Rumor Spreading in Social Networks[*]

Flavio Chierichetti, Silvio Lattanzi, Alessandro Panconesi

Dipartimento di Informatica
Sapienza Università di Roma

Abstract. Social networks are an interesting class of graphs likely to become of increasing importance in the future, not only theoretically, but also for its probable applications to ad hoc and mobile networking. Rumor spreading is one of the basic mechanisms for information dissemination in networks, its relevance stemming from its simplicity of implementation and effectiveness. In this paper, we study the performance of rumor spreading in the classic preferential attachment model of Bollobás et al. which is considered to be a valuable model for social networks. We prove that, in these networks: (a) The standard `PUSH-PULL` strategy delivers the message to all nodes within $O(\log^2 n)$ rounds with high probability; (b) by themselves, `PUSH` and `PULL` require polynomially many rounds. (These results are under the assumption that m, the number of new links added with each new node is at least 2. If $m = 1$ the graph is disconnected with high probability, so no rumor spreading strategy can work.) Our analysis is based on a careful study of some new properties of preferential attachment graphs which could be of independent interest.

1   Introduction

Rumor spreading is one of the basic mechanisms for information dissemination in networks. In this paper we analyze the performance of rumor spreading in the Preferential Attachment model [7]. We show that, while neither `PUSH` nor `PULL` by themselves guarantee fast information dissemination, with `PUSH-PULL` the information reaches all nodes in the network within $O(\log^2 n)$ rounds with high probability, n being the number of nodes in the network.

The study of information dissemination in social networks is an important endeavour, encompassing a variety of questions ranging from the purely technological to the spread of viruses and the diffusion of ideas in human communities. In order to gain insight into these and other related questions, a lot of activity has been devoted to studying stochastic generative models for social networks. The well-known preferential attachment model, defined precisely in the next section, is considered to be able to capture many of their salient features. Thus, it seems worthwhile to study how fundamental primitives of information dissemination behave in such a model. Rumor spreading is one of the very basic such primitives. Its simple, basic character makes it useful as a protocol and interesting theoretically, for one can hope to gain insight on more complex

phenomena by studying it. Perhaps surprisingly there seem to be no precise analysis in the literature of the speed with which rumor spreading terminates in the preferential attachment model (to the best of our knowledge) and thus this is the aim of this paper. Before discussing the relevant state-of-the-art, let us first describe our results precisely.

There is a danger of terminological confusion surrounding the term gossip that we shall now try to avoid. In this paper we study the *randomized gossip protocol*, also referred to in the literature with the terms rumor spreading or randomized broadcast (see for instance [11,19]). It should not be confused with the *gossip problem*, in which each node is to broadcast a piece of information and one seeks the most effective protocols to do it (see for instance [20]). The randomized gossip protocol is a widely used protocols in ad hoc networks to implement a broadcast service, and is defined as follows. Its aim is to broadcast a message, i.e. to deliver to every node in the network a message originating from one source node. The randomized gossip protocol proceeds in a sequence of synchronous rounds. At round $t \geq 0$, every node that knows the message, selects a random neighbour and sends message. This is the so-called `PUSH` strategy. The `PULL` variant is specular. At round $t \geq 0$ every node that does not yet have the message selects a neighbour uniformly at random and asks for the information, which is transferred provided that it is in possession of the queried neighbour. Finally, the `PUSH-PULL` strategy is a combination of both. At round $t \geq 0$, each node selects a random neighbour to perform a `PUSH` if it has the information or a `PULL` in the opposite case. One of the most studied question concerning rumor spreading is the following: how many rounds will it take for one of the above strategies to disseminate the information to all nodes in the graph, assuming a worst-case source?

We study this question for the preferential attachment model and show the following:

- regardless of the starting node, the `PUSH` strategy requires, with $\Omega(1)$ probability, polynomially many rounds;
- there are starting nodes such that the `PULL` strategy requires, with $\Omega(1)$ probability, polynomially many rounds;
- regardless of the starting node, the `PUSH-PULL` strategy requires, with probability $1 - o(1)$, $O(\log^2 n)$ many rounds.

From the technical point of view, Preferential Attachment networks (henceforth PA network) are quite interesting because in them coexist subgraphs that are very hard for rumor spreading, such as many high-degree stars, the so-called hubs, with subgraphs that are very good, such as small-degree expanders. These two act as opposing forces and only a careful analysis can ascertain which one will prevail. To this aim, we establish several strong structural properties of PA graphs that we believe will be useful for further study. In particular we show that a linear size portion of the graph behaves like a low-degree expander. This expander is not a subgraph made of nodes and edges. Rather it is a sort of cluster graph obtained by collapsing connected components into macronodes that are connected by short paths (as opposed to single edges). This core acts as a sort of fast information exchanger– it can be easily reached by every node and it can itself reach every node.

