Controlling Epidemic Spread using Probabilistic Diffusion Models on Networks

Amy Babay
University of Pittsburgh

Michael Dinitz
Johns Hopkins University

Aravind Srinivasan
University of Maryland

Leonidas Tsepenekas
University of Maryland

Anil Vullikanti
University of Virginia

Abstract

The spread of an epidemic is often modeled by an SIR random process on a social network graph. The MinInfEdge problem for optimal social distancing involves minimizing the expected number of infections, when we are allowed to break at most B edges; similarly the MinInfNode problem involves removing at most B vertices. These are fundamental problems in epidemiology and network science. While a number of heuristics have been considered, the complexity of these problems remains generally open. In this paper, we present two bicriteria approximation algorithms for MinInfEdge, which give the first non-trivial approximations for this problem. The first is based on the cut sparsification result of Karger (1999), and works when the transmission probabilities are not too small. The second is a *Sample Average Approximation (SAA)* based algorithm, which we analyze for the Chung-Lu random graph model. We also extend some of our results to tackle the MinInfNode problem.

1 INTRODUCTION

With the COVID-19 pandemic and future such pandemics in mind, *computational epidemiology*, powered by AI and efficient algorithms, has emerged as a vital discipline. There are two major sources of uncertainty in typical applications of computational epidemiology: how the disease will unfold probabilistically (we may have a good model for this, but have limited control over such stochasticity), and models for contact between members of a population (social-contact networks). In this work, we take a rigorous stochastic-optimization approach to develop provably-good approximation algorithms for budgeted epidemic control under such sources of uncertainty.

The most widespread tool for modeling the spread of an epidemic in a social-contact network $G = (V, E)$ are SIR random processes (Pastor-Satorras et al., 2015; Marathe and Vullikanti, 2013). According to those, the infection starts at a given set S of vertices, where without loss of generality we can assume that these nodes are merged into a single infectious vertex s (this is formally explained in Section 1.1). Afterwards, if any node $u \in V$ gets infected, it spreads the disease independently to each "healthy" neighbor $v \in V$ of it with probability $p_{u,v}$—also denoted p_e—where $e = (u, v) \in E$ is an edge of the given contact network.

In order to control and mitigate the spread of the disease, there are two primary interventions studied in the literature. The first involves *social distancing*, and is modeled as removing a subset $F \subseteq E$ of edges from the graph. The second corresponds to *vaccination*, and is modeled by removing a set $V' \subseteq V$ of nodes. There is a significant cost associated with removing nodes and edges, and this motivates the problems we study in this paper. In MinInfEdge, the goal is to choose a set F of edges for social distancing, so that the cost of F is at most some budget B, and the expected number of infections is minimized; similarly, MinInfNode involves removing a subset of at most B nodes to minimize the expected number of infections.

Despite the significant importance of the MinInfEdge and MinInfNode problems, they both remain quite open. A number of heuristics have been proposed, which choose edges or nodes based on local structural properties, e.g., degree, centrality and eigen-

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

vector components. However, these do not give any guarantees in general, except in very special random graph models, e.g., (Bollobás and Riordan, 2004). The only prior work on MinInfEdge with rigorous guarantees is for the case of deterministic graphs, where we assume $p_{u,v} = 1$ for all $(u, v) \in E$ (Hayrapetyan et al., 2005a; Eubank et al., 2004, 2006; Svitkina and Tardos, 2004). This scenario actually models a highly-contagious disease, and can be viewed as the SI model (Marathe and Vullikanti, 2013). In this paper, we obtain the first rigorous results for both MinInfEdge and MinInfNode.

1.1 Formal Problem Definition

Suppose we have an undirected graph $G = (V, E)$ with edge weights $c_e \geq 0$ for every edge $e \in E$ (representing the cost of removing the social connection e). Finally, let $n = |V|$ and $m = |E|$.

We assume an SIR model of disease spread, in which each node is in one of the states S (susceptible), I (infectious) or R (recovered). We also assume the infection starts at a subset $I_0 \subseteq V$. An infectious vertex v infects each susceptible neighbor $u \in N(v)$ once, independently with probability $p_{u,v} \in [0, 1]$, where $N(v) = \{u \in V : (u, v) \in E\}$. This is equivalent to a percolation process (Pastor-Satorras et al., 2015; Marathe and Vullikanti, 2013): consider a random subgraph $G(\vec{p}) = (V, E(\vec{p}))$ obtained by retaining each edge $e \in E$ independently with probability p_e (and thus removing each edge with probability $1 - p_e$). In particular, the probability that a set V_{inf} of vertices is reachable from I_0 in $G(\vec{p})$ is precisely equal to the probability that the set V_{inf} becomes infected during the SIR process. We will sometimes abuse notation and let $G(\vec{p})$ also represent the distribution over subgraphs thus obtained. Without loss of generality, we can assume that I_0 consists of a single vertex s, since we can add a meta-vertex s with edges to all vertices in I_0 with probability 1. Finally, some of our results assume a uniform probability setting in which $p_e = p$ for all $e \in E$; in this case we denote the random graph $G(\vec{p})$ by $G(p)$.

