undirected graphs: a vertex is active either if it belongs to a set of initially activated vertices or if at some point it has at least r active neighbors. A *contagious set* is a set whose activation results with the entire graph being active. Given a graph G, let $m(G, r)$ be the minimal size of a contagious set.

We study this process on the binomial random graph $G := G(n, p)$ with $p := \frac{d}{n}$ and $1 \ll d \ll \left(\frac{n \log \log n}{\log^2 n} \right)^{\frac{r-1}{r}}$. Assuming $r > 1$ to be a constant that does not depend on n, we prove that

$$m(G, r) = \Theta \left(\frac{n}{d^{\frac{r}{r-1}} \log d} \right),$$

with high probability. We also show that the threshold probability for $m(G, r) = r$ to hold is $p^* = \Theta \left(\frac{1}{(n \log^{r-1} n)^{1/r}} \right)$.

1

1. Introduction. In *r-neighbor bootstrap percolation* we are given an undirected graph $G = (V, E)$ and an integer $r > 1$. Every vertex is either *active* or *inactive*. A set of vertices composed entirely of active vertices is called active. Initially, a set of vertices A_0 is activated. These vertices are called *seeds*. A contagious process evolves in discrete steps where for $i > 0$,

$$A_i = A_{i-1} \cup \{v : |N(v) \cap A_{i-1}| \geq r\},$$

and $N(v)$ is the set of neighbors of v. In words, a vertex becomes active in a given step if it has at least r active neighbors. In this case we shall say the vertex is infected[1]. We refer to r throughout this paper as the *threshold*. Set

$$\langle A_0 \rangle := \bigcup_i A_i.$$

Definition 1 *Given $G = (V, E)$ and a threshold r, a set $A_0 \subseteq V$ is called contagious if $\langle A_0 \rangle = V$. That is, activating A_0 results with the entire graph being activated. The minimal cardinality of a contagious set in G is denoted in G by $m(G, r)$. The* number of generations *of a (not necessarily contagious) set A_0 which we denote by $\tau := \tau(A_0)$ is the minimal integer such that $\bigcup_{i \leq \tau} A_i = \langle A_0 \rangle$.*

Bootstrap percolation has been studied for a variety of graphs [5, 6, 7, 9, 20, 21]. Here we focus on the random graph $G(n, p)$ on n labeled vertices, where every possible edge appears independently with probability p. Our interest is in providing both upper and lower bounds on the typical size of a contagious set of minimal cardinality. We remark that the term "bootstrap percolation" is often used with respect to choosing vertices independently with some probability q to the set of seeds. In contrast, in this work we do not restrict ourselves to the study of randomly generated contagious sets.

Studying the behavior of combinatorial quantities in $G(n, p)$ has a long and rich history [10], and has resulted in a plethora of ideas which have proven useful in other contexts as well. In addition, there is much interest in studying computational problems on random graphs [18]. Furthermore, combinatorial and algorithmic ideas originating from the study of the model $G(n, p)$ of random graphs are often useful in the study of more general families of random graphs. Hence, beyond the intrinsic value of studying the value of $m(G, r)$ in $G(n, p)$ which we consider to be of interest of its own right, we believe the ideas in the current work may prove applicable in other contexts where contagious processes are studied.

---

[1]We alert the reader that in [21] the term *infected* is used in a sense different than ours. There a vertex is said to be infected if it has at least one active neighbor but fewer than r active neighbors.

1.1. *Our results.* Consider $G(n, p)$, and let $p := \frac{d}{n}$. We obtain a nearly tight characterization of the probable value of $m(G, r)$. We say an event in the probability space $G(n, p)$ occurs "typically" or "with high probability" (w.h.p.) if it occurs with probability $1 - o(1)$, where $o(1)$ represents a term that tends to 0 as n tends to infinity. For two integer-valued functions $f(n), g(n)$, we say that $f(n) \ll g(n)$ if $\lim_{n \to \infty} \frac{f(n)}{g(n)} = 0$.

Theorem 1.1 *Let $G \sim G(n, p)$ with $p := \frac{d}{n}$ and*

$$1 \ll d \ll \left(\frac{n \log \log n}{\log^2 n} \right)^{\frac{r-1}{r}} .$$

Then with high probability

$$m(G, r) = \Theta \left(\frac{n}{d^{\frac{r}{r-1}} \log d} \right).$$

The upper bound in Theorem 1.1 is constructive in the sense that it is derived by analyzing a polynomial time algorithm that typically finds a contagious set of size at most $O \left(\frac{n}{d^{\frac{r}{r-1}} \log d} \right)$.

Clearly it is always the case that $m(G, r) \geq r$. We examine how large p needs to be in order for $G(n, p)$ to satisfy that typically $m(G, r) = r$. The property of having a contagious set of size r is a monotone property, hence it has a sharp threshold [13]. We determine this threshold up to constant multiplicative factors:

Theorem 1.2 *Let $G \sim G(n, p)$ and suppose $r \geq 2$ is an integer. There exist $0 < c < C$, such that the following holds: if $p < \frac{c}{(n \log^{r-1} n)^{1/r}}$, then with high probability no set of size r is contagious. If $p > \frac{C}{(n \log^{r-1} n)^{1/r}}$, then with high probability there is a contagious set of size r. Moreover, with high probability there is a choice of a contagious set B_0 of size r for which $\tau(B_0) = O(\log \log n)$. This upper bound on $\tau(B_0)$ is best possible up to constant factors – with high probability there is no contagious set B of size r with $\tau(B) = o(\log \log n)$, as long as $p = o(n^{-1/r})$.*

1.2. *Related work.* Bootstrap percolation was introduced by Chalupa, Leath and Reich [15], motivated by applications in statistical physics. Other early works include [1, 27]. Initially, the study of bootstrap percolation focused mostly on lattices and grids. More recently, it has been studied on other families of graphs such as random d-regular graphs [9, 20], hypercubes [5] and several models of random graphs with a given degree sequence (e.g.,

4

[2, 3]). These works studied the case in which the set of seeds is selected independently at random. The smallest contagious set (the value of $m(G, r)$) was studied for some families of graphs such as hypercubes [5, 23] and grids [8].

The critical size of a *random* set needed for full activation (with high probability) of the binomial random graph $G(n, p)$ was first studied in [28]. The results in [28] were generalized and extended by [21] (using ideas from [25]), where the critical size of a random set required for complete activation of $G(n, p)$ for arbitrary constant threshold r is determined in great detail of precision. We shall apply the following theorem from [21] (which follows from Theorem 3.1, page 1996, and Theorem 3.10, page 2000, in [21]).

Theorem 1.3 *Let $r \geq 2$ be a fixed integer independent of n. Suppose $G \sim G(n, p)$ with $n^{-1} \ll p \ll n^{-1/r}$. Let*

$$a_c := \left(1 - \frac{1}{r}\right) \cdot \left(\frac{(r-1)!}{np^r}\right)^{1/(r-1)}.$$

Consider a fixed set A of vertices that are activated as seeds. Then for every fixed $\delta > 0$, with high probability the following holds:

1. *If $|A| = (1 + \delta)a_c$ then at least $n - O(n(pn)^{r-1}e^{-pn})$ vertices will be activated. Furthermore, $\tau(A) = \frac{\ln \ln(np)}{\ln r} + \frac{\ln n}{np} + O(1)$.*

2. *If $|A| \leq (1 - \delta)a_c$ then at most $2\left(\frac{(r-1)!}{np^r}\right)^{1/(r-1)}$ vertices will be activated.*

For example, Theorem 1.3 implies that when $G \sim G(n, p)$ with p as above, then with high probability $m(G, 2) \leq \frac{1+\delta}{2np^2}$. (Observe that $n^2 pe^{-pn} = o(n/d^2)$ for the range of p in Theorem 1.3, and hence the set of vertices not infected by A is small and can be added to the set of seeds with only negligible effect on the total number of seeds.) To the best of our knowledge, the upper bound $m(G, r) \leq (1 + \delta)a_c$ was the best upper bound known on $m(G, r)$ in random graphs prior to our work.

The lower bound of Theorem 1.3 implies that a *randomly chosen* set of $(1 - \delta)a_c$ vertices has only negligible probability of being contagious. Our upper bound in Theorem 1.1 (whose proof involves a more sophisticated choice of set of seeds) implies that for such graphs $m(G, r)$ is with high probability significantly smaller than a_c. This shows that choosing an initial set of seeds carefully (rather than uniformly at random) is typically beneficial for this key model of random graphs.