We now turn to a discussion of the relevant literature.

2 Related work

The literature on the gossip protocol and social networks is huge and we confine ourselves to what appears to be more relevant to the present work.

Clearly, at least diameter-many rounds are needed for the gossip protocol to reach all nodes. It has been shown that $O(n \log n)$ rounds are always sufficient for a connected graph of n nodes [11]. The problem has been studied on a number of graph classes, such as hypercubes, bounded-degree graphs, cliques and Erdös-Rényi random graphs (see [15,11,18]).

It is a common assumption that graphs generated by the model intuitively introduced by Barabási and Albert is a good representation of social networks [1]. In this model nodes arrive one after the other. Roughly speaking, when a new node arrives, m nodes are chosen randomly as neighbours, with probability proportional to their degree. Bollobás et al. formalize this model in [7]. We will use the terms "preferential attachment" and refer to it as the *PA model*. Analogously, we will speak of *PA graphs* and *PA networks*. The PA model has been the object of a great deal of rigorous study by a number of authors [2,4,5,6,7,8,14]. For instance, it is known that (a) its degree distribution follows a power-law [7], (b) its maximum degree is $\geq n^{\frac{1}{2}-\epsilon}$ [13], (c) its diameter is $\Theta(\log n/\log \log n)$ (for $m \geq 2$), and (d) its cover time is $\Theta(n \log n)$ (again for $m \geq 2$).

An interesting property of the PA graphs is their robustness with respect to node deletion. In [4,14] the authors study the size of the largest component of the PA graphs after random and adversarial node deletions. In [2] the authors study the virus-spreading problem on graphs very similar to PA graphs, relating it to spreading of computer viruses over the internet.

It is natural to ask whether expansion and high conductance imply that rumor spreading is fast. The graph in Figure 1 has high edge expansion, but rumor spreading takes linearly many rounds. The graph consists of $\sqrt{n}$ many independent sets, each of size $\sqrt{n}$. These independent sets are arranged in a cycle. Two neighbouring independent sets form a complete bipartite graph. The central node is connected to one vertex in each independent set.
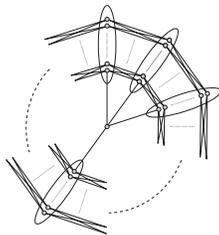

Fig. 1. Slow rumor spreading in spite of high edge expansion

The graph also has high edge expansion but `PUSH-PULL` requires polynomially many rounds in spite of the diameter being constant. Whether high conductance by itself implies the success of `PUSH-PULL` in general appears to be an intriguing and difficult open problem.

Mihail et al. [16] study the edge expansion and the conductance of graphs that are very similar to PA graphs. We shall refer to these as "almost" PA-graphs [2]. They show that edge expansion and conductance are constant in these graphs, when $m \geq 2$.

High conductance implies that *non-uniform* rumor spreading succeeds. By non-uniform we mean that, for every ordered pair of neighbours i and j, node i will select j with probability p_{ij} for the rumor spreading step (in general, $p_{ij} \neq p_{ji}$). Boyd et al. [3] consider the "averaging" problem on general graphs, which is closely related to the convergence of `PUSH-PULL`. A corollary of their main results is that, if the p_{ij} are suitably chosen, non-uniform `PUSH-PULL` rumor spreading succeeds within $O(\log n)$ rounds in almost-PA graphs. They also show that this distribution can be found efficiently using local computations in these graphs, but their method requires $\Omega(\log n)$ steps. While the contribution of [3] is noteworthy, this corollary is in our context somewhat trivial. That such a probability distribution exists is straightforward. Because of their high conductance, almost-PA graphs have diameter $O(\log n)$. Thus, in a synchronous network, it is possible to elect a leader in $O(\log n)$ many rounds and set up a BFS tree originating from it. By assigning probability 1 to the edge between a node and its parent one has the desired non uniform probability distribution. Thus, from the point of view of this paper the existence of non uniform problem is rather uninteresting. Boyd et. al [3] also show that this distributions can be found efficiently using local computations, but their method requires $\Omega(\log n)$ many steps. The local computations of each node, at every step, include a broadcasting of some value to all neighbours. Local broadcasting, used for diameter (that is, $O(\log n)$) many rounds, is a trivial information-dissemination strategy.