A social distancing strategy corresponds to the removal of a subset $F \subseteq E$ of edges; for such an F, we denote by $inf(V, E \setminus F, s)$ the number of vertices that are in the same connected component as s in the *residual* graph $G_F = (V, E \setminus F)$. For simplicity, we refer to F as a cut or a cut-set, though it need not always induce a cut in the graph for our problems of interest. The expected number of infected vertices in the percolation process is then $\mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus F, s)]$.

For the vaccination intervention, $F \subseteq V$ is a set of nodes to be removed, and in this case, $G_F = (V, E - \{(u, v) \in E : u \in F \text{ or } v \in F\})$ is the subgraph obtained by removing edges incident to nodes in F; here $inf(V, E \setminus \{(u, v) \in E : u \in F \text{ or } v \in F\}, s)$ will denote the number of infected vertices when the edges incident to vertices in F are removed. For the expected number of infections in the percolation process we use $\mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus \{(u, v) \in E(\vec{p}) : u \in F \text{ or } v \in F\}, s)]$.

The MinInfEdge Problem: Besides the already-described input, we are given a budget B, and the goal is to choose a set $F \subseteq E$ of edges such that:

1. $c(F) = \sum_{e \in F} c_e \leq B$, i.e., the total cost of the set F of edges to be removed, is at most B.

2. $\mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus F, s)]$, i.e., the expected number of nodes reachable from s when we remove the edges in F and conduct the disease percolation on the remaining graph, is minimized.

The MinInfNode Problem: Besides the already-described input, we are given a budget B, and the goal is to choose a set $F \subseteq V$ of vertices such that:

1. $|F| \leq B$, i.e., the total number of removed vertices is at most B.

2. $\mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus \{(u, v) \in E(\vec{p}) : u \in F \text{ or } v \in F\}, s)]$ is minimized.

(α, β)-approximation. As in (Hayrapetyan et al., 2005a; Eubank et al., 2004), we focus on bicriteria algorithms. We define this notion only for MinInfEdge, since the case of MinInfNode is almost identical. We say that a solution $F \subseteq E$ is an (α, β)-approximation if $c(F) \leq \alpha B$, and $\mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus F, s)] \leq \beta \mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus F^*, s)]$, where F^* is an optimal solution for the given instance.

Random Models for Networks: With the ever-growing importance of networks and network science, we need good random-graph models for predictive applications, simulations, testing of new algorithms etc.: see, e.g., (Barabási and Albert, 1999; Bollobás and Riordan, 2004).

In our context of social-contact networks, the random-graph model of Chung and Lu (2006) is particularly useful. In this model, we have vertices V, and a weight w_v for every node $v \in V$ that denotes its expected degree in the graph; let $w_{min} = \min_v w_v$ and $w_{max} = \max_v w_v$. The edges E of the graph are determined via the following random process. For all $u, v \in V$, the probability of having the (u, v) edge in E is

$$q_{u,v} = \frac{w_u w_v}{\sum_{r \in V} w_r},$$

where these edges are present independently and self-loops are allowed. A natural assumption here is that $w_{min} = O(1)$. A common instantiation of this model is with a power law, in which n_i, the number of nodes of weight i, satisfies $n_i = \Theta(n/i^{\beta})$, with $\beta > 2$ being a model parameter. In our paper, we use the power law instantiation every time we consider this model.

The random graphs captured by the Chung-Lu model are more realistic than those of the simple Erdős-Renyi model (Eubank et al., 2004). The reason for this is imposing a specified degree sequence that models the heavy-tailed nature of real-world degree distributions.

We refer to MinInfEdge and MinInfNode when the graph $G = (V, E)$ is from the Chung-Lu model as MinInf-CL and MinInfNode-CL, respectively. The random process for constructing the graph $G = (V, E)$ in the Chung-Lu model should not be confused with the percolation process occurring on G during the spread of the disease. In the case of MinInf-CL and MinInfNode-CL, the reader can view the whole process as happening in two steps. At first, $G = (V, E)$ is chosen randomly according to the Chung-Lu model. Afterwards, the disease starts its diffusion in the chosen network according to the probability vector $\vec{p}$.

1.2 Contributions and Outline

We mostly focus on MinInfEdge for the rest of the paper. In Appendix B we discuss which of our results extend to the case of MinInfNode.