It is proven in [21] that when $p \gg n^{-1/r}$, an arbitrary set of size r of activated vertices will activate the whole of $G(n, p)$ w.h.p. Similarly to The-

orem 1.1, Theorem 1.2 demonstrates that a careful choice of the seeds results in a contagious set of size r for p much smaller than $n^{-1/r}$.

Theorem 1.1 and the constructive nature of the upper bound there imply that there is a polynomial time algorithm that for most graphs (from the distribution specified in Theorem 1.1) returns a contagious set whose size is within a constant factor of the minimum possible. In contrast, on worst-case instances, approximating the minimal size of a contagious set within a ratio better than $O(2^{\log^{1-\delta} n})$ (n is the number of vertices) is intractable for every $\delta \in (0, 1)$, unless NP $\subseteq$ DTIME($n^{poly(\log n)}$) [16].

Theorem 1.3 (taken from [21]) considers also τ, the number of generations until complete activation. The parameter τ has been studied also in families of graphs such as grids [11, 12] and dense graphs [19]. We consider τ in the context of Theorem 1.2 but not in the context of Theorem 1.1. We briefly discuss τ further in Section 5.

The minimal number of edges that forces an n-vertex graph to satisfy $m(G, r) = r$ was considered in [19]. For example, it is proven that a graph having at least $\binom{n-1}{2} + 1$ edges must satisfy $m(G, 2) = 2$. This result is tight, as $m(G, 2) = 3$ (for $n \geq 3$) when G is a clique on $n - 1$ vertices along with an additional isolated vertex.

The current paper is one part of a larger body of work whose preliminary version is available in [14]. Other parts of that work will be published separately, and they concern contagious sets in d-regular graphs. For example, it is shown there that sufficiently strong expansion properties (e.g., spectral gap $d - O(\sqrt{d})$, or girth $\Omega(\log \log d)$) ensure that $m(G, 2) \leq O(n/d^2)$, where $n = |V(G)|$. (Recall in contrast that the best general upper bound for the value of $m(G, 2)$ for d-regular graphs on n vertices is $m(G, 2) \leq \frac{2n}{d+1}$ [24]; this bound is easily seen to be tight.) In addition, it is shown that when G is a random d-regular graph over n vertices (which with high probability is an excellent spectral expander, see [17]), it holds that $m(G, 2) \geq \Omega(\frac{n}{d^2 \log d})$ with high probability. That lower bound regarding random d-regular graphs is established using ideas similar to those used to establish the lower bound in Theorem 1.1.

1.3. *Overview of proof techniques.* The proof of the upper bound in Theorem 1.1 is based on the following observation (we consider $r = 2$ throughout this section – similar reasoning applies for $r > 2$). For a subset $A \subseteq V$, we denote by $N(A)$ the set of all vertices in $V \setminus A$ having a neighbor in A. Suppose we have an initial set A of seeds, and consider $N(A)$. Given that the graph is random, one can analyze the distribution of the sizes of the connected components of the subgraph induced by $N(A)$. Introducing a single

seed in a connected component of size k activates the whole component, thus giving $k-1$ infected vertices per investment of one seed. It turns out that we can activate a set of size $\frac{n}{d^2} = \frac{1}{np^2}$ in G by choosing $O\left(n\frac{\log\log d}{d^2\log d}\right)$ seeds in this way. Thereafter, the results of [21] can be used in order to deduce that G (apart from a set of negligible cardinality which can be activated separately) is activated with high probability.

To achieve the improved upper bound in Theorem 1.1, we repeat the procedure above iteratively. In iteration 0, choose an arbitrary set A_0 of seeds of size $\frac{n}{d^2\log d}$. Next, for each $1 \le i \le \log\log d$, consider the external neighborhood of the vertices activated in iteration $i-1$. Within this neighborhood identify the largest connected component, and activate a set B_i that includes one vertex from each component (thus activating the whole component), until the sum of sizes of activated components reaches $\frac{2^i n}{d^2\log d}$. After $\log\log d$ iterations we have $\frac{n}{d^2}$ active vertices, which as previously noted suffices to activate the whole of G (apart from a set of negligible cardinality treated separately). The total number of activated vertices is $|A_0| + \sum_{i=1}^{\log\log d}|B_i|$. We show that the latter sum is bounded by $O(\frac{n}{d^2\log d})$, with high probability.

Our lower bound in Theorem 1.1 is based on observing that if there is a contagious set of size t_0, then adding to it the first $t-t_0$ infected vertices gives an induced subgraph with t vertices and at least $2(t-t_0)$ edges. For a choice of $t_0 < \frac{n}{6d^2\log d}$ and $t = \frac{n}{3d^2}$, a simple probabilistic argument shows that a random graph with high probability does not contain any such subgraph.

For Theorem 1.2, the proof of the lower bound on the threshold probability for $m(G,2) = 2$ follows the same principles as the lower bound for Theorem 1.1 (but with $t_0 = 2$). For the proof of the upper bound (the typical existence of a contagious set of size 2 when $p \ge \frac{C}{\sqrt{n\log n}}$) we represent G as a union of two random graphs G_1 and G_2 with edge probabilities $\frac{1}{\sqrt{n\log n}}$ and $\frac{C_1}{\sqrt{n\log n}}$, respectively. We first show that in G_1, a random set of two vertices has probability significantly higher than $\frac{\log n}{n}$ of infecting $\Omega(\log n)$ additional vertices. We then show that this implies that with high probability, there is at least one pair of vertices that infects a set S of size $\Omega(\log n)$ in G_1. Finally, the results of [21] are used to prove that with high probability S will activate the whole of G_2, and thus the whole of G.

1.4. *Preliminaries and notation.* Let $H = (V, E)$ be an undirected graph. For $A, B \subseteq V$, we define $E(A)$ to be the set of all edges spanned by A and $E(A, B)$ the set of all edges with one endpoint in A and one endpoint in B. The notation log denotes logarithms in base 2 and ln denotes natural logarithms. The set of integers $\{1\ldots\ell\}$ for $\ell \ge 1$ is denoted by $[\ell]$. We

will reserve the notation G for $G(n, p)$ throughout this paper, omitting the dependency on n, p when clear from the context.

We shall use the term *infected* vertex to describe an activated vertex that is not one of the *seeds*, but has rather become activated by having at least 2 active neighbors. We say a set $S \subseteq V$ is *activated* if every vertex of S becomes active (either as a seed, or by becoming infected).

We close this section with a version of Chernoff's inequality (see, e.g., [22]).

Lemma 1.1 *Suppose that $X = \sum_{i=1}^{m} X_i$, where every X_i is a $\{0, 1\}$-random variable with $\Pr(X_i = 1) = p$ and the X_is are jointly independent. Then for arbitrary $\eta \in (0, 1)$, it holds that*

$$\Pr(X < (1 - \eta)pm) \leq \exp(-pm\eta^2/2),$$

and

$$\Pr(X > (1 + \eta)pm) \leq \exp(-pm\eta^2/3).$$

1.5. *Organization.* We first present our results when $r = 2$ as this case is more transparent, making it easier to present the main ideas behind the proofs. In Section 2 we prove that with high probability $m(G, 2) = \Theta(\frac{n}{d^2 \log d})$, dealing first with the upper bound and then establishing a lower bound. In Section 3 we determine the asymptotic threshold of having a contagious set of size 2. In Section 4 we discuss how to generalize the results of Sections 2 and 3 to the case where $r > 2$. In Section 5 we present some concluding comments.

2. $m(G, 2)$ in random graphs. In this section we prove Theorem 1.1 for the case $r = 2$. Unless explicitly stated, we will always focus on $G(n, p)$ where $p := \frac{d}{n}$ is as in the range of Theorem 1.1.

2.1. *Upper bound.* The following lemma can be derived from known results (e.g., [10]) but we present a self-contained proof for completeness.

Lemma 2.1 *Let $H := G(n_0, q)$ be the binomial random graph with n_0 vertices and edge probability q (we assume n_0 is large enough). Let $k = O(\log n_0)$ be an integer and $q = \frac{c}{n_0}$. Then for every $c < 1/20$, the probability a given vertex v belongs to a connected component of size at least k is at least $\left(\frac{c}{3}\right)^{k-1}$. Furthermore, with probability at least $1 - \exp\left(-\Omega\left(\frac{n_0}{k}\left(\frac{c}{3}\right)^{k-1}\right)\right)$ the number of vertices lying in components of size at least k is at least $\left(\frac{c}{3}\right)^{k-1} \cdot n_0/4$.*

Proof: Consider the following iterative procedure of exposing edges in H. Every vertex has a mark: either it is *used* or it is *unused*. In the beginning of the algorithm, all vertices are marked as unused and all edges of H are not exposed. We continue the process as long as the number of unused vertices is at least $\frac{n_0}{2} + k$ (or, differently put, the number of used vertices is at most $\frac{n_0}{2} - k$).