Also, Mosk-Aoyama and Shah [17] consider the problem of computing separable functions. In particular, they consider the uniform rumor spreading problem on graphs weighted by a high-conductance doubly-stochastic matrix "that assigns equal probability to each of the neighbours of any node" (that is, if p_{ij} is the probability that node i initiates a connection to node j in the generic round t, then $\forall ij \in E(G)$ $p_{ij} = p_{ji} = \Delta^{-1}$, where Δ is the maximum degree in the graph). Their work implies that if the conductance (or the edge expansion) of a graph is $\Omega(1)$ then rumor spreading ends in $O(\Delta \log^2 n)$ many rounds — this, while being a good bound for constant-degree graphs, is polynomially large for PA graphs (where the bound would be larger than $\Omega(n^{\frac{1}{2}-\epsilon})$).

3 Preliminaries

Preferential attachment graphs were intuitively introduced in [1] and formally defined in [7], from which the following definition is taken.

Definition 1. [PA model]. *Let G_n^m, m being a fixed parameter, be defined inductively as follows:*

- G_1^m consists of a single vertex with m self-loops.
- G_n^m is built from G_{n-1}^m by adding a new node u together with m edges $e_u^1 = (u, v_1), \ldots, e_u^m = (u, v_m)$ inserted one after the other in this order. Let M_i be the sum of the degree of all the nodes right before the edge e_u^i is added. The endpoint v_i is selected with probability $\frac{\deg(v_i)}{M_i+1}$, with the exception of node u that is selected with probability $\frac{\deg(u)+1}{M_i+1}$.

In other words, if a vertex $v \neq u$ has degree d when e_u^i is inserted, it will be chosen with probability proportional to d, while u will be chosen with probability proportional to its current degree plus one. Note that definition allows for self-loops. The rich-get-richer effect is clear, since the higher the degree of a node, the higher the probability that it will be chosen as the endpoint of a new edge.

In the sequel we will say that an event holds *with high probability* if it occurs with probability $1 - o(1)$, where $o(1)$ will be a quantity going to zero as n, the number of vertices of the graph under consideration, grows.

Definition 2. *Consider a rumor spreading strategy (*PUSH, PULL *or* PUSH-PULL*). Given a (random) graph of n nodes, we say that the strategy* succeeds *if the message is delivered within poly-logarithmically, in n, many rounds, regardless of the starting node, with probability $1 - o(1)$. It* fails *if, with non-vanishing probability, there exist a node from which the message requires polynomially-many rounds to be delivered to all nodes (i.e. it requires $\Omega(n^\alpha)$ many rounds for some $\alpha > 0$).*

4   Rumor spreading in Social Networks

We begin by showing some simple lower bounds for the performance of PUSH and PULL acting alone, and that the condition $m \geq 2$ is necessary. This discussion will provide the motivation to analyse the PUSH-PULL strategy.

The requirement $m \geq 2$ is due to the fact that G_n^1 is disconnected with high probability. The next lemma is implicit in [6].

Lemma 1. G_n^1 *is connected with probability*

$$\frac{\sqrt{\pi}}{2} \cdot \frac{\Gamma(n)}{\Gamma(n+1/2)} = \Theta\left(\frac{1}{\sqrt{n}}\right),$$

where Γ denotes the gamma function ($\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$).

Proof. The graph is connected iff no node, except the first, has a self-loop. The probability of this event is $\prod_{i=2}^n \frac{2i-2}{2i-1}$, which is equivalent to the expression in the claim.   □

Next we characterize the performance of PUSH and PULL. Fix $\epsilon > 0$. We say that a node is of *high degree* if its degree is $> n^\epsilon$. The next lemma says that, for a suitably small ϵ, there are lots of high degree nodes in the graph. More precisely, with high probability, the sum of their degrees is $\Omega(n^{1-\epsilon})$. To prove the lemma we need a sharp estimate of $E[X_k^n]$, where X_k^n is the number of nodes of degree k in G_n^1. [7] gives the bound $E[X_i^t] = (1 \pm o(1))\frac{4t}{i(i+1)(i+2)}$ but this is not sufficient for our purposes. We require a bound of the form $E[X_i^t] \leq \frac{4t}{i(i+1)(i+2)} + c$, for some absolute constant c (say, $c = 2$).

Lemma 2. *Fix $\epsilon > 0$ sufficiently small. Then, with high probability, the sum of the degrees of nodes that in G_n^1 have degree $> n^\epsilon$, is $\Omega(n^{1-\epsilon})$.*

The proof is omitted from this extended abstract for lack of space.

Theorem 1. *Both* PUSH *and* PULL *fail on* $\{G_n^m\}$.

Proof. If $m = 1$ the claim is implied by Lemma 1. We assume $m \geq 2$. Let us consider G_{2n}^m. We show next that, if we consider the nodes inserted after time n, at least $\Omega(n^\alpha)$ of them are connected only to high degree nodes, for some constant $\alpha > 0$.