In Section 2, we study the unit edge-cost case of MinInfEdge (all edges of G have cost 1), and we present an $(O(1), O(1))$-approximation for it. This result is for the uniform p probability setting, in the regime where Karger's cut sparsification result holds. However, even this simple setting is not trivial, and to the best of our knowledge, this is the first rigorous result when the transmission probability is not 1. Let $\hat{G}$ be the weighted graph obtained by setting the weight w_e of each edge e equal to p, and let $\hat{c}$ denote the weight of the minimum cut in $\hat{G}$. Karger's result (Theorem 2.1) states that if $\hat{c} \geq 9 \ln n$, then the size of every cut in $G(p)$ is close to the corresponding cut in $\hat{G}$. In this case, we are able to reduce MinInfEdge to a problem from (Hayrapetyan et al., 2005a), using just one random sample $G(p)$.

In Section 3 we present a sampling framework for MinInfEdge that utilizes the powerful sample-average-approximation (SAA) approach (Kleywegt et al., 2002; Ruszczynski and Shapiro, 2003; Shapiro, 2003; Swamy and Shmoys, 2012). Specifically, we sample a polynomial number of graphs from $G(\vec{p})$ and then formulate a linear program (LP) that describes the empirical estimate of the optimal solution of those samples. Afterwards, we solve this LP and provide a randomized-rounding procedure that transforms its fractional solution into an integral one. Let F_0 be the solution (set of edges to remove) that we compute, OPT the value of the optimal solution, and Γ the expected number of simple paths[1] in a randomly drawn graph from $G(\vec{p})$, where the randomness also includes the random choice of G, in case G is drawn from a random-graph model.

Three different sources of randomness: Our statements will refer to (combinations of) three distinct sources of randomness/uncertainty:

- **Type 1:** This randomness is over the random choice, if any, of our network $G = (V, E)$ (such as randomness resulting from choosing G according to the Chung-Lu model). If the network G is deterministic, Type 1 is vacuous: there is no randomness.
- **Type 2:** This randomness arises from the choices of our randomized rounding algorithm.
- **Type 3:** This type of randomness refers to the random percolation/diffusion of the disease, governed by $\vec{p}$.

Our main theorem for the SAA approach of Section 3 is summarized in the following, where "log" denotes the natural logarithm throughout.

Theorem 1.1. *For any chosen constants $\epsilon > 0$ and $\gamma > 1$, the following hold:*

- *with probability at least $1 - O(n^{-\gamma})$, where the randomness is solely of Type 2, we have $c(F_0) \leq O(\frac{\gamma}{\epsilon}) \log n \cdot B$;*
- *there exists an event $\mathcal{A}$ with $\Pr[\mathcal{A}] \geq 1 - O(\frac{1}{n^2}) - O\big(\frac{\Gamma \log n}{\epsilon^2 n^{\gamma}}\big)$ and $\mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s) \mid \mathcal{A}] \leq (1+\epsilon) \cdot OPT$. Here, randomness is with respect to Type 1 (if applicable), Type 2, and Type 3.*

Observe now that if $\Gamma \leq \text{poly}(n)$[2], we can choose γ to be large enough, such that $\Pr[\mathcal{A}] \geq 1 - O(1/n^2)$. As we show in Section 3 this immediately implies the following corollary.

Corollary 1.2. *When $\Gamma \leq \text{poly}(n)$, we easily get $\mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s)] \leq (1 + O(\epsilon) + O(1/n))OPT$, where the randomness is with respect to Type 1 (if applicable), Type 2, and Type 3.*

Hence, in Section 4 we prove that a family of Chung-Lu random-graphs satisfies the $\Gamma \leq \text{poly}(n)$ property

[1] "Paths" will refer throughout to *simple* paths: ones in which no nodes or edges are repeated.

[2] Throughout, "poly" will denote an arbitrary univariate or bivariate polynomial.

(recall this model captures realistic social-contact networks well (Eubank et al., 2004, 2006)). Under this property, are main result informally says that *we can approximate the budget to within a factor $O(\log n)$ with high probability, and the expected number of infected people to within a constant factor*.

A remark regarding Section 3 is that our goal is to present a "proof of concept", so we do not optimize the constants in our algorithms, and we are content with polynomial running times. In particular, we do not spell out the actual running times of our algorithms: these will easily be seen to be bounded by polynomials of n and m. We remark that most of the prior work on this problem has been experimental, and that our paper is the first to give rigorously-proven results.

Section 4 develops the above-mentioned poly(n) bound on Γ for a realistic Chung-Lu family of graphs. More generally, it shows a *phase-transition* phenomenon for the expected number of paths of any length k, as a function of the model parameter β: this is proved to be at most poly$(n, 2^k)$ for $\beta > 3$, and to be at least $(k!)^{\Omega(1)}$ for $\beta < 3$. This leads to our provably-good approximation algorithms for the Chung-Lu family of graph models when $\beta > 3$.

