Distributing antidote using PageRank vectors

Fan Chung*
University of California, San Diego

Paul Horn
Emory University

Alexander Tsiatas
University of California, San Diego

Abstract

We give an analysis of a variant of the contact process on finite graphs, allowing for non-uniform cure rates, modeling antidote distribution. We examine an inoculation scheme using PageRank vectors which quantify the correlations among vertices in the contact graph. We show that for a contact graph on n nodes we can select a set H of nodes to inoculate such that with probability at least $1-2\epsilon$, any infection from any starting infected set of s nodes will die out in $c \log s + c'$ time, where c and c' depend only on the probabilitstic error bound ϵ and the infection rate, and the size of H depends only on s, ϵ and the topology around the initially infected nodes, independent of the size of the whole graph.

1 Introduction

The spreading and containment of epidemics on networks is a widely-studied problem with many applications in modeling both disease outbreaks in human and animal populations as well as the spread of viruses and worms on technological networks such as the Internet, online social networks, and email. Many analytical models have been used to address numerous crucial problems, such as the conditions for disease spreading, the critical threshold for the infection rate, the duration of persistent epidemics, and the effective distribution of limited amounts of antidote. We will examine a well-studied contact process model [4, 14], coupled with an inoculation scheme using PageRank vectors. In this paper, we give a complete analysis of our scheme,

*Research supported in part by ONR MURI N000140810747, and AF/SUB 552082.

showing the improved efficiency of the inoculation scheme without affecting the performance guarantee as in previous results in [4].

A contact graph consists of a set of nodes together with prescribed pairs of nodes where direct contact can take place and infections can spread (see [6, 9, 14]). Analysis of spreading on the contact graph is performed with the contact process, a continuous-time Markov process originally studied in the first half of the twentieth century [10]. Since then, it has been applied specifically to network epidemics in many contexts, including social networks [17], Internet viruses [3], and crop disease [8].

Previously, most analysis of network infection models concerned determining the critical infection threshold [3, 9, 14]. There is a parameter, known as the infection rate, that models the virulence or resistance of a given epidemic, and with it comes a threshold: if the infection rate exceeds that point, then an epidemic will persist indefinitely. In these analyses, the infected nodes became healthy all at the same rate. In the contact process, this occurs when an equal amount of antidote is sent indiscriminately to all nodes, requiring a large amount of antidote. In practice, this is often undesirable; in this paper, we will give a model that avoids such widespread antidote distribution.

Another approach is to combat epidemics by using contact tracing, or inoculating neighbors of infected nodes, using a total amount of antidote that depends only on the sum of the degrees of the infected nodes. However, both simulation and mathematical analysis have shown that contact tracing can be ineffective, especially on large real-world graphs that exhibit small-world phenomena and power-law degree distributions [7, 11, 12, 15, 16, 17].

In [4], Borgs et al. show that for the contact process on a contact graph G, inoculating every node with antidote equal to its degree will result in any infection dying out in $O(\log n)$ time with high probability, where n is the number of nodes in G. This scheme uses a total amount of antidote equal to the sum of the degrees in G. For special graphs such as the expanders, it was shown [4] that such a large amount of antidote is necessary (up to a constant factor) when a constant proportion of the nodes are initially infected. Our proposed model will not improve this result on expander graphs, but for many classes of graphs, we will require antidote only on smaller portions of the network.

In this paper, we analyze an inoculation scheme using PageRank vectors. PageRank was first introduced by Brin and Page [5] for Web search algorithms. Although the original definition is for the Web graph, PageRank is well defined for any graph, including the contact graphs that we study. Here, we will use a modified version of PageRank, known as personalized

PageRank, for the contact graph with the initially infected nodes as seeds. PageRank captures the quantitative correlation between pairs or subsets of nodes. For example, if the contact graph has some small cuts or bottlenecks, it is likely that an infection will not propagate through them, and nodes on the other side will have low PageRank. Our inoculation scheme using PageRank specifies the selected nodes for sending antidote and provides a probabilistic guarantee for the termination of any epidemic. Furthermore, the number of selected nodes for inoculation is usually much smaller than the total size of the contact graph, thus improving the previous schemes which inoculated all the nodes. The number of selected nodes depends only on the number of initially infected nodes, the probabilistic guarantee bound and the isoperimetric invariant of the graph, the *Cheeger ratio* (to be defined later). Hence, it is independent of the total size of the contact network.

Previously, an empirical study [13] found that inoculating nodes according to their PageRank works well in combating epidemics for certain examples of contact networks. Our analysis complements this experimental work and is applicable to any given general contact network. The analysis in Section 3 provides a trade-off between the probabilistic guarantee of termination and the time required.