We will think of every edge uv as a pair of "half-edges", going out of u and v respectively. A half-edge is *good* if, at time n, it goes out of a high degree node. Consider now a node u arrived at time $t > n$. Choosing a neighbour for u is equivalent to choosing a half-edge uniformly at random in G_{t-1}^m. We say that u is *good* if the half-edges it chooses are all good. Note that the events of being good for each of the nodes from $n + 1$ to $2n$ are independent.

As in [7], we can think of G_n^m in the following, equivalent way. Generate G_{nm}^1 and then collapse into one node each sequence of m consecutive nodes. Clearly, the degree of a node in G_n^m is at least as large as the degree of the nodes G_{nm}^1 it comes from. The sum of degrees of high degree nodes in G_n^m is no less than the analogous sum in G_{nm}^1.

By Lemma 2, the probability of a node being good is at least

$$\left(\Omega \left(\frac{n^{1-\epsilon}}{n} \right) \right)^m \geq \Omega(n^{-m\epsilon}).$$

We choose ϵ in such a way that $(m + 1)\epsilon < 1$. Say, $0 < \epsilon \leq 1/(m + 2)$.

Now take the last $\Theta(n^{(m+1)\epsilon})$ nodes. Among those, by Chernoff-Hoeffding bound, at least $\Omega(n^\epsilon)$ are good, with high probability. Let v_t be one of them and suppose that it was inserted at time $t \geq 2n - \Theta(n^{(m+1)\epsilon})$.

The probability that v_t not chosen as a neighbour by nodes inserted at later times is at least

$$\prod_{i=mt+1}^{2mn} \left(1 - \frac{m}{2i - 1} \right) > \prod_{i=mt+1}^{2mn} \left(1 - \frac{m}{2mt} \right) \geq \left(1 - \frac{1}{2n} \right)^{m\Theta(n^{(m+1)\epsilon})} = 1 - o(1).$$

In other words, with high probability, v_t is only connected to m nodes of high degree.

So, suppose the PULL algorithm is being used. If the source of the message is v_t, then the message will not be passed to any other node in time $< o(n^\epsilon)$ with high probability because its m neighbours all have high degree — PULL does not work.

Analogously, if we place the message in $u \neq v_t$, since the only way to reach v_t is via m high degree nodes, PUSH will not be able to route the message to v_t in $o(n^\epsilon)$ time. □

The previous theorem provides strong motivation to analyse the PUSH−PULL strategy.

5 Push and Pull acting together

In view of Lemma 1 we assume $m \geq 2$ from now on. Our aim is to show the following.

Theorem 2. *Let* $m \geq 2$. *Then,* PUSH–PULL *succeeds for* $\{G_n^m\}$ *within* $O(\log^2 n)$ *many rounds.*

The proof will be based on several structural facts concerning PA graphs. Some are taken from the existing literature, while new ones will have to be established. We will use the following from [6].

Theorem 3. *Let* $m \geq 2$. *The diameter of* G_n^m *is* $O(\log n / \log \log n)$ *with high probability.*

Consider a connected graph with n nodes in which every node has degree $O(\log n)$ and that has $O(\log n)$ diameter. It is easy to see that in such a graph rumor spreading succeeds within $O(\log^2 n)$ rounds. The next two lemmas say that a PA graph contains a linear size subgraph of this kind. Their proofs are postponed to the next section. Here they are used to prove Theorem 2.

Lemma 3. *Let* $m \geq 2$ *and let* $\epsilon > 0$ *be sufficiently small. Then, there exists a set of vertices* $W \subseteq V(G_n^m)$, *such that: (a)* W *only contains nodes added after time* ϵn *and before time* $n/2$, *(b)* $|W| \geq \epsilon n$, *and (c) every pair of vertices* $x, y \in W$ *are connected by a path of length* $O(\log n)$ *consisting solely of edges inserted between time* ϵn *and* $3/4n$.

The next lemma roughly says that if a vertex is not a hub by time ϵn it will never be (this includes the case of nodes inserted after that time).

Lemma 4. *Let* $m \geq 2$ *and fix any* $\epsilon > 0$. *Then, with high probability the following holds: (a) Every node added after time* ϵn *has degree* $O(\log n)$ *in* G_n^m, *and (b) For every* $c' > 0$ *there exists* $c > c'$ *such that, if a node has degree* $\leq c' \log n$ *in* $G_{\epsilon n}^m$ *its degree in* G_n^m *will be* $< c \log n$.

The following lemma follows from lemmas 3 and 4. It says that hubs are at distance at most 2 from W.

Lemma 5. *Let* W *be as in Lemma 3 and let* $m \geq 2$. *Then, there exists a sufficiently large constant* c *such that, with high probability, every node* $v \in V(G_n^m)$ *with degree* $\geq c \log n$ *is at distance* ≤ 2 *from* W.