In the beginning of every iteration, we choose an unused vertex v in H, and attempt to expose a simple path in H containing v as follows. Consider a set S of exactly $n_0/2$ unused vertices. Expose all edges between v and S. If there is a vertex $v_2 \in S$ connected to v, add v_2 to the path. Continue in this fashion (attaching a vertex to the last vertex appended to the path) until either one of two cases occurs: a *success*, meaning that the size of the path containing v reaches k, or a *failure*, meaning that we have failed to find a path of k vertices containing v (namely, we constructed a path P of $l < k$ vertices, and the last vertex on the path has no edge to any of the vertices of the corresponding set S). Finally, proceed by marking all the vertices that are in the path rooted at v as used.

The probability that all vertices in a set U of $n_0/2$ unused vertices are not connected to a vertex $w \notin U$ is $(1 - q)^{n_0/2}$. It follows that the probability that during an iteration we can append a new vertex to a path of length smaller than k is

$$1 - (1 - q)^{n_0/2} \geq \frac{n_0 q}{3},$$

(here we use our assumption on q, c) and this holds *independently* of the length of the path we have constructed thus far. Hence the probability we succeed in growing a path of length k (and hence in a connected component of size at least k) at a given iteration is at at least

$$\left(\frac{n_0 q}{3}\right)^{k-1} = \left(\frac{c}{3}\right)^{k-1}.$$

As $\lfloor n_0/(3k) \rfloor \leq \frac{\frac{n_0}{2} - k}{k}$ (recall we assume $k = O(\log n_0)$), it follows that the distribution of the number of successes (until less than $\frac{n_0}{2} + k$ unused vertices remain) stochastically dominates the binomial distribution with $\lfloor n_0/(3k) \rfloor$ trials and success probability $\left(\frac{c}{3}\right)^{k-1}$ (the exact number of trials depends on the number of failures, but failures only increase the number of trials). Furthermore, standard concentration results concerning the binomial distribution imply that probability at least $1 - \exp\left(-\Omega(\frac{n_0}{k} \left(\frac{c}{3}\right)^{k-1})\right)$, the number of successes is at least $\frac{n_0}{4k} \left(\frac{c}{3}\right)^{k-1}$. Since every success places k vertices (rather than just one) in a component of size at least k the lemma is proven. $\square$

We shall also rely on the following lemma.

Lemma 2.2 *Let the activation threshold be $r = 2$ and let d_0 be a sufficiently large constant. Then for $d = d(n) \geq d_0$, a random graph $G \sim G(n, p)$ is w.h.p. such that activating any set of size at least $n/2$ activates all but at most n/d^3 vertices.*

Proof: Let A_0 be an initially activated set, and let $U = [n] - \langle A_0 \rangle$. Then $|U| \leq n/2$, and every vertex of U has at most one neighbor outside of U, implying that the number of edges crossing between U and its complement is at most $|U|$. The probability of having such a set U of cardinality $|U| \geq n/d^3$ in $G(n, p)$ can be estimated from above through Lemma 1.1 as follows:

$$\sum_{k=max\{1, n/d^3\}}^{n/2} \binom{n}{k} Pr[\text{Bin}(k(n-k), p) \leq k] \leq \sum_{k=max\{1, n/d^3\}}^{n/2} \binom{n}{k} e^{-\frac{knp}{8}}$$

$$\leq \sum_{k=\max\{1, n/d^3\}}^{n/2} \left(\frac{en}{k}\right)^k e^{-\frac{knp}{8}}$$

We now distinguish between two cases: if $d \leq n^{1/3}$ then we upper bound the summation above by

$$\sum_{k=n/d^3}^{n/2} (ed^{1/3} \cdot e^{-\frac{np}{8}})^k = o(1).$$

Otherwise, if $d > n^{/3}$ the summation can be upper bounded by

$$\sum_{k=1}^{n/2} (en \cdot e^{-n^{2/3}/8})^k = o(1),$$

as desired.

□

Theorem 2.1 *If $d = np$ satisfies $1 \ll d \ll \left(\frac{n \log \log n}{\log^2 n}\right)^{1/2}$ and $G \sim G(n, p)$, then whp $m(G, 2) \leq \frac{13n}{d^2 \log d}$.*

We give a constructive proof for Theorem 2.1. Namely, we provide an algorithm that finds a contagious set that is not larger than the upper bound in this Theorem. Our algorithm is composed of three stages described below.

Stage I. Set
$$\ell = \log \log d \,.$$

Initialize $B_0 = C_0 = D_0$ to be a fixed subset of $[n]$ of size $\frac{n}{d^2 \log d}$.

For $i = 1, \ldots, \ell$ repeat:

Set
$$s_i = \frac{\log d - 4}{\ell - i + 4} \,.$$

Step (i1). Expose edges of G between C_{i-1} and $V \setminus \bigcup_{j=0}^{i-1} B_j$. Let B_i be an arbitrary set of $\frac{d|C_{i-1}|}{2}$ neighbors of C_{i-1} in $V \setminus \bigcup_{j=0}^{i-1} B_j$. If there is no such set – declare a failure;

Step (i2). Expose edges of G inside B_i. Let x_i be the number of connected components of $G[B_i]$ and define $y_i = \min \left\{ x_i, \frac{n}{d^2 2^{\ell-i} s_i} \right\}$. Let $T_{i1}, \ldots, T_{iy_i}$ be the y_i largest components of $G[B_i]$ (breaking ties arbitrarily). If

$$\left| \bigcup_{j=1}^{y_i} T_{ij} \right| < \frac{n}{d^2 2^{\ell-i}}$$

– declare a failure. Otherwise form D_i by choosing one arbitrary vertex from each T_{ij}. Clearly

$$|D_i| = y_i \leq \frac{n}{d^2 2^{\ell-i} s_i} \,.$$

Let C_i be an arbitrary subset of $\bigcup_{j=1}^{y_i} T_{ij}$ of size

$$|C_i| = \frac{n}{d^2 2^{\ell-i}} \,.$$

Assume that Stage I was successful for every $1 \leq i \leq \ell$. Denote

$$A_{01} = D_0 \cup D_1 \cup \ldots \cup D_\ell \,.$$

The algorithm activates all vertices in A_{01}. Finally let A_{02} the set of vertices that remain inactive after A_{01} is activated. We activate all vertices in A_{02} We now prove a series of propositions that upper bound the size of $A_{01} \bigcup A_{02}$.

Proposition 2.1 *Activating A_{01} activates $\bigcup_{j=0}^{\ell} C_j$.*

Proof: We prove by induction that activating $D_0 \cup \ldots D_i$ activates $\bigcup_{j=0}^{i} C_j$. Induction basis follows from the definition of C_0, D_0. For the induction step, assume that C_{i-1} is already activated. Recall that each vertex in $T_{i1}, \ldots, T_{iy_i}$ has a neighbor in C_{i-1} by the definition of B_i. Activating in addition the

vertex v, where $\{v\} = D_i \cap T_{ij}$, activates all of T_{ij}, implying that all of C_i gets activated. $\square$

Let us now estimate the probability of failure of each round of Stage I, and see what it delivers assuming its success. Notice first that for all $i \in [l]$ we have that $|C_i| \leq \frac{n}{d^2}$, $|B_i| \leq \frac{d}{2}|C_{i-1}|$, implying that during Stage I the union $B_0 \cup B_1 \cup \ldots$ always has cardinality at most $(\ell + 1) \cdot d \cdot n/d^2 \leq n/10$.

Next observe that if there does not exist an index $i \in [\ell]$ for which a failure occurs in steps $i1, i2$, then by the definition of the algorithm above we have that for every i the following equalities hold : $|C_i| = \frac{n}{d^2 2^{\ell-i}}$ and $|B_i| = \frac{n}{d \cdot 2^{\ell-i+2}}$ which implies that, $|D_i| \leq \frac{n}{d^2 2^{\ell-i} s_i}$ holds for every $i \in [\ell]$ as well. Indeed, observe that if $x_i < \frac{n}{d^2 2^{\ell-i} s_i}$, then the union of T_{ij} is the whole set B_i, and thus all of B_i will be activated. Otherwise $y_i = \frac{n}{d^2 2^{\ell-i} s_i}$. If the y_i largest components of $G[B_i]$ do not contain all vertices in components of size at least s_i, then $|T_{ij}| \geq s_i$ for $j = 1, \ldots, y_i$, implying $|\bigcup_{j=1}^{y_i} T_{ij}| \geq \frac{n}{d^2 2^{\ell-i}}$; in the opposite case we also have the same outcome. Thus Step (i2), if successful, results indeed in a subset C_i of cardinality $|C_i| = \frac{n}{d^2 2^{\ell-i}}$, as declared.