2 Preliminaries

We model an epidemic spreading on a general undirected contact graph $G = (V, E)$ with vertex set V and edge set E. For a vertex v, let d_v denote the *degree* of v which is the number of *neighbors* of v. Suppose that the graph G has n nodes with a *degree sequence* $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, where d_i is the degree of vertex v_i. For a set of nodes $T \subseteq V$, the *volume* of T is defined to be $\mathrm{vol}(T) = \sum_{v \in T} d_v$. Let D denote the *diagonal degree matrix* $\mathrm{diag}(d_1, \ldots, d_n)$ and A the *adjacency matrix* of G, where

$$A_{ij} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

We consider a typical random walk on G with the *transition probability matrix* defined by $W = D^{-1}A$. Personalized PageRank vectors are based on random walks and W, with two governing parameters: a seed vector $\mathbf{s}$, representing an initial distribution over V, and a jumping constant α, which controls the rate of diffusion. Being a distribution, the entries of $\mathbf{s}$ sum to 1. The personalized PageRank $\mathrm{pr}(\alpha, \mathbf{s})$ is defined to be the solution to the

following recurrence relation:

$$\mathrm{pr}(\alpha, \mathbf{s}) = \alpha\mathbf{s} + (1-\alpha)\mathrm{pr}(\alpha, \mathbf{s})W. \tag{1}$$

Here, $\mathbf{s}$ (and all other vectors) will be treated as row vectors. The original definition of PageRank defined in [5] is the special case where the seed vector is the the uniform distribution (as used in [13]).

From (1), an alternative expression for the personalized PageRank $\mathrm{pr}(\alpha, \mathbf{s})$ is a geometric sum of random walks (see [2]):

$$\mathrm{pr}(\alpha, \mathbf{s}) = \alpha \sum_{t=0}^{\infty} (1-\alpha)^t \mathbf{s} W^t. \tag{2}$$

For a subset of nodes H in a graph G, the *Cheeger ratio* $h(H)$ is a measure of the cut between H and its complement $\bar{H}$:

$$h(H) = \frac{e(H, \bar{H})}{\min(\mathrm{vol}(H), \mathrm{vol}(\bar{H}))},$$

where $e(H, \bar{H})$ denotes the number of edges $\{u, v\}$ with $u \in H$ and $v \in \bar{H}$. For a given value h, we say that H is an *h-cluster* if its Cheeger ratio $h(H)$ satisfies $h(H) \leq h$.

For an h-cluster H and a given α, the *α-core* C of H is the set of all vertices u so that the personalized PageRank on H, with seed u and jumping constant α, is at least $1 - \frac{h}{\alpha}$:

$$C = \left\{ u \in V \mid \mathrm{pr}(\alpha, \mathbf{1}_u^*)\mathbf{1}_H \geq 1 - \frac{h}{\alpha} \right\}. \tag{3}$$

It has been shown [2] that if C is the α-core of H, then $\mathrm{vol}(C) \geq \frac{1}{2}\mathrm{vol}(H)$. This indicates that there are many nodes $u \in H$ for which the personalized PageRank vector $\mathrm{pr}(\alpha, \mathbf{1}_u^*)$ has very little mass outside of H if α is larger than h..

3 Infection model and inoculation scheme

We use the following contact process (see [10]) as our infection model, which is also used in [3]. The contact process is a continuous-time Markov process parametrized by β, the infection rate, with $0 \leq \beta < 1$ and $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, the cure vector. We assume that at time $t = 0$, a seed set $S \subseteq V$ is infected. We will use $\mathbf{1}_S$ to denote the indicative vector associated with S, and $\mathbf{x}(0) = \mathbf{1}_S$.

Each node v_i has an infection state $x_i(t)$; a node is considered "healthy" if $x_i(t) = 0$, and "infected" if $x_i(t) = 1$. Thus, the entire process is characterized by a state vector $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))$. The state transitions are as follows:

- If a node x_j is infected, an adjancent node x_i becomes infected at rate β. We refer to this transition as a *spread event*.
- An infected node v_i becomes healthy at rate c_i. We refer to this transition as a *cure event*.

In any continuous-time Markov process, for a transition (e.g., spread or cure event) that occurs with rate λ, the elapsed time until that transition takes place assumes an exponential random variable with parameter λ, which is independent of any state information. Such a random variable has a probability density function $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and 0 otherwise. We denote such a random variable as $\text{Expo}(\lambda)$.

Using this contact process as a model, our goal is to choose $\mathbf{c}$ such that with high probability, the infection dies out quickly, and the total amount of antidote used is small. Furthermore, we want $\mathbf{c}$ to only depend on the seed set S and the degree distribution $\mathbf{d}$, but not on t or $\mathbf{x}(t)$. Our main theorem describes how to find such a cure vector $\mathbf{c}$. First, we establish the relationship between PageRank and the infection starting from S but leaving a specified area.

Theorem 1 *Suppose that an infection starts in $S \subseteq H \subseteq V$ with infecction rate β, and each node $v \in H$ is inoculated with $c_v = d_v$. Let $\mathcal{E}_H$ denote the event that an infection started in S ever leaves the set H. Then $\mathcal{E}_H$ can be upper bounded by the PageRank vector as follows:*

$$\mathbf{P}(\mathcal{E}_H) \leq \frac{s}{\beta} \text{pr}\left(1-\beta, \frac{\mathbf{1}_S}{s}\right) \mathbf{1}^*_{\bar{H}}.$$

The proof of Theorem 1 will be given in Section 5. Using Theorem 1, we can further derive the following:

Theorem 2 *Let G be a contact graph with n nodes, S be an initial set of infected nodes with $|S| = s$, and β be the infection rate with $0 \leq \beta < 1$. Suppose that H is an h-cluster that contains S in its $(1-\beta)$-core. If all nodes in H are inocluated with antidote equal to their degrees, then with probability at least $1 - 2\frac{sh}{\beta(1-\beta)}$, any infection starting from S will die out in at most $c\log(1/h) + c'$ time, where c and c' depend only on β and not on n.*

Theorem 2 will be proved in Section 5.