In Section 5, we show a slightly different SAA approach combined with a deterministic rounding, which achieves an $(O(n^{2/3}), O(n^{2/3}))$-approximation for any graph, without any dependence on the parameter Γ.

1.3 Further Related Work

There has been much work on heuristics for interventions for the SIR model (Yang et al., 2019; Eames et al., 2009; Cohen et al., 2003; Miller and Hyman, 2007; Barabási and Albert, 1999; Sambaturu et al., 2020). In particular, heuristics based on degree or centrality, e.g., (Cohen et al., 2003; Miller and Hyman, 2007), have been shown to be quite effective in many classes of networks (including random graphs), but these do not provide any guarantees. The work of Sambaturu et al. (2020) explores the use of the sample average approximation method, but has worst-case approximation bounds as large as $O(n)$.

However, as mentioned earlier, rigorous results are only known for the setting where $p_{u,v} = 1$ for all $(u, v) \in E$ (Bollobás and Riordan, 2004; Hayrapetyan et al., 2005a; Eubank et al., 2006; Svitkina and Tardos, 2004). The MinInfEdge problem is known to be NP-hard even in this setting (Hayrapetyan et al., 2005a; Svitkina and Tardos, 2004), and constant factor bicriteria approximation algorithms are known.

Another related direction of work has been on reducing the first eigenvalue, referred to as the spectral radius, based on a characterization of the time to die out in SIS models (in which, unlike the SIR model, an infected node switches back to state S) (Ganesh et al., 2005). There has been much work on reducing the spectral radius, e.g., (Preciado et al., 2014b, 2013, 2014a; Saha et al., 2015; Ogura and Preciado, 2017). However, these results do not imply any guaranteed bounds for MinInfEdge or MinInfNode.

2 MinInfEdge WITH UNIT EDGE-COSTS AND UNIFORM PROBABILITIES

In this section we are going to consider a special case of MinInfEdge. Specifically, we assume that the edge costs of the network $G = (V, E)$ are all 1, i.e., $c_e = 1$ for all $e \in E$. Moreover, we will work under he uniform transimition probability setting.

For a random graph $G(p) = (V, E(p))$ and any $F \subseteq E$, let $F(p) = F \cap E(p)$ be the random cut corresponding to F in $G(p)$. Let also c_{min} be the size of the smalletst cut in G. We are going to use a cut sparisification result of Karger (1999).

Theorem 2.1. *(Karger, 1999) Let $\epsilon = \sqrt{\frac{3(d+2)(\ln n)}{c_{min} \cdot p}}$ for some $d > 0$. If $\epsilon \leq 1$ then, with probability at least $1 - O(1/n^d)$, we have $\big||F(p)| - \mathbb{E}_{G(p)}[|F(p)|]\big| \leq \epsilon \mathbb{E}_{G(p)}[|F(p)|]$ for every $F \subseteq E$.*

Observation 2.2. When $c_{min} \cdot p \geq 9 \ln n$, the statement of Theorem 2.1 holds with high probability, i.e., with probability at least $1 - O(1/n)$.

Observation 2.2 basically determines the regime where the results of this section hold. However, notice that $c_{min} \cdot p \geq 9 \ln n$ is a realistic assumption, since for most real-life scenarios the transmission probability will be some constant, and the size of the minimum cut in G can very well be $\Omega(\ln n)$.

To tackle MinInfEdge in the current setting, we are going to reduce it to a problem from (Hayrapetyan et al., 2005b), namely the Minimum-Size Bounded-Capacity Cut problem (MinSBCC). In this problem, we are given a graph $G = (V, E)$, a source vertex $s \in V$, and a budget B. We are then asked to find a set $F \subseteq E$ of at most at most B edges, which minimizes the number of nodes in the same component as s in $G_F = (V, E \setminus F)$, i.e., $inf(V, E \setminus F, s)$. The main result of Hayrapetyan et al. (2005b) follows.

Theorem 2.3. *For any $\lambda \in (0, 1)$, there exists a polytime $(\frac{1}{\lambda}, \frac{1}{1-\lambda})$-approximation algorithm for MinSBCC: it finds a cut of size at most $\frac{1}{\lambda}B$, in which the number of nodes in the same component as s in the resulting subgraph is at most $\frac{1}{1-\lambda}$ times the value of the optimal solution with size B.*

Our approach for solving MinInfEdge goes as follows. At first, we sample a graph $H = (V, E')$ from $G(p)$. Then, we create an instance of MinSBCC, where the graph under consideration is H, the source vertex is s, and the budget is γBp for a small constant γ which we set later. Finally, we run the $(\frac{1}{\lambda}, \frac{1}{1-\lambda})$-approximation of Hayrapetyan et al. (2005b) on the created instance of MinSBCC, and get a solution $F' \subseteq E'$. Let now S be all the vertices that are in the same connected component as s in $H_{F'} = (V, E' \setminus F')$. Our returned solution for the original instance of MinInfEdge is $\bar{F} = \{\{u, v\} \in E : u \in S, v \notin S\}$.