Proposition 2.2 *With high probability there is no $i \in [\ell]$ such that step $i1$ fails.*

Proof: For Step (i1), the probability that C_{i-1} has less than $\frac{d}{2}|C_{i-1}|$ neighbors outside of $\bigcup_{j=0}^{i-1} B_j$ is bounded from above by

$$Pr\left[\mathrm{Bin}\left(\frac{9n}{10}, 1 - \left(1 - \frac{d}{n}\right)^{|C_{i-1}|}\right) \leq \frac{d|C_{i-1}|}{2}\right] = \exp\left\{-\Theta(d|C_{i-1}|)\right\}$$
$$= \exp\left\{-\Theta\left(\frac{n}{d \cdot 2^{\ell-i}}\right)\right\},$$

and the sum of these estimates for $i = 1, \ldots, \ell$ is obviously $o(1)$. $\square$

Proposition 2.3 *With high probability, there is no $i \in [\ell]$ such that step $i2$ fails.*

Proof: Apply Lemma 2.1 with parameters

$$n_0 = |B_i| = \frac{n}{d \cdot 2^{\ell-i+2}}, \quad q = \frac{d}{n} = \frac{1}{2^{\ell-i+2}n_0}, \quad k = s_i\,.$$

We derive that with probability $1 - \exp\left\{-\Omega\left(\frac{n_0}{s_i}\left(\frac{1}{2^{\ell-i+4}}\right)^{s_i}\right)\right\}$ the set B_i has at least $\left(\frac{1}{3 \cdot 2^{\ell-i+2}}\right)^{s_i} \frac{n}{d \cdot 2^{\ell-i+4}}$ vertices in connected components of size at

least s_i. The (absolute value of the) exponent in the exceptional probability above can be estimated as follows:

$$
\begin{aligned}
\frac{n_0}{s_i}\left(\frac{1}{2^{\ell-i+4}}\right)^{s_i} &= \frac{n(\ell-i+4)}{2^{\ell-i+2}d(\log d - 4)} \cdot 2^{-\frac{(\ell-i+4)(\log d - 4)}{\ell-i+4}} \\
&\geq \frac{4n}{d^2\log d}\cdot\frac{\ell-i}{2^{\ell-i}}.
\end{aligned}
$$

Set $K := \frac{4n}{d^2\log d}$, and observe that the requirement $d \ll \left(\frac{n\log\log n}{\log^2 n}\right)^{1/2}$ implies that $K \gg \frac{\log n}{\log\log n}$. By the calculations above, we can upper bound the probability there is failure in one of the rounds by

$$
\sum_{j=1}^{\ell}\exp\left\{-\frac{K\cdot j}{2^j}\right\}.
$$

Denoting the jth summand by $f(j)$ we see that $\frac{f(j+1)}{f(j)} = \exp(\frac{K}{2^j}\frac{j-1}{2})$ which tends to infinity with n for all $1 \leq j \leq \ell = \log\log d$. Therefore $\sum_{j=1}^{\ell}f(j) = \Theta(f(\ell)) = \exp\left\{-\Theta(\frac{n}{d^2\log d}\cdot\frac{\log\log d}{\log d})\right\}$. Hence recalling our assumed upper bound on $d(n)$, the union bound implies that except for probability $o(1)$, step $(i2)$ is completed for every $i \in [\ell]$. Given that there are no failures in step $(i2)$, the number of vertices of B_i in components of size at least s_i is at least

$$
\begin{aligned}
\left(\frac{1}{3\cdot 2^{\ell-i+2}}\right)^{s_i}\frac{n}{d\cdot 2^{\ell-i+4}} &\geq \left(\frac{1}{2^{\ell-i+4}}\right)^{s_i}\frac{n}{d\cdot 2^{\ell-i+4}} = 2^{-\log d + 4}\cdot\frac{n}{d\cdot 2^{\ell-i+4}} \\
&= \frac{n}{d^2\cdot 2^{\ell-i}}.
\end{aligned}
$$

$\square$

Now we estimate the size of the set $A_{01} = \bigcup_{i=0}^{\ell} D_i$. Recall that $|D_0| = \frac{n}{d^2\log d}$, and using Propositions 2.2 and 2.3 we get that with probability $1 - o(1)$, $|D_i| \leq \frac{n}{d^2 2^{\ell-i}s_i} = \frac{n(\ell-i+4)}{d^2 2^{\ell-i}(\log d - 4)}$. It thus follows that

$$
\begin{aligned}
|A_{01}| &= \left|\bigcup_{i=0}^{\ell} D_i\right| \leq \frac{n}{d^2\log d} + \frac{n}{d^2(\log d - 4)}\sum_{i=1}^{\ell}\frac{\ell-i+4}{2^{\ell-i}} \\
&= \frac{n}{d^2\log d} + \frac{n}{d^2(\log d - 4)}\left[4\sum_{i=1}^{\ell}2^{-\ell+i} + \sum_{i=1}^{\ell}\frac{\ell-i}{2^{\ell-i}}\right].
\end{aligned}
$$

Obviously $\sum_{i=1}^{\ell} 2^{-\ell+i} < 2$. Also, $\sum_{i=1}^{\ell} \frac{\ell-i}{2^{\ell-i}} \leq \sum_{j=1}^{\infty} \frac{j}{2^j} = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \frac{1}{2^j} = \sum_{i=1}^{\infty} \frac{2}{2^i} = 2$. Altogether,

$$|A_{01}| = \left| \bigcup_{i=0}^{\ell} D_i \right| \leq \frac{n}{d^2 \log d} + \frac{n}{d^2 (\log d - 4)}(8+2) \leq \frac{12n}{d^2 \log d}\,.$$

To complete the analysis of Stage I observe that assuming it was successful, the set C_ℓ is activated, and no edges between C_ℓ and $V - \bigcup_{i=0}^{\ell} B_i$ and inside $V - \bigcup_{i=0}^{\ell} B_i$ have been exposed.

Stage II. Denote $G_2 = G[C_\ell \cup \left(V - \bigcup_{i=0}^{\ell} B_i \right)]$. We can view G_2 as a random graph with edge probability p, in which the initial seed C_ℓ of size $|C_\ell| = \frac{n}{d^2}$ is activated. Then according to Theorem 1.3, w.h.p. all but $O(nde^{-d/2}) < \frac{n}{4}$ vertices of G_2 will be infected. Recalling that $\left| \bigcup_{i=0}^{\ell} B_i \right| \leq \frac{n}{10}$, we arrive at the conclusion that w.h.p. after Stage II at least $n/2$ vertices of G are infected, when activating the initial seed A_{01}.

Stage III. According to Lemma 2.2 above, the random graph $G \sim G(n,p)$ is w.h.p. such that activating any set of size $n/2$ results in all but at most n/d^3 vertices being activated. Apply this Lemma to the outcome of Stage II, and denote by A_{02} the set of vertices that remain unactivated, w.h.p $|A_{02}| \leq n/d^3$. Define

$$A_0 = A_{01} \cup A_{02}\,,$$

then $|A_0| \leq \frac{12n}{d^2 \log d} + \frac{n}{d^3} < \frac{13n}{d^2 \log d}$, and $\langle A_0 \rangle = [n]$.

2.2. *Lower bound.* In our analysis, we shall include two parameters α and β that can simultaneously be optimized to give the best possible lower bound provable with our current approach. For simplicity of the presentation, rather than optimizing α and β, we shall fix $\alpha = 3$ and $\beta = 2 - \frac{1}{\log d}$.

Let G be a random graph sampled from $G(n,p)$. Let $t = \frac{n}{3d^2}$. We assume that d is bounded from below by some sufficiently large constant (that can be computed explicitly from the proof of Lemma 2.3), and bounded from above by $o(\sqrt{n})$.