We remark that Theorem 2 implies a tradeoff between the Cheeger ratio h and the probabilistic bound. If the initial set of infected nodes S lie within the $(1-\beta)$-core of a h-cluster H, the probability of the infection dying out in $O(\log s)$ time is high, as long as the product sh is small. In particular, if the seed set S lies on one side of a small cut, it will likely lie within the core of an h-cluster with small Cheeger ratio h. If there is no such small cut, then the infection is likely to spread about the graph. This leads to the following corollary:

Corollary 1 *For any $\epsilon > 0$ and an infection starting from a seed set S, if S lies within the $(1-\beta)$-core of an h-cluster H, and $h \leq \frac{\epsilon}{s}\beta(1-\beta)$, then with probability at least $1-2\epsilon$, the infection will die out in $c \log s + c'$ time, where c and c' depend only on β and ϵ and not on n.*

The proof of the corollary follows from applying the bound on ϵ to Theorem 2. We note that the above corollary explicitly relates the desired probabilistic guarantee ϵ with the Cheeger ratio h of the h-cluster containing S in its core.

The above theorems suggest the following inoculation scheme:
InoculationScheme(G, S, β, ϵ)
Input: a contact graph G, an initial set S of s infected nodes, the infection rate β and the error bound ϵ.

- Set $h = \frac{\epsilon}{s}\beta(1-\beta)$.
- Use **PageRank-Nibble** from [2] or **Local Partition** from [1] to find an h-cluster H containing S, if one exists.
- Check to see if S is in the $(1-\beta)$-core of H. If so, then inoculate each node $v \in H$ with $c_v = d_v$.
- If S is not in the core of H, or an h-cluster could not be found, then let $H = G$ and inoculate every node with $c_v = d_v$.

This inoculation scheme relies on being able to find a cluster H that has small Cheeger ratio h and contains S in its core. If such an H does not exist, then the algorithm will terminate with the entire graph inoculated. For example, in the case of expander graphs, the algorithm will soon terminate and is reduced to same scenario as in [4]. Nevertheless, it is likely for a general contact graph to contain small h-clusters. In such cases, only a small portion

of the graph needs to be inoculated, while the desired performance guarantee is maintained.

Next, we consider the case that the initially infected nodes are randomly distributed in an h-cluster H. We might expect that the infection is not likely to escape H. This is not strictly true, because if S contains some nodes near the boundary of H, it is still quite likely that the infection will escape. Nevertheless, we will be able to establish an upper bound for such probability by proving the following theorem:

Theorem 3 *Suppose $H \subseteq G$ is an h-cluster, and the set S of initially infected nodes consists of s nodes randomly and independently selected from H with probability proportional to their degrees. Suppose the infection rate is β. Then, for a given ϵ satisfying $sh \leq \epsilon$ and $s \geq \log(1/\epsilon)/\epsilon$, if all nodes in H are inocluated with antidote equal to their degrees, then with probability at least $1 - \epsilon$, any infection starting from S will die out in $c \log s + c'$ time, where c and c' depend only on ϵ and β.*

The proof of Theorem 3 will be given in Section 5. This randomized model is relevant when a disease outbreak originates in a subpopulation, which can be represented as an h-cluster in a larger population graph. Theorem 3 implies that if the subpopulation is relatively isolated from the rest of the population (i.e. the Cheeger ratio h is small), then we can effectively combat the infection by only attacking the epidemic within that h-cluster.

4 An outline of the analysis of the inoculation scheme and several useful facts

In order to prove Theorems 1 and 2, we will prove several basic facts which are need in the following brief outline of our analysis of the inoculation scheme:

- The probability of the infection spreading to a distant node is small (Lemma 1).
- The probability that a nearby node will remain infected for a long time is small (Lemmas 2, 3).
- The probability that an infection persists within the inoculated nodes is small (Lemma 4).

- The probability that an infection escapes the inoculated nodes depends on the PageRank on the uninoculated nodes $\bar{H}$, proving Theorem 1.
- After proving Theorem 1, Theorem 2 follows by showing that the personalized PageRank on $\bar{H}$ is small.

In this section, we will first give some definitions and then proceed to prove Lemmas 1–4.

Consider that if a vertex v_k is infected at some time t, then the infection must have traversed some walk in the graph from a vertex $v_0 \in S$. Suppose $\pi = (v_0, \ldots, v_k)$ is a path in G of length k. Let $\mathcal{S}_\pi$ denote the event that v_0 is infected at time 0, and the infection spreads to v_k before time t along the path π. It is important to note that if $\mathcal{S}_\pi$ occurs, v_k is not necessarily infected at time t, because it could have been cured before time t; however, v_k cannot be infected at time t if no $\mathcal{S}_\pi$ occurred.