Proof. Let H be the set of vertices inserted before time ϵn and let R be the set of vertices inserted after time $\frac{3}{4}n$. Recall that W is a subset of the vertices inserted after H and before R. A vertex $v \in R$ is *good* if it is connected to W with its first edge. Our aim is to show that R contains $\Theta(n)$ good vertices. Now, given any vertex in W we will consider only the m half-edges going out of it when this vertex joined the graph. Even with this limitation, the first edge choice of $v \in R$ hits W with probability at least

$$\frac{|W|m}{2mn}.$$

By Lemma 3(b) this is at least some constant $1 > p > 0$. By our previous choice concerning the half-edges of vertices in W, being good is an independent Bernoulli trial. It follows from the Chernoff-Hoeffding bounds (and stochastic domination) that the number of good nodes is $\Theta(n)$ with high probability.

So far we have exposed only the random choice of the first edge of every good node. Thus, there remain $\Theta(n)$ other edges going out of good nodes that are to be fixed. We will use this to prove that every vertex in H whose degree is $\geq c\log n$ in G_n^m is connected to a good node with probability at least $1 - O\left(n^{-2}\right)$. Once this is done the claim will follow from the union bound, since $|H| < n$.

Let v' be of degree $\geq c\log n$ (the value of c will be fixed later). By Lemma 4 this vertex must have been added before time ϵn, i.e. $v' \in H$, and must have had degree $\geq c'\log n$ at that time. Let z be a good node inserted at time $t > \frac{3}{4}n$. The second edge choice of z selects v' with probability at least

$$\frac{\text{(degree of } v' \text{ at time } t) \text{ - } m}{2m(\# \text{ edges at time } t)} \geq \frac{c'\log n - m}{2mn} = \frac{c''\log n}{n}$$

These choices are Bernoulli trials with total expectation $\geq c'''\log n$ (that is the number of good nodes times $c''\frac{\log n}{n}$). It follows from the Chernoff-Hoeffding bounds that c (and consequently c', c'', c''') can be chosen in such a way that the probability that v' has no neighbour among the good vertices is at most $O\left(\frac{1}{n^2}\right)$. The claim follows by union bound. □

Given the previous lemmas, we can prove Thm. 2 in the following way.

Proof (of Thm. 2). Let us partition $V(G_n^m)$ into three classes. Let $H \subseteq V$ be composed of the nodes of degree $\Omega(\log n)$, let $W \subseteq V$ be as in lemma 3 (note that by lemma 4, w.h.p. $W \cap H = \varnothing$) and let $R = V - H - W$.

Suppose that the information starts in some node of H. Then, by the PULL strategy the information will be taken by at least one node in W in time $O(\log^2 n)$ (by the maximum degree of nodes in W, a coupon collector argument and lemma 5).

Suppose instead that the information starts in some node of R. The distance between one node of R and one node in $V - R$ is at most $O(\log n)$ by the diameter bound of theorem 3. Also, each of the nodes in R has degree $O(\log n)$ by definition. Thus, in time $O(\log^2 n)$ (= maximum distance $\times$ maximum degree, see [11]) the information will reach W (either directly or by passing through H) by the PUSH and PULL strategies.

So, we can assume that after $O(\log^2 n)$ steps the information is in W. Each node added after time ϵn has degree $O(\log n)$ and the diameter of W, even after projecting on $W \cup R$, is $O(\log n)$. Thus if at some point a node in W has the information, after $O(\log^2 n)$ steps the information will have reached all nodes in W by the PUSH strategy.

By lemma 5, if each node in W has the information, after $O(\log^2 n)$ steps the information will have been passed to each of the nodes in H by the PUSH strategy.

After each node in $W \cup H$ has the information, the PULL strategy employed will give the information to each of the nodes in R in time $O(\log^2 n)$. □

6 Proofs

In this section we prove the two main lemmas from the previous section, Lemma 3 and 4. Recall that Lemma 4 was a statement about degrees in G_n^m. The next lemma is the key technical ingredient.