Lemma 2.3 *For the setting above, w.h.p. G does not have a subgraph with $t = \frac{n}{3d^2}$ vertices and βt edges, where $\beta = 2 - \frac{1}{\log d}$.*

Proof: There are $\binom{n}{t} \leq (3ed^2)^t$ possible choices of a set T of t vertices in G. There are $\binom{\binom{t}{2}}{\beta t} \leq (\frac{et}{2\beta})^{\beta t}$ ways of choosing βt edge locations in T. The

probability that all these choices are indeed edges is $\left(\frac{d}{n}\right)^{\beta t} = (\frac{1}{3dt})^{\beta t}$. Hence the probability that G has a subgraph with t vertices and βt edges is upper bounded by:

$$(3ed^2)^t \left(\frac{et}{2\beta}\right)^{\beta t} \left(\frac{1}{3dt}\right)^{\beta t} = \left(\frac{e^{\beta+1}d^{2-\beta}}{3^{\beta-1}2^\beta \beta^\beta}\right)^t.$$

Now in the exponent for d substitute $\beta = 2 - \frac{1}{\log d}$, obtaining $d^{2-\beta} = 2$. For the other terms we can substitute an approximation $\beta \simeq 2$ assuming d is large enough. The expression $\frac{e^{\beta+1}d^{2-\beta}}{3^{\beta-1}2^\beta \beta^\beta}$ is then roughly $\frac{2e^3}{48}$ and is strictly smaller than 1. Raising to the power of t, the probability tends to 0 as n grows. $\square$

Corollary 2.1 *For the parameters as above, $m(G, 2) > \frac{n}{6d^2 \log d}$ w.h.p.*

Proof: Suppose otherwise. Then for $t = \frac{n}{3d^2}$, the set of $t_0 = \frac{n}{6d^2 \log d}$ seeds and first $t - t_0$ infected vertices induces a subgraph with t vertices and $2(t - t_0) = (2 - \frac{1}{\log d})t$ edges, contradicting Lemma 2.3. $\square$

3. The asymptotic threshold for $m(G, 2) = 2$.

Lemma 3.1 *Let $p < \frac{c}{\sqrt{n \log n}}$ for some sufficiently small $c > 0$. Then with high probability, $m(G, 2) > 2$.*

Proof: Otherwise, for every $1 \le t \le n - 2$, there are two vertices a,b and a set of t vertices disjoint from $\{a, b\}$, such that the subgraph spanned on $G[\{a, b\} \cup T]$ spans at least $2t$ edges. The probability such a subgraph exists is upper bounded by

$$\binom{n}{2}\binom{n}{t}\binom{(t+2)^2/2}{2t}p^{2t},$$

which (for large t) is at most

$$n^2 \left(en/t\right)^t \left(e(t+2)p\right)^{2t} \le n^2(25c)^t = o(1),$$

when $t \ge \log n$ and c is sufficiently small. $\square$

We will now prove that if $p = \frac{C}{\sqrt{n \log n}}$ and C is large enough, then w.h.p. $G \sim G(n, p)$ satisfies $m(G, 2) = 2$.

Lemma 3.2 *Let $X \sim \mathrm{Bin}(n, p)$ with $np \le 1$. Then $Pr[X > 0] > \frac{np}{2}$.*

Proof: By Bonferroni's Inequality,

$$Pr[X > 0] \geq np - \binom{n}{2}p^2 = np\left(1 - \frac{(n-1)p}{2}\right) > \frac{np}{2}\,.$$

□

We expose $G(n,p)$ in two stages: $G = G_1 \cup G_2$, where $G_i \sim G(n,p_i)$, $p_1 = \frac{1}{\sqrt{n \log n}}$, $p_2 = \frac{C_1}{\sqrt{n \log n}}$ with C_1 being a large enough constant, to be set later. We will argue that w.h.p. G_1 contains two vertices u_1, u_2 infecting a set U of size $k = \Theta(\log n)$. Then we will use G_2 and Theorem 1.3 to argue that if C_1 is large enough then with high probability the set U activates all of V in G_2 and thus in G.

Lemma 3.3 *Let $k = c_1 \log n$, where $0 < c_1 < 1$ is a small enough constant. Let G_1 be distributed as $G\left(n, \frac{1}{\sqrt{n \log n}}\right)$. Then with high probability there are two vertices in G_1 that activate a set of size k.*

Proof: Initialize $V_0 = V = [n]$. We describe an algorithm that has at most $\frac{n}{2k}$ iterations, indexed by $i = 1, \ldots, \frac{n}{2k}$. Every iteration has at most $k - 2$ steps, indexed by $j = 3, \ldots, k$. We now describe iteration i.

Let u_1, u_2 be arbitrary vertices of V_0. For simplicity of the proof (and at the expense of requiring a smaller constant c_1 in the statement of the lemma), partition $V_0 - \{u_1, u_2\}$ into $k - 2$ sets $U_{i,1}, \ldots, U_{i,k-2}$, each of size at least $\lfloor \frac{|V_0|-2}{k-2} \rfloor$. For $j = 3, \ldots, k$, if there is a vertex $v_j \in U_{i,j-2}$ with at least two neighbors in $\{u_1, u_2, \ldots, u_{j-1}\}$, then set $u_j := v_j$. Otherwise, abort iteration i, dump $\{u_1, \ldots, u_{j-1}\}$, update $V_0 := V_0 - \{u_1, \ldots, u_{j-1}\}$, and move to iteration $i + 1$.

Observe crucially that during iteration i we have only exposed edges of G_1 touching $\{u_1, \ldots, u_{j-1}\}$, so the rest (i.e. edges whose both endpoints belong to V_0) are not exposed and fully retain their randomness. Also, at each iteration we dump less than k vertices; since we perform at most $\frac{n}{2k}$ iterations, the size of V_0 is always at least $n/2$.

Let us now estimate the probability that the i-th iteration succeeds. When looking for u_j inside this iteration, the probability that such a vertex is found is at least the probability that there is a vertex in a set of size $n/2k$ having at least 2 neighbors in a set of size $j - 1$, where the edge probability is p_1. The probability for a given vertex of $U_{i,j}$ to have at least two neighbors in the set of size $j - 1$ can be estimated from below by $\binom{j-1}{2}\frac{p_1^2}{2} > \frac{(j-2)^2 p_1^2}{4}$. Thus the probability that there is a required vertex in $U_{i,j}$ is at least $Pr[\text{Bin}\left(\frac{n}{2(k-2)}, \frac{(j-2)^2 p_1^2}{4}\right) > 0]$, and the latter is at least

$\frac{n(j-2)^2 p_1^2}{16k}$ by Lemma 3.2, as $\frac{n}{2(k-2)} \cdot \frac{(j-2)^2 p_1^2}{4} < 1$ (recall we assume $k = c_1 \log n$ where c_1 is sufficiently small); this estimate is valid independently of what happened in the current iteration. Thus the probability that the i-th iteration succeeds is at least (using Stirling's approximation)

$$\prod_{j=3}^{k} \frac{n(j-2)^2 p_1^2}{16k} = \left(\frac{np_1^2}{16k}\right)^{k-2} \cdot ((k-2)!)^2 \geq \left(\frac{np_1^2(k-2)^2}{16e^2 k}\right)^{k-2} \geq \left(\frac{np_1^2 k}{200}\right)^k.$$

Substituting the expressions for k and p_1, we obtain that the above expression is at least $(c_1/200)^{c_1 \log n}$, and this is more than $\frac{2k \log n}{n}$ for $c_1 > 0$ small enough.

Since we are ready to perform $\frac{n}{2k}$ iterations, with each being successful independently with probability at least $\frac{2k \log n}{n}$, w.h.p. one of them will indeed succeed – resulting in a set of size k which can be activated by two vertices in it. $\quad \square$

The equality $m(G, 2) = 2$ now follows from Theorem 1.3. Namely, in G_2 there is an active set S of cardinality $c_1 \log n$ (generated by choosing two "correct" vertices to start the process in G_1). Hence when $p_2 = \frac{C_1}{\sqrt{n \log n}}$ for $C_1 > \frac{1}{\sqrt{c_1}}$, we get that with high probability S is a contagious set in G_2.

We now deal with the number of generations. We first consider the upper bound. To this end, we upper bound the number of generations until activation of the contagious set constructed in Lemma 3.3. We analyze first the number of generations it takes to activate k vertices in the infection process occurring in G_1. For this, consider the following random directed graph which we denote by $H_{2,k}$. There are k vertices numbered from 1 to k. Each vertex $i \geq 3$ has two outgoing arcs to two random vertices of index less than i. It is implicit in the proof of Lemma 3.3, that the length of the longest directed path of $H_{2,t}$ is an upper bound on the number of generations, which we denote by $l(H_{2,k})$. The parameter $l(H_{2,k})$ was studied in several previous works (e.g., [4, 26]) and shown to be of order $\Theta(\log k)$. Here we present a simple self contained proof that $l(H_{2,k}) = O(\log k)$ (the leading constant in the O-term in our proof is not optimal).