For a vertex v, let $C(v, t)$ denote the time of the first cure event at vertex v after time t. From this, we define a *realization* of a walk as the sequence of random variables X_v as follows:

- A spread event from vertex v_i to v_{i+1} occurs at time X_{i+1}.
- $0 < X_1 < C(v_0, 0)$.
- For all $i \geq 1$, $X_i < X_{i+1} < C(v_i, X_i)$.

For $\mathcal{S}_\pi$ to occur, the infection must spread to v_k before time t; therefore, $\mathcal{S}_\pi$ occurs if and only if there is a realization of π with $X_k < t$. There are many possible realizations of π, but in our analysis, we will be concerned with a specific realization: the *canonical realization*. In this realization, given the times of all the cure and spread events, X_i is the maximum over all possible realizations with those cure and spread times. Thus, the canonical realization is the latest possible infection path along π.

With X_i as in the canonical realization, we also define an event $\mathcal{S}'_\pi$ which occurs when at least one spread event from v_i to v_{i+1} occurs between X_i and $C(v_i, X_i)$. Thus, if $\mathcal{S}'_\pi$ occurs, then the infection spreads along π to v_k, but not necessarily before time t. While we are primarily concerned with the event $\mathcal{S}_\pi$, when the spread occurs before time t, it is clear that $\mathcal{S}_\pi \subseteq \mathcal{S}'_\pi$, and using the canonical realization allows us to prove the following lemma that indicates that the probability that an infection follows a long path is small:

Lemma 1 *For any path π of length k,*

$$\mathbf{P}(\mathcal{S}_\pi) \leq \mathbf{P}(\mathcal{S}'_\pi) \leq \beta^k \prod_{j=0}^{k-1} \frac{1}{c_j}.$$

Proof: Let $\mathcal{S}_j$ denote the event that there is a spread event from v_j to v_{j+1} in between times X_j and $C(v_j, X_i)$. Due to the Markov property of the contact process, the probability of $\mathcal{S}_j$ occurring is

$$\mathbf{P}(\mathcal{S}_j) \leq \frac{\beta}{c_j}.$$

Since the curing process at every node is independent, we can write

$$\mathbf{P}(\mathcal{S}_\pi) \leq \mathbf{P}(\mathcal{S}'_\pi) = \prod_{j=0}^{k-1} \mathbf{P}(\mathcal{S}_j) = \beta^k \prod_{j=0}^{k-1} \frac{1}{c_j}.$$

□

For a walk π of length k with canonical realization $(X_i)_{i=1}^k$, we define the *canonical end time* of π to be

$$Z_\pi = \begin{cases} X_k & \text{if } \mathcal{S}_\pi \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, Z_π is the last time that v_k could become infected via the path π, or 0 if it is never infected via π. The following lemma states that the probability that Z_π is large is small:

Lemma 2 *Suppose for a path π of length k, Z_π is its canonical end time. Then,*

$$\mathbf{P}(Z_\pi > t) \leq \frac{1}{(2k)!} t^{2k-1} e^{-\beta t} \beta^k \prod_{j=0}^{k-1} \frac{1}{c_j}.$$

Proof: Let $(X_i)_{i=1}^k$ denote the canonical realization of π. Then,

$$\begin{aligned} \mathbf{P}(Z_\pi > t) &= \mathbf{P}(Z_\pi > t, \mathcal{S}_\pi) \\ &= \mathbf{P}(X_k > t, \mathcal{S}_\pi) \\ &\leq \mathbf{P}(X_k > t, \mathcal{S}'_\pi) \\ &= \mathbf{P}(X_k > t | \mathcal{S}'_\pi) \mathbf{P}(\mathcal{S}'_\pi). \end{aligned}$$

Applying Lemma 1, we have

$$\mathbf{P}(Z_\pi > t) \leq \mathbf{P}(X_k > t|\mathcal{S}'_\pi)\beta^k \prod_{j=0}^{k-1} \frac{1}{c_j}.$$

We further observe that

$$\begin{aligned}\mathbf{P}(X_{k-1} > t|\mathcal{S}'_\pi) &= \mathbf{P}\left(\sum_{i=1}^{k}(X_i - X_{i-1}) > t|\mathcal{S}'_\pi\right) \\ &\leq \mathbf{P}\left(\sum_{i=1}^{k}(C(v_i, X_{i-1}) - X_{i-1}) > t|\mathcal{S}'_\pi\right).\end{aligned}$$

We consider the time between X_{i-1} and the first cure event at v_i after X_{i-1} subject to the condition $\mathcal{S}'_\pi$: at least one spread event occured before the cure at $C(v_i, X_{i-1})$. Therefore, the time between X_{i-1} and $C(v_i, X_{i-1})$ is at least the time for one spread event, namely, the exponential random variable $\text{Expo}(\beta)$, plus the time for one cure event, $\text{Expo}(c_i)$. Thus, we have

$$\sum_{i=1}^{k}(C(v_i, X_{i-1}) - X_{i-1}) \geq \sum_{i=1}^{k}(\text{Expo}(\beta) + \text{Expo}(c_i)).$$

Because $\beta \leq c_i$, $\text{Expo}(c_i)$ is stochastically dominated by $\text{Expo}(\beta)$. We can write

$$\sum_{i=1}^{k}(C(v_i, X_{i-1}) - X_{i-1}) \geq \sum_{i=1}^{k}(\text{Expo}(\beta) + \text{Expo}(\beta)).$$