Lemma 2.4. *When the assumption of Observation 2.2 holds, $|\bar{F}| \leq \frac{\gamma}{(1-\epsilon)\lambda}B$ with probability at least $1 - O(1/n)$, where $\epsilon < 1$ is as in Theorem 2.1.*

Proof. Notice that the vertices that are in the same connected as s in $(V, E \setminus \bar{F})$, are exactly those that are connected to s in $(V, E' \setminus F')$. Therefore, the random cut corresponding to $\bar{F}$ in $G(p)$ is F', i.e., $F' = \bar{F}(p)$. Hence, $\mathbb{E}_{G(p)}[F'] = \mathbb{E}_{G(p)}[\bar{F}(p)] = p|\bar{F}|$. Therefore, using Theorem 2.1, we have that with probability at least $1 - O(1/n)$:

$$\big||F'| - p|\bar{F}|\big| \leq \epsilon p|\bar{F}| \implies |F'| \geq (1-\epsilon)p|\bar{F}|$$

Since $|F'| \leq \frac{\gamma Bp}{\lambda}$ (we ran the algorithm of Theorem 2.3 with budget γBp), we get $\frac{\gamma Bp}{\lambda} \geq |F'| \geq (1-\epsilon)|\bar{F}|p$ with probability at least $1 - O(\frac{1}{n})$. Rearranging terms implies that $|\bar{F}| \leq \frac{\gamma}{(1-\epsilon)\lambda}B$. □

Lemma 2.5. *$|S| \leq \frac{\gamma}{1-\lambda}OPT$ with probability at least $1 - \frac{2}{\gamma}$, where OPT is the value of the optimal solution (the expected number of nodes infected).*

Proof. Let F^* denote the optimal solution (so $|F^*| \leq B$), and let $\hat{F} = F^* \cap E'$ be a random variable denoting the edges of F^* that are present in E'. Let also $S_{\hat{F}}$ be the random variable denoting the nodes that are in the same connected component as s in $(V, E' \setminus \hat{F})$. We say that there was a "success" in the process of sampling H if the following two conditions are satisfied: **1)** $|\hat{F}| \leq \gamma Bp$ and **2)** $|S_{\hat{F}}| \leq \gamma \cdot OPT$. If either condition is not true we say that there was a "failure".

Suppose that there was a success. Then the first condition implies that $\hat{F}$ was a feasible solution for the MinSBCC instance (since its size was within the given budget), and hence $|S| \leq \frac{1}{1-\lambda}|S_{\hat{F}}|$. Then the second condition implies $|S| \leq \frac{\gamma}{1-\lambda}OPT$ as desired.

Finally, we need to show that the probability of success is at least $1 - \frac{2}{\gamma}$, or equivalently that the probability of failure is at most $\frac{2}{\gamma}$. Clearly $\mathbb{E}_{G(p)}[|\hat{F}|] = p|F^*| \leq pB$, so by Markov's inequality $\Pr[|\hat{F}| > \gamma Bp] \leq \frac{1}{\gamma}$. Similarly, $\mathbb{E}[|S_{\hat{F}}|] = OPT$ by the definition of OPT, and so by Markov $\Pr[|S_{\hat{F}}| > \gamma \cdot OPT] \leq \frac{1}{\gamma}$. A union bound implies a failure probability of at most $2/\gamma$. □

Theorem 2.6. *When all edges have unit cost and the transmission probabilities are uniform, there exists an $(O(1), O(1))$-approximation for the* MinInfEdge *that works with high probability, as long as the assumption of Observation 2.2 holds.*

Proof. If we set γ to be a large enough constant (say, 4), then with probability at least $1/2 - O(1/n)$ we return a solution $\bar{F}$ which violates the budget by at most $O(1)$ (Lemma 2.4), and the size of the connected component in $(V, E \setminus \bar{F})$ which contains s is at most $O(1) \cdot OPT$ (Lemma 2.5). Clearly this implies that $\mathbb{E}_{G(p)}[inf(V, E(p) \setminus \bar{F}, s)]$ is also at most $O(1) \cdot OPT$. Thus, our algorithm gives the bounds in Theorem 2.6 with constant probability. By repeating this process $O(\log n)$ times and taking the best solution, this algorithm can be made to work with high probability. □

3 THE SAA PATH-DEPENDENT FRAMEWORK FOR ARBITRARY NETWORKS

Consider a general instance of MinInfEdge. For a suitable number $N \leq \text{poly}(n, m)$ that is going to be set later, we simulate the disease-percolation process on G independently N times. In other words, we independently sample N graphs $G_j = (V, E_j)$, $j = 1, 2, \ldots, N$, where $E_j \subseteq E$ is the subset of edges acquired in the j^{th} simulation (or sample), when each edge is retained with probability p_e. The heart of our approach is to then show how these "typical" samples G_j can guide us towards computing a provably-good solution for our given probabilistic percolation model.