Lemma 3.4 *With high probability, $l(H_{2,k})$ is at most $40 \log k$.*

Proof: Let $0 < \rho < 1$ be a constant to be optimized later. Call an arc (i, j) in $H_{2,k}$ *good* if $j \leq \rho i$ and *bad* if $j > \rho i$ (note that necessarily $j < i$). A path can have at most g good arcs, where g is largest number satisfying $k\rho^g \geq 1$. Given a vertex i, the probability that a random outgoing arc is bad is at most $(1 - \rho)$. For arbitrary $t \geq 2g$, let us upper bound the probability

that there is a path of length t. There are (less than) k possible starting points. From each vertex there are two outgoing arcs to choose. So the number of candidate paths is at most $k2^t$. For each candidate path, there are $\sum_{i=1}^{g} \binom{t}{i} \leq 2^t$ possible locations for the good arcs. For the rest of the arcs to be bad, the probability is at most $(1 - \rho)^{t-g} \leq (1 - \rho)^{t/2}$. Hence the probability that some candidate path actually reaches length t is at most $k2^{2t}(1 - \rho)^{t/2}$.

Choose $\rho = \frac{19}{20}$. Then $g \simeq 20 \ln k$, and we can choose $t = 40 \ln k$. For these parameters, $k2^{2t}(1 - \rho)^{t/2} = k\frac{2^{40 \ln k}}{20^{10 \ln k}} = k(\frac{4}{5})^{10 \ln k} = o(1)$. Hence w.h.p. $l(H_{2,k})$ does not exceed $40 \ln k$. $\quad\square$

Lemma 3.4 implies that with high probability the number of generations until B_0 infects a set of size $c_1 \log n$ in G_1 is at most $O(\log k) = O(\log \log n)$. Thereafter, Theorem 1.3 implies that with high probability all the vertices in G_2 are activated within $O(\log \log n)$ generations.

Now we establish a lower bound on the number of generations. We first claim that with high probability no set of size $k \geq \log n$ can infect too many vertices in a single round.

Lemma 3.5 *Suppose that $p \leq \frac{1}{\sqrt{2en}}$. Then, with high probability every set of size $k \geq \log n$ in $G(n, p)$ infects (in one round) a set of size smaller than k^2.*

Proof: Given that a set S of size k is active, the probability that a vertex outside S is infected by S in one round is $Pr(\text{Bin}(k, p) \geq 2) \leq (pk)^2$. Therefore, the probability for a fixed k there is a set of size k infecting k^2 additional vertices in one round is at most

$$\binom{n}{k} \cdot \binom{n}{k^2} (p^2 k^2)^{k^2},$$

Which can be upper bounded (when $k \geq \log n$) by

$$\left(\frac{en}{k}\right)^k \left(\frac{1}{2}\right)^{k^2} = o(1/n).$$

Taking a union bound over all $k \geq \log n$ concludes the proof. $\quad\square$

Lemma 3.5 implies that with high probability, for every contagious set of size $\log n$, the number of generations required to activate $G(n, p)$ for $p \leq \frac{1}{\sqrt{2en}}$ is at least $\log \log n - \log \log \log n$. The same must hold for contagious sets of size 2 (because every contagious set of size 2 is contained in a contagious set of size $\log n$). This concludes the analysis of the number of generations and the proof of Theorem 1.2 (for the case $r = 2$).

18

4. Generalizing the results for $r > 2$. In this section we study the case where the threshold of every vertex is r, where $r > 2$ is a fixed constant. As the proofs are similar to the $r = 2$ case, we sketch the main ideas without going into every detail.

4.1. *The asymptotic value of $m(G, r)$ in $G(n, p)$.* Here we explain that a similar reasoning to the case $r = 2$ implies that in $G(n, p)$, w.h.p. $m(G, r) \sim \frac{n}{d^{\frac{r}{r-1}} \log d}$. We begin by discussing the upper bound. First we have the following lemma.

Lemma 4.1 *Let $r > 2$ be a fixed integer (independent of n). For $d \geq 100r$, a random graph $G \sim G(n, p)$ is w.h.p. such that activating any set of size at least $n/2$ activates all but at most n/d^3 vertices.*

Proof: This follows from the fact that if a set U is disjoint from a set $\langle A_0 \rangle$, then U can have at most $r|U|$ neighbors in $\langle A_0 \rangle$. Using the assumption that $\frac{n}{d^3} \leq |U| \leq n/2$ and choosing $d_0 > 100r$, implies the lemma along similar lines to Lemma 2.2 – details omitted. $\square$

Our goal is to have the initial set of seeds infect a set of size $\frac{C_1 n}{d^{\frac{r}{r-1}}}$, where C_1 is large enough. Then by applying Theorem 1.3, we conclude that with high probability at least $n/2$ vertices are activated. Finally, using Lemma 4.1 we will be able to deduce that with high probability all of G becomes activated. To achieve this goal, it suffices to make some modest changes to the algorithm presented in Theorem 2.1. The main difference is that now we look in the ith iteration for large connected components in the set of all vertices having $r - 1$ neighbors in C_{i-1}. As in the proof of Theorem 2.1 we initially set $\ell = \log \log d$ and $B_0 = C_0 = D_0$ to be a fixed subset of $[n]$ of size $\frac{C_1 n}{d^{\frac{r}{r-1}} \log d}$. The algorithm is ran for ℓ iterations, where ℓ drops by 1 in every iteration – terminating once $\ell = 0$. Specifically,

Iterating: We aim to get a set C_i of size

$$|C_i| = \frac{C_1 n}{d^{\frac{r}{r-1}} 2^{\ell-i}} .$$

Given a set C_{i-1}, we find a subset B_i of vertices in $V \setminus \bigcup_{j=0}^{i-1} B_j$, all having at least $r - 1$ neighbors in C_{i-1}, and

$$b_i := |B_i| = \frac{C_1^{r-1}}{2^{(\ell-i+1)(r-1)} \cdot d \cdot (r-1)^{r-1}} \frac{n}{2} =: n_0 .$$

Since the probability a vertex (disjoint from C_{i-1}) has at least $r - 1$ neighbors in C_{i-1} is asymptotically equal to $\binom{|C_{i-1}|}{r-1} p^{r-1} \geq \left(\frac{|C_{i-1}|}{r-1} p \right)^{r-1}$, we get

using the Chernoff bound that the size of B_i is indeed lower bounded by $\left(\frac{|C_{i-1}|}{r-1}p\right)^{r-1} 2(n-o(1))/3 > b_i$ with probability at least $1-\exp\left(-O(\frac{n}{d(\log d)^{r-1}})\right)$, where terms depending only on r are treated as constants (as we assume r is a constant not depending on n). It is therefore straightforward to verify that with high probability for all $i \in [\ell]$ it holds that $|B_i| \geq b_i$.

Analogously to the $r = 2$ case, if a set S is connected, and every vertex in S has at least $r - 1$ active neighbors, it suffices to activate a single vertex in S in order to activate the whole of S. Hence we estimate the number of "large" connected components in B_i.

Set

$$s_i = \frac{\log d}{(\ell - i + 1)r^2}$$

Apply Lemma 2.1 to $G[B_i]$ with parameters

$$n_0, \quad q = \frac{d}{n} = \frac{C_1^{r-1}}{2^{(\ell-i+1)(r-1)+1}(r-1)^{r-1}n_0}, \quad k = s_i .$$

We need to verify:

$$\left(\frac{C_1^{r-1}}{3 \cdot 2^{(\ell-i+1)(r-1)+1}(r-1)^{r-1}}\right)^{s_i} \frac{n_0}{4} \geq \frac{C_1 n}{d^{\frac{r}{r-1}}2^{\ell-i}},$$

which amounts to

$$\left(\frac{3 \cdot 2^{(\ell-i+1)(r-1)+1}(r-1)^{r-1}}{C_1^{r-1}}\right)^{s_i} \leq \frac{n_0 d^{\frac{r}{r-1}}2^{\ell-i}}{4C_1 n} = \frac{C_1^{r-2}d^{\frac{1}{r-1}}2^{\ell-i-2}}{2^{(\ell-i+1)(r-1)+1}(r-1)^{r-1}} .$$

This follows from $\left(\frac{6}{C_1^{r-1}}\right)^{s_i} \leq \frac{C_1^{r-2}}{16(r-1)^{r-1}}$ and $2^{(\ell-i+1)(r-1)s_i} \leq \frac{d^{\frac{1}{r-1}}}{2^{(\ell-i+1)(r-2)+4}}$. The former inequality is valid when $C_1 > 6 \cdot 17(r-1)^{r-1}$ and large enough d (e.g., d such that $s_i > 1$). For the latter inequality, we need to satisfy

$$s_i \leq \frac{\frac{1}{r-1}\log d - (\ell - i + 1)(r - 2) - 4}{(\ell - i + 1)(r - 1)}$$

$$= \frac{\log d}{(\ell - i + 1)(r - 1)^2} - \frac{r - 2}{r - 1} - \frac{4}{(\ell - i + 1)(r - 1)}$$

– which is indeed valid for our choice of s_i.