The sum of $2k$ independent exponential random variables has a gamma distribution $\Gamma(2k, \beta)$. Therefore,

$$\begin{aligned}\mathbf{P}(X_{k-1} > t|\mathcal{S}'_\pi) &\leq \frac{1}{(2k)!}\int_t^\infty x^{2k-1}e^{-x}dx \\ &\leq \frac{t^{2k-1}e^{-\beta t}}{(2k)!}.\end{aligned}$$

Putting all of this together, the lemma immediately follows. □

The next lemma addresses the question of whether or not a vertex v is infected at time t. Note that $\mathcal{S}_\pi$ only addresses whether or not a vertex was infected via π at some time before t; it could be cured thereafter.

If $(X_i)_{i=1}^k$ is the canonical realization of π, then we say that v is infected at time t via path π if $\mathcal{S}_\pi$ occurs and the first cure event at v_k after X_k does not occur until after time t. We denote this event by $\mathcal{T}_{\pi,t} = \mathcal{S}_\pi \cap \{C(v_k, X_k) > t\}$.

Lemma 3 *Suppose π is a walk of length k, and the amount of antidote at v_k is $c_k \geq \beta$. Then,*

$$\mathbf{P}(\mathcal{T}_{\pi,t}) < e^{-\beta t/2}\left(1 + \frac{1}{(2k)!}(t/2)^{2k-1}\right)\beta^k \prod_{j=0}^{k-1}\frac{1}{c_j}.$$

Proof:

We first note that the elapsed time from X_k to the cure event $C(v_k, X_k)$ is an exponential random variable with parameter c_k, independent of $\mathcal{S}_\pi$. Thus, we can write

$$\begin{aligned}
\mathbf{P}(\mathcal{T}_{\pi,t}) &= \mathbf{P}(C(v_k, X_k) > t, \mathcal{S}_\pi) \\
&= \mathbf{P}(\text{Expo}(c_k) > t - X_k, \mathcal{S}_\pi) \\
&\leq \mathbf{P}(\text{Expo}(c_k) > t/2, \mathcal{S}_\pi, X_k \leq t/2) + \mathbf{P}(X_k > t/2, \mathcal{S}_\pi) \\
&\leq \mathbf{P}(\text{Expo}(c_k) > t/2)\mathbf{P}(\mathcal{S}_\pi) + \mathbf{P}(X_k > t/2, \mathcal{S}_\pi).
\end{aligned}$$

From the definition of the canonical end time Z_π, if $\mathcal{S}_\pi$ occurs, then $X_k = Z_\pi$. Therefore, we have

$$\mathbf{P}(\mathcal{T}_{\pi.t}) \leq \mathbf{P}(\text{Expo}(c_k) > t/2)\mathbf{P}(\mathcal{S}_\pi) + \mathbf{P}(Z_\pi > t/2, \mathcal{S}_\pi).$$

Using Lemma 2, we can write

$$\mathbf{P}(Z_\pi > t/2, \mathcal{S}_\pi) \leq \frac{1}{(2k)!}(t/2)^{2k-1}e^{-\beta t/2}\beta^k \prod_{j=0}^{k-1}\frac{1}{c_j}.$$

Meanwhile, from the exponential distribution and Lemma 1,

$$\mathbf{P}(\text{Expo}(c_k) > t/2)\mathbf{P}(\mathcal{S}_\pi) \leq e^{-c_k t/2}\beta^k \prod_{j=0}^{k-1}\frac{1}{c_j} \leq e^{-\beta t/2}\beta^k \prod_{j=0}^{k-1}\frac{1}{c_j}.$$

This implies

$$\begin{aligned}
\mathbf{P}(\mathcal{T}_{\pi,t}) &\leq \mathbf{P}(\text{Expo}(c_k) > t/2)\mathbf{P}(\mathcal{S}_\pi) + \mathbf{P}(Z_\pi > t/2, \mathcal{S}_\pi) \\
&\leq e^{-\beta t/2}\left(1 + \frac{1}{(2k)!}(t/2)^{2k-1}\right)\beta^k \prod_{j=0}^{k-1}\frac{1}{c_j}.
\end{aligned}$$

□

For a path $\pi = (v_0, \ldots, v_k)$, we say that π is *safe* if $c_{v_i} \geq d_{v_i}$ for $0 \leq i \leq k$.

We denote by $\mathcal{P}_k$ the set of paths originating in S of length exactly k, and we define $\mathcal{P}'_k$ correspondingly for safe paths. We will prove the following lemma which states that the probability that an infection persists within the inoculated nodes can be made arbitrarily small by choosing an appropriate length of time:

Lemma 4 *Suppose G is a contact graph on n nodes with seed set S of size s. Then for any $h > 0$ and any infection rate β, then*

$$\mathbf{P}\left(\bigcup_{\pi\in\mathcal{P}'} \mathcal{T}_\pi\right) \leq \sum_{\pi\in\mathcal{P}'} \mathbf{P}(\mathcal{T}_\pi) \leq \frac{sh}{\beta(1-\beta)}.$$

if

$$t \geq 8\left(\frac{\log(1/h)+\log(2\beta)}{\min(\beta,\beta\log(1/\beta))}\right) = c\log(1/h)+c',$$

where c and c' only depend on β and not on n or s.