We start by presenting the linear program LP (1)-(4). This LP models an "empirical" solution to the problem, when the diffusion process can only result in the graphs G_j, and each of these graphs materializes with probability $1/N$. We use $\mathcal{P}(s, v, G_j)$ to denote the set of paths from s to v in the graph G_j, and $[k]$ to denote the set $\{1, 2, \ldots, k\}$ for any positive integer k. For the integral version of our LP, x_e is the indicator variable for removing edge e, and y_{vj} the indicator for vertex v *not becoming reachable* from s in G_j after our edge-removal. Then, constraint (2) makes sure that v is disconnected from s in G_j iff for every path of $\mathcal{P}(s, v, G_j)$ at least one edge of the path has been removed. Constraint (3) captures the budget constraint, and the objective function (1) measures exactly the expected number of infections, when each G_j appears with probability $1/N$. Finally, in order to be able to efficiently solve the system, the $\{0,1\}$–variables are relaxed to lie in $[0,1]$.

$$\min \frac{1}{N} \sum_{j\in[N]} \sum_{v\in V} (1 - y_{vj}) \text{ such that} \tag{1}$$

$$\sum_{e\in P} x_e \geq y_{vj}, \quad \forall j,\ \forall v,\ \forall P \in \mathcal{P}(s, v, G_j) \tag{2}$$

$$\sum_{e\in E} c_e x_e \leq B \tag{3}$$

$$x_e, y_{vj} \in [0, 1], \quad \text{for all } j \in [N], v \in V, e \in E \tag{4}$$

Our algorithm involves the following steps:

1. Solve LP (1)-(4), and let x, y be the optimal fractional solution. This solution can be computed in polynomial time via the ellipsoid method, with a *separation oracle* that checks if the shortest-path distance from s to v in G_j (with edge weights x_e) is less than y_{vj} (Grötschel et al., 1988).

2. For some user-specified constant $\epsilon \in (0, 1)$, define the following sets for the sake of analysis:

   $$S(j) = \{v \in V : y_{vj} \geq \epsilon\} \text{ for every } j \in [N]$$
   $$\mathcal{P}_{hit} = \cup_j \cup_{v\in S(j)} \mathcal{P}(s, v, G_j)$$

3. Let F_0 denote the set of edges which will constitute our returned solution. For some constant γ that will be defined later, put each edge $e \in E$ independently in F_0, with probability

   $$x'_e = \min\left\{\frac{(\gamma + 5)x_e \log n}{\epsilon}, 1\right\}$$

For any fixed $F \subseteq E$, we define random variables $h(G_j, F)$ and $h(G, F)$, where the randomness here is over the choice of the G_j's, i.e., the randomness is of Type 3. Let $h(G_j, F) = inf(V, E_j \setminus F, s)$ and $h(G, F) = \frac{1}{N}\sum_{j=1}^{N} h(G_j, F)$; the former represents the number of infections in the j-th sample if F are the edges to be removed, and the latter represents the average number of infections over the N sampled graphs if again F is the set of edges removed.

For the small user-defined constant $\epsilon > 0$, we now choose $N = \frac{3n}{\epsilon^2} \log\left(n^2 \cdot 2^{m+1}\right)$ and present a simple concentration result in Lemma 3.1; note that for this choice we have $N = \text{poly}(n, m)$ and hence our algorithm runs in polynomial time.

Lemma 3.1. *For the chosen value $N = \frac{3n}{\epsilon^2} \log\left(n^2 \cdot 2^{m+1}\right)$, with probability at least $1 - \frac{1}{n^2}$, we have $h(G, F) \in \left[(1-\epsilon)\mathbb{E}[h(G, F)], (1+\epsilon)\mathbb{E}[h(G, F)]\right]$ for all sets $F \subseteq E$. The expectation here is over randomness of Type 3, and specifically over the random sampling of the N graphs G_j.*

Let $F^* = \arg\min_F \mathbb{E}[h(G, F)]$, where the expectation is again over the random sampling of the graphs G_j (Type 3 randomness). Since for every F we have $\mathbb{E}[h(G_j, F)] = \mathbb{E}_{G(\vec{p})}[inf(V, E(\vec{p}) \setminus F, s)]$ for all G_j, and $\mathbb{E}[h(G, F)] = \frac{1}{N}\sum_j \mathbb{E}[h(G_j, F)]$, we see that F^* is actually the optimal edge set for MinInfEdge. Also, we define the random variable $\hat{F} = \arg\min_F h(G, F)$, denoting the optimal integral solution of LP (1)-(4); $\hat{F}$ is actually the optimal empirical solution for the sampled set of graphs. Recall now that F_0 is the subset of edges computed by our LP rounding algorithm, and recall the parameter Γ from Section 1, indicating the expected number of paths in a randomly-drawn graph (with randomness being of types 1 and 3).