From the calculations outlined in the paragraph above and Lemma 2.1, the probability that $|C_i| < \frac{C_1 n}{d^{\frac{r}{r-1}}2^{\ell-i}}$ is upper bounded by

$$e_i := \exp\left(-\frac{C_1 n}{s_i d^{\frac{r}{r-1}}2^{\ell-i}}\right) = \exp\left(-\frac{C_1 n(\ell - i + 1)r^2 \log d}{d^{\frac{r}{r-1}}2^{\ell-i}}\right).$$

Hence the probability that a failure will occur in one of the rounds is upper bounded by $\sum_{i=1}^{\ell} e_i$. Analogous reasoning to the $r = 2$ case implies that

$$\sum_{i=1}^{\ell} e_i = \exp\left(-\Theta\left(\frac{n \log\log d}{d^{\frac{r-1}{r}} \log^2 d}\right)\right).$$

Substituting $d = o\left(\left(\frac{n \log\log n}{\log^2 n}\right)^{\frac{r-1}{r}}\right)$, we have that the probability there exists $i \in [\ell]$ such that $|C_i| < \frac{C_1 n}{d^{\frac{r}{r-1}} 2^{\ell - i}}$ is $o(1)$.

The total size of the seed of Stage I is then

$$\frac{C_1 n}{d^{\frac{r}{r-1}} \log d} + \sum_{i=1}^{\ell} \frac{C_1 n}{d^{\frac{r}{r-1}} 2^{\ell - i} s_i} = \frac{C_1 n}{d^{\frac{r}{r-1}} \log d} + \frac{C_1 n}{d^{\frac{r}{r-1}} \log d} \sum_{i=1}^{\ell} \frac{\ell - i + 1}{2^{\ell - i}}$$

$$= O\left(\frac{n}{d^{\frac{r}{r-1}} \log d}\right).$$

This concludes the proof that for d satisfying the condition in Theorem 1.1, with high probability $m(G, r) \leq O\left(\frac{n}{d^{\frac{r-1}{r}} \log d}\right)$.

Now we turn to the lower bound. We shall use the following auxiliary Lemma:

Lemma 4.2 Let $\beta = r - \frac{r-1}{\log d}$. Set $t = \frac{n}{3d^{\frac{r}{r-1}}}$ and assume $d = np = o(n)$ is larger than an appropriate constant d_0 (that may depend on r). Then with high probability no set of vertices of size t spans βt edges.

Proof: Using the equality $p = \frac{1}{3d^{\frac{1}{r-1}}t}$ we conclude that the probability that $G(n, p)$ contains a set of size t that spans at least βt edges is upper bounded by

$$\binom{n}{t}\binom{\binom{t}{2}}{\beta t} p^{\beta t} \leq (e\alpha d^{\frac{r}{r-1}})^t \left(\frac{et}{2\beta}\right)^{\beta t} \left(\frac{1}{3d^{\frac{1}{r-1}}t}\right)^{\beta t} = \left(\frac{e^{\beta+1} d^{1/\log d}}{3^{\beta-1} 2^\beta \beta^\beta}\right)^t.$$

It can be verified that the latter expression is $o(1)$. □

Corollary 4.1 Let G be distributed as $G(n, p)$, where $d_0 < d = np \ll n^{1-1/r}$ and d_0 is a large enough constant that may depend on r but not on n. Then with high probability $m(G, r) > \frac{(r-1)n}{3rd^{\frac{r}{r-1}} \log d}$.

Proof: Suppose there exists a contagious set A_0 of size $t_0 = \frac{(r-1)n}{3rd^{\frac{r}{r-1}} \log d}$. Setting $t = \frac{n}{3d^{\frac{r}{r-1}}}$, we get that A_0 together with the first $t - t_0$ infected

vertices, would produce a set A of size $\frac{n}{3d^{\frac{r}{r-1}}}$ spanning at least $r(t - t_0) = (r - \frac{r-1}{\log d})t$ edges. As we have just shown, w.h.p. such a set A does not exist. This concludes the proof. $\square$

4.2. *The threshold for $m(G, r) = r$.* We will show that the threshold for the emergence of a contagious set of size r in $G(n, p)$ is $p \sim (n \log^{r-1} n)^{-1/r}$. We begin by proving that contagious sets of size r are unlikely to exist when $p = c(n \log^{r-1} n)^{-1/r}$, for some appropriate constant $c > 0$.

Lemma 4.3 *Suppose that $p \leq c(n \log^{r-1} n)^{-1/r}$ for some $c > 0$ that is sufficiently small. Then with high probability $m(G, r) > r$, when G is distributed as $G(n, p)$.*

Proof: By a similar reasoning to the case $r = 2$, if $G(n, p)$ has a contagious set of size r then for every $1 < t \leq n - r$ it has a set of $t + r$ vertices spanning at least rt edges. The probability that such a set exists is upper bounded by

$$\binom{n}{r}\binom{n}{t}\binom{(t+r)^2/2}{rt}p^{rt}.$$

For large enough t, the expression above can be upper bounded by

$$n^r \left(\frac{en}{t}\right)^t \left(\frac{e(t+r)^2 p}{2rt}\right)^{rt} < n^r \left(n\frac{(etp/r)^r}{t}\right)^t,$$

where we used the fact that $(t+r)^2 < 2t^2$ for large enough t. Setting $t = \log n$ and substituting the value of p, we can upper bound the expression above by $n^r(c')^t$ for some $c' > 0$ that tends to 0 as $c \to 0$. Hence, taking c to be sufficiently small we can ensure that $(c')^{\log n} < n^{-r+1}$, implying that the probability there exists a contagious set of size r is at most $1/n$. This concludes the proof of the Lemma. $\square$

We now proceed and prove that if $p > C(n \log^{r-1} n)^{-1/r}$ for a suitable constant C, then with high probability there is a contagious set of size r in $G(n, p)$.

Theorem 4.1 *Suppose that $p > C(n \log^{r-1} n)^{-1/r}$, where C is a sufficiently large constant that may depend on r. Let G be distributed as $G(n, p)$. Then with high probability $m(G, r) = r$.*

Proof: It suffices to prove that for $p_1 = (n \log^{r-1} n)^{-1/r}$, a random graph $G_1 \sim G(n, p_1)$ is typically such that activating appropriately chosen r vertices will infect $c_1 \log n$ vertices. Thereafter, exposing the remaining edges of $G = G(n, p)$ with probability $p_2 = C_2(n \log^{r-1} n)^{-1/r}$, where C_2 is a large

enough constant, and using Theorem 1.3 implies that that the whole of G gets activated with high probability.

We use ideas similar to those appearing in Lemma 3.3. Let $k = c_1 \log n$ where $0 < c_1 < 1$ is constant that will be determined later. Run the same iterative procedure as in Lemma 3.3, but choose initially a set I of r vertices from V_0. Now partition $V_0 \setminus I$ to $k - r$ sets each of size at least $\lfloor \frac{|V_0| - r}{k - r} \rfloor$. We now run an iterative procedure identical to the one in Lemma 3.3, but search for a vertex in $v_j \in U_{i, j-r}$ having at least r neighbors in $\{u_1, u_2, \ldots, u_{j-1}\}$ (in the procedure in Lemma 3.3, $r = 2$). If found, set $u_j := v_j$. If such a u_j is not found, we stop iteration j, delete $\{u_1, \ldots, u_{j-1}\}$ and update $V_0 := V_0 - \{u_1, \ldots, u_{j-1}\}$. The probability that the jth iteration succeeds can be lower bounded by $\binom{j-1}{r} \frac{p_1^r}{2} \geq \left(\frac{(j-r)p_1}{r} \right)^r / 2$ (recall that r is a fixed constant and $p = o(1)$). Assuming c to be a sufficiently small constant that may depend on r, we reason that $\Pr[\mathrm{Bin}(\frac{n}{2(k-r)}, \left(\frac{(j-r)p_1}{r} \right)^r / 2) > 0] \geq \frac{n}{2(k-r)} \cdot \left(\frac{(j-r)p_1}{r} \right)^r / 2)$. Therefore, the probability the ith iteration succeeds is at least

$$
\prod_{j=r+1}^{k} \frac{np_1^r}{2r^r k} \cdot (j-r)^r = \left(\frac{np_1^r}{2r^r k} \right)^{k-r} \cdot ((k-r)!)^r \geq \left(\frac{np_1^r (k-r)^r}{2er^r k} \right)^{k-r} .
$$

Choosing c_1 to be small enough and plugging in k and p, the aforementioned probability can be lower bounded by $(c_1/10)^{c_1 \log n} > n^{-1/3}$. The rest of the argument is essentially identical to that of Lemma 3.3. $\square$

Similar arguments to those presented in Section 3 imply that the number of generations for a contagious set as above to activate G is $\Theta(r \log \log n)$. We omit the details.