Lemma 6. *Consider a Polya urn process lasting for n steps. Suppose that this Polya urn starts with $R_0 \geq \alpha n$ red balls and $B_0 \leq g(n)$ blue balls, for $\alpha > 0$ and some function $g(n) \in o(n)$. Then, with probability $1 - o(1/n)$, the number of blue balls after the n-th extraction, B_n, will be $B_n \leq c \max\{g(n), \log n\}$, for some constant $c > 0$.*

Proof. The probability that, overall, k blue balls will be added to the urn is

$$
\begin{aligned}
P[B_{n+B_0+R_0} = k] &= \binom{n}{k} \cdot \frac{B_0 \cdots (B_0 + k - 1) \cdot R_0 \cdots (R_0 + n - k - 1)}{(B_0 + R_0) \cdots (B_0 + R_0 + n - 1)} \\
&= \frac{(B_0 + k - 1)!}{k! \, (B_0 - 1)!} \cdot \frac{(R_0 + n - k - 1)!}{(R_0 - 1)! \, (n - k)!} \cdot \frac{n! \, (B_0 + R_0 - 1)!}{(B_0 + R_0 + n - 1)!} \\
&= \frac{\binom{B_0 + k - 1}{k} \cdot \binom{R_0 + n - k - 1}{n - k}}{\binom{B_0 + R_0 + n - 1}{n}} \\
&= f(k; B_0 + R_0 + n - 2, B_0 + k - 1, n) \frac{B_0 + R_0 - 1}{B_0 + R_0 + n - 1}.
\end{aligned}
$$

Where $f(k; s, t, u)$ is the probability of getting exactly k good elements from a sample without replacement of u elements, from a set of s elements, t of which are good.

Consider the numerator of the second to the last expression, is $h(k) = \binom{B_0 + k - 1}{k} \cdot \binom{R_0 + n - k - 1}{n - k}$. By simple algebraic manipulation, we obtain that for integer $k \geq c' \cdot g(n)$, for some $c' > 0$, it holds that $h(k) > h(k + 1)$. Therefore, the whole expression is decreasing for k in that range.

Now, let us bound the "bad" event using r for $r = c' \cdot (g(n) + \log n)$, for some sufficiently large constant $c' > 0$,

$$
\begin{aligned}
P[B_{n+B_0+R_0} \geq r] &= \sum_{i=r}^{n+B_0} P[B_{n+B_0+R_0} = i] \\
&\leq (n + B_0) P[B_{n+B_0+R_0} = r] \\
&= (n + B_0) f(r; B_0 + R_0 + n - 2, B_0 + r - 1, n) \frac{B_0 + R_0 - 1}{B_0 + R_0 + n - 1} \\
&\leq (n + B_0) \sum_{i=r}^{\infty} f(i; B_0 + R_0 + n - 2, B_0 + r - 1, n) \frac{B_0 + R_0 - 1}{B_0 + R_0 + n - 1}
\end{aligned}
$$

where the last step allows us to use the well-known tail bound for the hypergeometric sum [10]. Let P denote the probability that at least k good elements are in a sample (without replacement) of u elements, from a set of s elements, t of which are good. Then

$$
P \leq 2 \exp\left(-\frac{(k - 1 - ut/s)^2}{k - 1}\right).
$$

Note that, in our case,

$$ut/s \leq \frac{n\left(B_0 + r - 1\right)}{B_0 + R_0 + n - 2} \leq (1 \pm o_n(1)) \frac{c' + 1}{1 + \alpha} \left(g(n) + \log n\right).$$

So, we get that $k - 1 - ut/s \geq (1 \pm o_n(1)) \frac{c'\alpha - 1}{1 + \alpha} \left(g(n) + \log n\right)$. Thus it is possible to make this lower bound arbitrarily large, by choosing c' bounded away from $\frac{1}{\beta}$. The statement follows. □

We show how Lemma 6 implies lemma 4. Take any node v having degree less than $c \log n$ at time ϵn. The graph process can be seen as the following urn process. At time ϵn, the urn contains a number of blue balls equal to d_v, the degree of v, and $\epsilon n - d_v$ red balls. The urn process works as follows. At each new time step, a red ball is added to the urn. Then a random ball is extracted. The time step ends after this ball, and a new one of the same color, are added to the urn.

The number of blue balls of the a Polya urn process (with the same starting conditions, and the same running time) dominates the number of blue balls of the urn process just described.

This proves the second part of lemma 4. The first part follows by noting that, just after having added the generic node j, the degree of j is upper bounded by $2m$. Suppose that v was added after time ϵn. Then, the Polya urn process of lemma 6, with an initial urn of $\log n$ blue balls and ϵn red ones, trivially dominates the degree of v.

We now move on to proving Lemma 3. The next lemma says that, given any positive integer k, any tree can be partitioned into connected components of size k (with the exception of one component) and diameter at most $2k$. Later we will use this to cluster a linear size subgraph of G_n^m.