Proof: Our strategy is to analyze short paths and long paths separately. For long paths, we will use Lemma 1, and for short paths, we will use Lemma 3.

Building off our definitions of $\mathcal{P}_k$ and $\mathcal{P}'_k$, we define $\mathcal{P}_{\geq k} = \bigcup_{j=k}^{\infty} \mathcal{P}_j$, and $\mathcal{P}'_{\geq k}, \mathcal{P}_{<k}$, and $\mathcal{P}'_{<k}$ accordingly. Let k_0 be the cutoff between long and short paths to be determined later. For paths of length at least k_0, we observe that

$$\begin{aligned}
\sum_{\pi\in\mathcal{P}'_{\geq k_0}} \mathbf{P}(\mathcal{T}_\pi) &\leq \sum_{k=k_0}^{\infty} \sum_{\pi\in\mathcal{P}'_k} \mathbf{P}(\mathcal{S}_\pi) \\
&\leq \sum_{k=k_0}^{\infty} \sum_{v_0\in S} \sum_{v_1\sim v_0} \cdots \sum_{v_k\sim v_{k-1}} \mathbf{P}(\mathcal{S}_{(v_0,\ldots,v_k)}) \\
&\leq \sum_{k=k_0}^{\infty} \sum_{v_0\in S} \sum_{v_1\sim v_0} \cdots \sum_{v_k\sim v_{k-1}} \beta^k \prod_{j=0}^{k-1} \frac{1}{c_j} \\
&\leq \sum_{k=k_0}^{\infty} \sum_{v_0\in S} \sum_{v_1\sim v_0} \cdots \sum_{v_k\sim v_{k-1}} \beta^k \prod_{j=0}^{k-1} \frac{1}{d_j} \\
&= \frac{s\beta^{k_0}}{1-\beta}.
\end{aligned}$$

On the other hand, for paths of length less than k_0, we have

$$\begin{aligned}
\sum_{\pi\in\mathcal{P}'_{<k_0}} \mathbf{P}(\mathcal{T}_\pi) &\le \sum_{k=0}^{k_0-1} \sum_{\pi\in\mathcal{P}'_k} \mathbf{P}(\mathcal{T}_\pi) \\
&\le \sum_{k=0}^{k_0-1} \sum_{v_0\in S} \sum_{v_1\sim v_0} \cdots \sum_{v_k\sim v_{k-1}} \mathbf{P}(\mathcal{T}_\pi) \\
&\le \sum_{k=0}^{k_0-1} \sum_{v_0\in S} \sum_{v_1\sim v_0} \cdots \sum_{v_k\sim v_{k-1}} e^{-\beta t/2}\left(1+\frac{1}{(2k)!}(t/2)^{2k-1}\right)\beta^k \prod_{j=0}^{k-1}\frac{1}{c_j} \\
&\le s\sum_{k=0}^{k_0-1} e^{-\beta t/2}\beta^k\left(1+\frac{(t/2)^{2k-1}}{(2k)!}\right) \\
&\le sk_0 e^{-\beta t/2}\left(1+\frac{(t/2)^{2k_0-1}}{(2k_0)!}\right).
\end{aligned}$$

Combining the bounds for short and long paths yields

$$\sum_{\pi\in\mathcal{P}'} \mathbf{P}(\mathcal{T}_\pi) \le s\left(\frac{\beta^{k_0}}{1-\beta} + k_0 e^{-\beta t/2}\left(1+\frac{(t/2)^{2k_0-1}}{(2k_0)!}\right)\right)$$

We choose $k_0 = \beta t/8$. Because $t \ge 8/\beta$, we can use Stirling's approximation to derive the bound

$$e^{\beta t/4} \ge k_0 + \frac{(t/2)^{2k_0}}{(2k_0)!}.$$

Thus we have

$$\begin{aligned}
\sum_{\pi\in\mathcal{P}'} \mathbf{P}(\mathcal{T}_\pi) &\le s\left(\frac{\beta^{k_0}}{1-\beta} + e^{-\beta t/4}\right) \\
&\le s\left(\frac{\beta^{\beta t/8}}{1-\beta} + e^{-\beta t/4}\right) \\
&\le \frac{sh}{\beta(1-\beta)}
\end{aligned}$$

by using the assumption $t \ge 8\left(\frac{\log(1/h)+\log(2\beta)}{\beta\log(1/\beta)}\right)$, $\beta^{\beta t/8} \le \frac{sh}{2\beta(1-\beta)}$. This completes the proof of Lemma 4.