5. Conclusions. The discussion below concerns the case $r = 2$.

Theorems 1.1 and 1.2 both show that the smallest contagious set has size $\Theta\left(\frac{n}{d^2 \log d} \right)$ w.h.p., but address two different ranges of degrees. The negative results (nonexistence of small contagious sets) in both theorems are based essentially on the same argument (lower bounds on the size of the smallest subgraph of average degree $4 - O(\frac{1}{\log d})$). However, our proofs of the upper bounds in the two theorems are based on different principles. Our proof for Theorem 1.1 is based on an algorithm that performs $\log \log d$ iterations, where in every iteration additional vertices are designated as seeds. Such a proof produces a contagious set of size at least $\Omega(\log \log d)$ (in fact, our proof of Proposition 2.3 requires values of d for which the contagious set is even larger), and hence is inappropriate for Theorem 1.2, in which the total

number of seeds allowed is only 2. In contrast, the proof of Theorem 1.2 is based on examining disjoint pairs of vertices until some pair is found to be contagious. Disjointness implies that the proof examines much fewer than n candidate sets for being contagious. Such a proof is inappropriate for Theorem 1.1, because for the range of degrees considered in Theorem 1.1 a random set of size $\Theta\left(\frac{n}{d^2 \log d}\right)$ has probability much less than $1/n$ of being contagious.

In this work we did not handle two ranges of degrees. One is when d is a large constant. It is not difficult to extend Theorem 1.1 also to the case of large constant degrees. This range of degrees is omitted from the current work mainly for the reasons of simplicity, as we are using Theorem 1.3 as a blackbox, and that Theorem requires $d = d(n)$ to tend to infinity with n. The more challenging range of parameters omitted from our paper is when $d = o\left(\sqrt{\frac{n}{\log n}}\right)$ but still too large for Theorem 1.1 to apply. It would be interesting to prove that the smallest contagious set has typically size $\Theta\left(\frac{n}{d^2 \log d}\right)$ also in this regime.

The positive results in Theorem 1.2 implicitly establish one specific average degree $d = \Theta\left(\sqrt{\frac{n}{\log n}}\right)$ that suffices with high probability for two related problems: one is the existence of a contagious set of size 2, and the other is the existence of such a set for which the number of generations is $O(\log \log n)$ (which is best possible up to constant multiplicative factors). For both problems, this value of d is best possible up to constant factors. Nevertheless, it would be interesting to determine whether there is some $d' < d$ for which with high probability there is a contagious set of size 2, but every contagious set of size 2 requires more than $O(\log \log n)$ generations.

Theorem 1.1 does not explicitly address the number of generations. An upper bound on the number of generations implicit in our proof of Theorem 1.1 is $O(\log d \log \log d)$, but we doubt that it is tight.

REFERENCES

[1] M. Aizenman and J. L. Lebowitz. Metastability effects in bootstrap percolation. *J. Phys. A.* 21: 3801–3813, 1988.

[2] H. Amini. Bootstrap percolation and diffusion in random graphs with given vertex degrees. *Electronic Journal of Combinatorics*, 17: Research paper 25, 2010.

[3] H. Amini and N. Fountoulakis. What I tell you three times is true: bootstrap percolation in small worlds. In *Proceedings of the 8th Workshop on Internet and Network Economics* (WINE '12) (P. Goldberg, Ed.), Lecture Notes in Computer Science: 7695, 462–474, 2012.

[4] S. Arya, M. Golin, and K. Mehlhorn. On the expected depth of random circuits. *Combinatorics Probability and Computing* 8: 209-228, 1999.

[5] J. Balogh and B. Bollobás. Bootstrap percolation on the hypercube, *Probabilty Theory and Related Fields*, 134: 624–648, 2006.

[6] J. Balogh, B. Bollobás, H. Duminil-Copin and R. Morris. The sharp threshold for bootstrap percolation in all dimensions. *Trans. Amer. Math. Soc.*, 364: 2667–2701, 2012.

[7] J. Balogh, Y. Peres and G. Pete. Bootstrap percolation on infinite trees and non-amenable groups. *Combinatorics, Probability and Computing*, 15: 715–730, 2006.

[8] J. Balogh and G. Pete. Random disease on the square grid. *Random Structures and Algorithms*, 13: 409–422, 1998.

[9] J. Balogh and B. Pittel. Bootstrap percolation on the random regular graph. *Random Structures and Algorithms*, 30: 257–286, 2007.

[10] B. Bollobás. *Random graphs*. Cambridge Stud. Adv. Math. 73, Cambridge University Press, 2nd Edition, Cambridge, 2001.

[11] B. Bollobás, C. Holmgren, P.J. Smith, and A.J. Uzzell. The time of bootstrap percolation for dense initial sets. *Annals of Probability*, 42: 1337–1373, 2014.

[12] B. Bollobás, P.J. Smith, and A.J. Uzzell. The time of bootstrap percolation with dense initial sets for all thresholds. *Random Structures and Algorithms* 47: 1–29, 2015.

[13] B. Bollobás and A. Thomason. Threshold functions. *Combinatorica* 7: 35-38, 1987.

[14] A. Coja-Oghlan, U. Feige, M. Krivelevich and D. Reichman. Contagious sets in expanders. *SODA*, 1953–1987, 2015.

[15] J. Chalupa, P. L. Leath and G. R. Reich. Bootstrap percolation on a Bethe lattice. *J. Phys. C: Solid State Phys.*, 12, p L31, 1979.

[16] N. Chen. On the approximability of influence in social networks. *SIAM Journal of Discrete Math*, 23: 1400–1415, 2009.

[17] J. Friedman. A proof of Alon's second eigenvalue conjecture and related problems. *Memoirs of the AMS*, 910, 2008.

[18] A. Frieze and C. McDiarmid. Algorithmic theory of random graphs. *Random Structures and Algorithms*, 10: 5–42, 1997.

[19] D. Freund, M. Poloczek and D. Reichman. Contagious Sets in Dense Graphs. arXiv preprint arxiv:1503.00158, 2015.

[20] S. Janson. On percolation in random graphs with given degree sequence. *Electronic Journal of Probability*, 14: 86–118, 2009.

[21] S. Janson, T. Łuczak, T. Turova and T. Vallier. Bootstrap percolation on the random graph $G_{n,p}$. *Annals of Appied Probability*, 22: 1989–2047, 2012.

[22] M. Mitzenmacher and E. Upfal. *Probability and Computing, randomized algorithms and probabilistic analysis*. Cambridge University Press, Cambridge, 2005.

[23] N. Morrison and J. Noel. Extremal bounds for bootstrap percolation in the hypercube. Preprint, http://arxiv.org/abs/1506.04686v1.

[24] D. Reichman. New bounds for contagious sets. *Discrete Math.*, 312: 1812–1814, 2012.

[25] G.-P. Scalia-Tomba. Asymptotic final-size distribution for some chain-binomial processes. *Adv. in Appl. Probab.*, 17: 477–495, 1985.

[26] T. Tsukiji and F. Xhafa. On the depth of randomly generated circuits. *ESA*, 208-220, 1996.

[27] A. C. D. van Enter. Proof of Straley's argument for bootstrap percolation. *J. Stat. Phys.*, 48: 943–945, 1987.

[28] T. Vallier. Random graph models and their applications. Ph.D. thesis, Lund University, 2007.

Uriel Feige
Department of Computer Science and Applied Mathematics
The Weizmann Institute
7610001, Rehovot
Israel
E-mail: uriel.feige@weizmann.ac.il

Michael Krivelevich
Tel Aviv University
School of Mathematical Sciences
Faculty of Exact Sciences
6997801, Tel Aviv
Israel
E-mail: krivelev@post.tau.ac.il

Daniel Reichman
Electrical Engineering and Computer Science
University of California at Berkeley
Berkeley, CA 94709
USA
E-mail: daniel.reichman@gmail.com