Lemma 7. *Fix some integer $k \geq 1$. The nodes of any tree can be partitioned into connected components called macronodes in such a way that: (a) each macronode, except at most one, contains k nodes, and (b) the diameter (in the tree) of each macronode is at most $2k$.*

Proof. Fix a root, and label each node with the number of nodes of the subtree rooted there. Pick a node v having the smallest label ℓ greater than or equal k. If $\ell = k$, then the subtree rooted at v will form a macronode. If $\ell > k$ then, consider the levels of the subtree rooted at v. We will put into the macronode the nodes of the levels, in ascending order, in such a way that the number of nodes in the macronode will be k. If the number of nodes in current level, plus the already inserted nodes, exceeds k, take any subset of the nodes of the current level in order to reach k. What is the diameter of a macronode? First of all note that the height of the tree rooted at v is at most k (otherwise, the subtree rooted in a child of v would have a number of nodes no less than k but smaller than the subtree rooted at v). Thus, the maximum distance between two nodes in the subtree is bounded by $2k$. □

We finally come to the proof of Lemma 3. The main thrust of the proof is to show that G_n^m contains a linear-size, low degree expander of type $G(N, M)$. Note that the existence of a graph of linear size that is "almost" of type $G(N, M)$ was already proven

in [5], where "almost" means that a linear, albeit small, number of edges may be added and/or deleted from $G(N, M)$. These edges introduce complications, but what makes Lemma 3 is that the proof in [5] holds only if m is a very large constant, while we assume $m \geq 2$.

Proof (of Lemma 3). By theorem 3.2 of [4], w.h.p. there exists a subset W' of the nodes of $V(G_n^m)$ such that: (a) W' contains nodes added after time ϵn and before time $n/2$, (b) $|W'| \in \Theta(n)$, and (c) $V(G_n^m)$ projected on W' is connected.

Let us fix such a W'. Take any spanning tree of W' and partition it in macronodes of size $s = \lceil ((4mn)/|W'|)^2 \rceil \in O(1)$ as shown in lemma 7. (Recall that the last macronode may not have the required size. In that case, we remove its nodes from W'.) By the lemma, the diameter of each of the macronodes in $V(G_n^m)$ is at most $O(1)$.

We will show that the macronodes will be connected thanks to an Erdös-Renyi-like random graph $G(N, M)$ (a graph having N nodes and M edges chosen UAR between those with these properties) and some other edges. Also, $M \geq (1/2 + \epsilon)N$: by a theorem of [12] this implies that, this $G(N, M)$ will contain a giant component (i.e., a component of size $\Omega(n)$) of diameter $O(\log n)$. As the diameter of a macronode is $O(1)$, this will imply the main statement.

Consider nodes added between $\lceil n/2 \rceil$ and $3/4n - 1$. Each of those nodes will choose the first 2 of its m neighbours by selecting the outgoing edges of the nodes in W' with probability at least $(|W'|/(2mn))^2$. If such an event occurs we say that a "pseudo-edge" is added between the macronodes containing the two selected vertices. The macronodes, and the pseudo-edges, will compose the Erdös-Renyi-like random graph $G(N, M)$.

Consider the auxiliary graph in which macronodes are vertices and two of them are connected by an edge if there is a pseudo-edge connecting them. The number t of such macronodes in W' is $t \leq |W'|/s$. The number of pseudo-edges added between macronodes is, with high probability, at least $(|W'|/(2mn))^2 \cdot n/4 = |W'|^2/(16m^2 n)$. As the latter is greater than $(1/2 + \epsilon)|W'|/s$, it follows from the results of [12] that in the auxiliary graph there exists a giant component having diameter $O(\log n)$ and size $\Omega(n)$. The claim follows by choosing W as the set of nodes that make up the macronodes. □

7 Conclusion

We have shown how fast the `PUSH-PULL` strategy disseminates some information throughout the nodes of a PA graph, and how slow the `PUSH`, `PULL` strategies obtain the same result. In doing so, we have proved some lemmas that strengthen the connection between the PA random graphs and classical Erdös-Renyi random graphs.

We believe that our results might offer some insights on real rumor spreading among humans. Namely, it seems plausible that in a social network there exists a "core" of people that might not be VIPs, and yet collectively are able to reach a majority of their community in a few steps. Also, our proofs indicate how VIPs are only important for relaying the information and not as originators of rumours (that is, even if they never started a communication themselves, the information would still spread through the network — just by people asking and telling them what they know) .

8 Acknowledgements

We would like to thank Benjamin Doerr and D. Sivakumar for several useful discussions on this problem.