Proof. (**Theorem 1.1.**) Showing the first part of the theorem is easy. Since each edge e is removed (independently) with probability x'_e, the expected cost of the removed edges is

$$\mathbb{E}[c(F_0)] \leq \sum_e c_e x'_e \leq \frac{(\gamma + 5)\log n}{\epsilon} \sum_e c_e x_e$$
$$\leq \frac{((\gamma + 5)\log n)B}{\epsilon}$$

where the last inequality follows from constraint (3). Next, we can assume w.l.o.g. that $B = 1$. To do so, we first hard-wire $x_e = 0$ for all edges e with $c_e > B$, thus ignoring these edges in our edge-removal problem. Then, we uniformly scale all remaining c_e's and the budget by a factor of $1/B$. Using the second statement of Lemma C.1 with $R = (6(\gamma + 5)\log n)/\epsilon$ gives:

$$\Pr[c(F_0) \geq (6(\gamma + 5)\log n)/\epsilon] \leq O(1/n^{\gamma})$$

We next prove the second part of the theorem. The event $\mathcal{A}$ that is a function of the randomness of types 1, 2, 3 is the conjunction of the following three events:

- $\mathcal{A}_1$: For each $P \in \mathcal{P}_{hit}$, there exists an edge $e \in P$, such that $e \in F_0$.
- $\mathcal{A}_2$: $h(G, F^*) \leq (1 + \epsilon)\mathbb{E}[h(G, F^*)]$.
- $\mathcal{A}_3$: $h(G, F_0) \geq (1 - \epsilon)\mathbb{E}[h(G, F_0)]$.

We will first show that $\mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s) \mid \mathcal{A}] \leq (1 + O(\epsilon))OPT$, and then lower-bound $\Pr[\mathcal{A}]$.

Let us first condition on $\mathcal{A}$. Consider any $j \in [N]$. By $\mathcal{A}_1$ and the definition of the set $\mathcal{P}_{hit}$, the only vertices in $(V, E_j \setminus F_0)$ that are reachable from s can be those in $V \setminus S(j)$; these vertices are exactly the ones getting infected in the j-th sample. Further, by definition we have $y_{vj} < \epsilon$ for every $v \in V \setminus S(j)$. Therefore, the empirical number of infections over all the samples is:

$$\begin{aligned} h(G,F_0) &\le \frac{1}{N}\sum_{j\in[N]}\sum_{v\notin S(j)} 1 \\ &\le \frac{1}{N}\sum_{j\in[N]}\sum_{v\notin S(j)} \frac{1-y_{vj}}{1-\epsilon} \\ &\le \frac{h(G,\hat{F})}{1-\epsilon} \le \frac{h(G,F^*)}{1-\epsilon} \end{aligned} \tag{5}$$

The second inequality above follows because the LP value is a lower bound on $h(G,\hat{F})$, and the last inequality follows since $\hat{F}$ minimizes $h(G,F)$. Combining (5) and the definitions of $\mathcal{A}_2$, $\mathcal{A}_3$ yields

$$\begin{aligned} \mathbb{E}[h(G,F_0)] &\le \frac{h(G,F_0)}{1-\epsilon} \le \frac{h(G,F^*)}{(1-\epsilon)^2} \\ &\le \frac{(1+\epsilon)}{(1-\epsilon)^2}\mathbb{E}[h(G,F^*)] \\ &= (1+O(\epsilon))\mathbb{E}[h(G,F^*)] \end{aligned}$$

To conclude the proof we need to lower-bound $\Pr[\mathcal{A}]$. First, Lemma 3.1 shows that each of $\mathcal{A}_2$ and $\mathcal{A}_3$ holds with probability at least $1-1/n^2$. Let us consider $\mathcal{A}_1$.

Let $\mathcal{B}$ be a random variable denoting the number of paths over all the samples G_j. Since Γ is the expected number of paths in a single graph, linearity of expectation gives $\mathbb{E}[\mathcal{B}] = \Gamma N = O(\frac{n^3\Gamma}{\epsilon^2})$, since $m = O(n^2)$. Thus, by using Markov's inequality we have $\Pr[\mathcal{B} = \Omega(\frac{n^5\Gamma}{\epsilon^2})] \le O(1/n^2)$. Equivalently, $\Pr[\mathcal{B} = O(\frac{n^5\Gamma}{\epsilon^2})] \ge 1 - O(1/n^2)$. The randomness in the previous statements is of types 1 and 3.