□

5 Proof of the main theorems

We are now ready to prove Theorem 1. Suppose that an infection starts in $S \subseteq H \subseteq V$, and each node $v \in H$ is inoculated with $c_v = d_v$. Recall that $\mathcal{E}_H$ denotes the event that an infection started in S ever leaves the set H. Theorem 1 states that $\mathcal{E}_H$ satisfies

$$\mathbf{P}(\mathcal{E}_H) \leq \frac{s}{\beta}\mathrm{pr}\left(1-\beta, \frac{\mathbf{1}_S}{s}\right)\mathbf{1}_{\bar{H}}^*.$$

Proof of Theorem 1: Let $\mathcal{B}_k$ denote the set of all paths of length k from S to $\bar{H}$ such that the first $k-1$ steps are in H. We define $\mathcal{B}$ to be the union of all $\mathcal{B}_k$. Note that if $u \in \bar{H}$ is ever infected, then $\mathcal{S}_\pi$ occurs for some $\pi \in \mathcal{B}$. We will bound that probability using the union bound:

$$\begin{aligned}
\sum_{\pi\in\mathcal{B}} \mathbf{P}(\mathcal{S}_\pi) &\leq \sum_k \sum_{\pi\in\mathcal{B}_k} \mathbf{P}(\mathcal{S}_\pi) \\
&\leq \sum_k \sum_{v_0\in S} \sum_{v_k\in\bar{H}} \sum_{\pi=(v_0,\ldots,v_k)\in\mathcal{B}_k} \mathbf{P}(\mathcal{S}_\pi) \\
&\leq \sum_k \sum_{v_0\in S} \sum_{v_k\in\bar{H}} \sum_{\pi=(v_0,\ldots,v_k)\in\mathcal{B}_k} \beta^k \prod_{j=0}^{k-1} \frac{1}{d_j} \\
&= \sum_k \mathbf{1}_S \beta^k (D^{-1}A)^k \mathbf{1}_{\bar{H}}^* \\
&= \sum_k \mathbf{1}_S \beta^k W^k \mathbf{1}_{\bar{H}}^* \\
&= \frac{s}{\beta}\mathrm{pr}\left(1-\beta, \frac{\mathbf{1}_S}{s}\right)\mathbf{1}_{\bar{H}}^*,
\end{aligned}$$

where we use (1), the assumption on H and apply Lemma 1.

Note that if π is a path that contains vertices in $\bar{H}$, then it has an initial segment $\bar{\pi} \in \mathcal{B}$, and $\mathcal{S}_\pi \subseteq \mathcal{S}_{\bar{\pi}}$. The set of such walks is $\mathcal{P}\setminus\mathcal{P}'$; we have shown that

$$\begin{aligned}
\mathbf{P}\left(\bigcup_{\pi\in\mathcal{P}\setminus\mathcal{P}'} \mathcal{T}_\pi\right) &= \mathbf{P}\left(\bigcup_{\pi\in\mathcal{B}} \mathcal{T}_\pi\right) \\
&\leq \sum_{\pi\in\mathcal{B}} \mathbf{P}(\mathcal{S}_\pi) \\
&\leq \frac{s}{\beta}\mathrm{pr}\left(1-\beta, \frac{\mathbf{1}_S}{s}\right)\mathbf{1}_{\bar{H}}^*.
\end{aligned}$$

Thus, we have shown that the probability that the infection leaves H depends on the personalized PageRank on $\bar{H}$. □

We are now ready to prove our next theorem.

Proof of Theorem 2: By the assumptions, H is a cluster with S contained in its $(1-\beta)$-core. Thus, for $u \in S$, we have from (3)

$$\mathrm{pr}(1-\beta, \mathbf{1}_u^*)\mathbf{1}_{\bar{H}}^* \leq \frac{h}{1-\beta}.$$

Summing over all $u \in S$ gives

$$\begin{aligned} \mathrm{pr}(1-\beta, \frac{\mathbf{1}_S}{s})\mathbf{1}_{\bar{H}}^* &= \frac{1}{s}\sum_{u\in S} \mathrm{pr}(1-\beta, \mathbf{1}_u)\mathbf{1}_{\bar{H}}^* \\ &\leq \frac{h}{1-\beta}. \end{aligned}$$

Applying this bound to Theorem 1 gives

$$\mathbf{P}(\mathcal{E}_H) \leq \frac{sh}{\beta(1-\beta)}.$$

Thus, the probability that the infection escapes H is at most $\frac{sh}{\beta(1-\beta)}$. Because $c_i = d_i$ for $v_i \in H$, all paths within H are safe paths, and we can apply Lemma 4 to bound the probability that the infection persists in H. Lemma 4 implies that the probability that the infection persists in H for longer than $c\log(1/h) + c'$ time is also at most $\frac{sh}{\beta(1-\beta)}$. Combining these two results, the probability that the infection persists anywhere on the contact graph for longer than $c\log s + c'$ time is at most $2\frac{sh}{\beta(1-\beta)}$. □

Theorem 3 implies that if S is chosen randomly from an h-cluster H, then there is a high probability that it is also in the core of H. This is important, because if S is in the core of H, then we can effectively combat any infection starting from S by only inoculating H. The proof is similar to the analysis involved in local partitioning algorithms using PageRank [2].

Proof of Theorem 3: Suppose we are given an h-cluster H, and S is formed by selecting s random vertices from H, independently with probability proportional to their degrees. Suppose v is one of those s nodes, and let X be a random variable that marks the amount of personalized PageRank contained in $\bar{H}$:

$$X = \mathrm{pr}(\alpha, \mathbf{1}_v)\mathbf{1}_{\bar{H}}^*.$$

From [2], we have

$$\mathbf{E}(X) \leq \frac{h}{2\alpha},$$

where the expectation is over the possible nodes $v \in H$. Futhermore, since $X \leq 1$, we can bound the variance by $\mathrm{Var}(X) \leq \mathbf{E}(X)$.