References

1. A. L. Barabási, R. Albert, "Emergence of Scaling in Random Networks", Science 1999, Vol. 286. no. 5439, pp. 509-512.
2. N. Berger, C. Borgs, J. T. Chayes, A. Saberi, "On the spread of viruses on the internet", SODA 2005: 301-310.
3. S. P. Boyd, A. Ghosh, B. Prabhakar, D. Shah, "Gossip algorithms: design, analysis and applications", INFOCOM 2005: 1653-1664.
4. B. Bollobás, O. Riordan, "Robustness and vulnerability of scale-free random graphs", Internet Mathematics 1 (2003), 1–35.
5. B. Bollobás, O. Riordan, "Coupling Scale-Free and Classical Random Graphs", Internet Mathematics 1(2): (2003).
6. B. Bollobás, O. Riordan, "The Diameter of a Scale-Free Random Graph", Combinatorica, Volume 24, Issue 1 (January 2004).
7. B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, "The degree sequence of a scale-free random graph process", Random Structures & Algorithms, Volume 18, Issue 3 (May 2001).
8. C. Cooper, A. M. Frieze, "The cover time of the preferential attachment graph", J. Comb. Theory, Ser. B 97(2): 269-290 (2007).
9. A. J. Demers, D. H. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. E. Sturgis, D. C. Swinehart, D. B. Terry, "Epidemic Algorithms for Replicated Database Maintenance", PODC 1987: 1-12.
10. D. Dubhashi, A. Panconesi, "Concentration of Measure for the Analysis of Randomised Algorithms", Cambridge University Press, 2009.
11. U. Feige, D. Peleg, P. Raghavan, E. Upfal, "Randomized Broadcast in Networks", Random Struct. Algorithms 1(4): 447-460 (1990).
12. D. Fernholz, V. Ramachandran, "The diameter of sparse random graphs", Random Struct. Algorithms 31(4): 482-516 (2007).
13. A. Flaxman, A. M. Frieze, T. I. Fenner, "High Degree Vertices and Eigenvalues in the Preferential Attachment Graph", Internet Mathematics, 2(1), 2005.
14. A. Flaxman, A. M. Frieze, J. Vera, "Adversarial deletion in a scale free random graph process", SODA 2005: 287-292.
15. A. Frieze, G. Grimmett, "The shortest-path problem for graphs with random arc-lengths", Discrete Applied Mathematics, 10:57–77, 1985.
16. M. Mihail, C. H. Papadimitriou, A. Saberi, "On Certain Connectivity Properties of the Internet Topology", FOCS 2003:28-35.
17. D. Mosk-Aoyama, D. Shah, "Fast Distributed Algorithms for Computing Separable Functions", Information Theory, IEEE Transactions on Publication, Volume 54 , Issue 7.
18. B. Pittel, "On spreading a rumor", SIAM Journal on Applied Mathematics, 47(1):213–223, 1987.
19. B. Doerr, T. Friedrich, T. Sauerwald. Quasirandom Broadcasting. In Proceedings of SODA 2008, pages 773-781.
20. J.Hromkovic, R.Klasing, A.Pelc, P.Ruzicka, Dissemination of Information in Communication Networks: Broadcasting, Gossiping, Leader Election, and Fault-tolerance. Springer-Verlag, 2005

Appendix

8.1 Proof of Lemma 2

Proof. For $t = 1$, we have $E[X_2^1] = 1$ and $E[X_{\neq 2}^1] = 0$. Thus, the following recursion for $t \geq 2$ obtains,

$$E[X_i^t] = \begin{cases} E[X_1^{t-1}]\left(1 - \frac{1}{2t-1}\right) + \left(1 - \frac{1}{2t-1}\right) & \text{if } i = 1 \\ E[X_2^{t-1}]\left(1 - \frac{2}{2t-1}\right) + E[X_1^{t-1}]\frac{1}{2t-1} + \frac{1}{2t-1} & \text{if } i = 2 \\ E[X_i^{t-1}]\left(1 - \frac{i}{2t-1}\right) + E[X_{i-1}^{t-1}]\frac{i-1}{2t-1} & \text{otherwise} \end{cases}$$

An easy induction shows that $E[X_i^t] \leq \frac{4t}{i(i+1)(i+2)} + c$, for some absolute constant c (say, $c = 2$).

As shown in [7], the X_i^n's are concentrated around their expectation within an error of $\sqrt{n} \log n$, for $i \leq n^{1/15}$. Thus, we have that, with high probability, the sum of the degrees of small-degree nodes is:

$$\sum_{i=1}^{n^\epsilon} i \cdot X_i^n \leq \sum_{i=1}^{n^\epsilon} i \cdot \left(E[X_i^n] + O(\sqrt{n} \log n)\right)$$

$$\leq O(n^{1/2+2\epsilon} \log n) + 4n \sum_{i=1}^{n^\epsilon} \frac{1}{(i+1)(i+2)}$$

$$= O(n^{1/2+2\epsilon} \log n) + 2n - \Omega(n^{1-\epsilon})$$

$$= 2n - \Omega(n^{1-\epsilon})$$

The sum of all degrees in G_n^1 after n steps is $2n$. Thus, the total degree of high degree nodes is at least $\Omega(n^{1-\epsilon})$, proving the claim. $\square$