Here we take $\alpha = 1 - \beta$. Since we are selecting s random vertices, we consider $Y = \sum_{i=1}^{s} X_i$ where X_i is a copy of X. We are interested in bounding

$$\mathbf{P}(Y \geq \frac{sh}{\alpha}) \leq \mathbf{P}(Y \geq 2\mathbf{E}(Y)).$$

Using Chernoff's inequality and the known bound for $\mathbf{E}(X)$, we have

$$\begin{aligned} \mathbf{P}(Y \geq 2s\mathbf{E}(X)) &\leq e^{-s\mathbf{E}(X)^2/(2\mathrm{Var}(X))} \\ &\leq e^{-sh/(4\alpha)} \\ &\leq \epsilon \end{aligned}$$

since $sh/(4\alpha) \geq \log(1/\epsilon)$.

By Theorem 1, with probability at most $1-\epsilon$, the event that the infection starting from S leaves H satisfies

$$\begin{aligned} \mathbf{P}(\mathcal{E}_H) &\leq \frac{s}{\beta}\mathrm{pr}\left(1-\beta, \frac{\mathbf{1}_S}{s}\right)\mathbf{1}_{\bar{H}}^* \\ &\leq \frac{sh}{4\alpha\beta} \\ &\leq \frac{sh}{4(1-\beta)\beta} \\ &\leq \epsilon \end{aligned}$$

by the assumption $sh \leq \epsilon$. This completes the proof of Theorem 3. □

6 Concluding remarks

There are many questions remaining, several of which we mention here:

1. In this paper, we show that if s infected nodes are in the core of a h-cluster H and the product of s and h is small, then we only need to inoculate nodes in H so that the infection will die out in $O(\log s)$ time with high probability. Is it possible to improve or replace the condition imposed on the product sh?

2. In our main theorems, our analysis involves the Cheeger ratio which is one of the parameters concerning the structure of a graph. It will be desirable if other structural parameters can help improve the probabilistic bounds in the statement of Theorem 2, for example.

3. In this paper, we consider a fixed infection rate β and ask how little antidote can be used while still ensuring the contact process dies out quickly. The other natural approach to this problem is to fix an amount of antidote and ask for what range of β will the disease necessarily die out quickly .

4. One can also consider alternative models of contact process where cured nodes may or may not susceptible to reinfection. In addition, the type of propagation on networks can be different.

Many related interesting questions remain to be answered.

References

[1] R. Andersen and F. Chung, Detecting sharp drops in PageRank and a simplified local partitioning algorithm, *Proceedings of Theory and Applications of Models of Computation 2007*, 1–12.

[2] R. Andersen, F. Chung and K. Lang, Local graph partitioning using PageRank vectors, *Proceedings of the 47th Annual IEEE Symposium on Foundation of Computer Science (FOCS 2006)*, 475–486.

[3] N. Berger, C. Borgs, J. Chayes and A. Saberi, On the spread of viruses on the internet, *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, 301–310.

[4] C. Borgs, J. Chayes, A. Ganesh and A. Saberi, How to distribute antidote to control epidemics, *Random Structures and Algorithms*, to appear.

[5] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems* 30 (1998), 107–117.

[6] L. Chen and K. Carley, The impact of network topology on the spread of anti-virus countermeasures, *Proceedings of NAACSOS Conference (2003)*.

[7] Z. Dezso and A. Barabasi, Halting viruses in scale-free networks, *Physical Review E* 65 (2002), 055103.

[8] G. Forster and C. Gilligan, Optimizing the control of disease infestations at the landscape scale, *Proceedings of the National Academy of Sciences*, 104 (2007), 4984–4989.

[9] A. Ganesh, L. Massoulie and D. Towsley, The effect of network topology on the spread of epidemics, *Proceedings of the 24th Annual IEEE Conference on Computer Communications (INFOCOM 2005)* 2, 1455–1466.

[10] W. Kermack and A. McKendric, A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London A*, 115 (1927), 700–721.

[11] I. Kiss, D. Green and R. Kao, Infections disease control using contact tracing in random and scale-free networks, it Journal of the Royal Society Interface 3 (2005), 55–62.

[12] R. May and A. Lloyd, Infection dynamics on scale-free networks, *Physical Review E* 64 (2001), 066112.

[13] J. Miller and J. Hyman, Effective vaccination strategies for realistic social networks, *Physica A* 386 (2007), 780–785.

[14] M. Newman, The spread of epidemic disease on networks, *Physical Review E* 66 (2002), 016128.

[15] R. Pastor-Sattoras and A. Vespignani, Epidemic spreading in scale-free networks, *Physical Review Letters* 86 (2001), 3200–3203

[16] R. Pastor-Sattoras and A. Vespignani, Epidemic dynamics in finite size scale-free networks, *Physical Review E* 65 (2002), 035108.

[17] L. Tsimring and R. Huerta, Modeling of contact tracing in social networks, *Physica A*, 325 (2003), 33–39.