Consider now a path $P \in \mathcal{P}_{hit}$. If there exists an $e \in P$ such that $x'_e = 1$, then this path is broken. Hence, assume that for all $e \in P$ we have $x'_e < 1$. By the definition of the paths in $\mathcal{P}_{hit}$ we also have $\sum_{e\in P} x'_e \ge (\gamma+5)\log n$. Therefore, the probability that all edges of P survive is at most

$$\prod_{e\in P}(1-x'_e) \le e^{-\sum_{e\in P} x'_e} \le e^{-(\gamma+5)\log n} \le n^{-(\gamma+5)}$$

In the end, a union bound over all $P \in \mathcal{P}_{hit}$ gives $\Pr[\mathcal{A}_1|\mathcal{B}] \ge 1 - \frac{\mathcal{B}}{n^{\gamma+5}}$. Combining everything gives $\Pr[\mathcal{A}_1] \ge (1-O(\frac{\Gamma}{\epsilon^2 n^\gamma}))(1-O(\frac{1}{n^2})) = 1 - O(\frac{\Gamma}{\epsilon^2 n^\gamma}) - O(\frac{1}{n^2})$. Hence, putting down all the lower bounds for $\mathcal{A}_1, \mathcal{A}_2$ and $\mathcal{A}_3$ yields $\Pr[\mathcal{A}] \ge 1 - O(\frac{\Gamma}{\epsilon^2 n^\gamma}) - O(\frac{1}{n^2})$. □

Corollary 3.2. *When $\Gamma \le \mathrm{poly}(n)$, we trivially get $\mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s)] \le (1 + O(\epsilon) + O(1/n))OPT$, where the randomness is with respect to Type 1 (if applicable), Type 2, and Type 3.*

4 COUNTING PATHS IN THE CHUNG-LU MODEL

Recall the random graph model of Chung and Lu (2006). Here we are given vertices V, where each vertex $v \in V$ comes with a positive integer w_v indicating its expected degree in the graph. For every pair of vertices u and v, the edge (u,v) is independently included in the graph with probability $q_{u,v} = w_u w_v / \sum_{r\in V} w_r$. Furthermore, we consider a power-law model, in which n_i, the number of nodes of weight i, satisfies $n_i = \Theta(n/i^\beta)$, where $\beta > 2$ is a given parameter. Finally, recall that $w_{max} = \max_v w_v$, $w_{min} = \min_v w_v$, and a common assumption in this setting is $w_{min} = O(1)$.

Take now any random graph $G = (V,E)$ that is produced by the above model. In that graph, we assume that a disease percolation process takes place, and this process is governed by some probability vector $\vec{p}$. We are interested in bounding the expected number of paths Γ in $G(\vec{p})$, where the randomness of Γ is obviously of both Types 1 and 3. To do so, we start by analyzing the expected number of paths of length k in G, where the randomness here is only of Type 1. In what follows, we are using ℓ_k to denote the latter.

Our first result is showing that when $\beta > 3$, we have $\ell_k \le \mathrm{poly}(n, 2^k)$. Furthermore, if $p_e \le c_0$ for all $e \in E$, where c_0 is a universal positive constant, we demonstrate how to utilize the bound on ℓ_k and eventually give a polynomial bound on Γ.

In addition, when $\beta < 3$ we provide a negative result, indicating that our SAA framework from Section 3 cannot be utilized for this case, as no polynomial bound on Γ can be guaranteed.

By an abuse of notation, we will let m denote the **expected** number (not actual number) of edges in the graph G. Trivially, $m = \sum_{v\in V} w_v/2$. Since $\beta > 2$, m can also be expressed as:

$$\begin{aligned} m = \Theta(\sum_i i\cdot n_i) &= \Theta\Big(\int_{w_{min}}^{w_{max}} \frac{n}{z^{\beta-1}}dz\Big) \\ &= \Theta\Big(\frac{n}{w_{min}^{\beta-2}}\Big) \end{aligned} \tag{6}$$

The following lemma is required for counting paths.

Lemma 4.1. *Fix some length k, and suppose that we are given a positive integer $D \ge w_{min}$. Let $S(D,k) \doteq \{(a(w_{min}), a(w_{min}+1), \ldots, a(D)) : (\forall i,\ a(i) \in \mathbb{Z}_{\ge 0})$ and $\sum_i a(i) = k\}$. Then,*

$$\ell_k \le n\cdot\Big(\frac{2^k k!}{m^k}\Big)\cdot \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \Big(\binom{n_i}{a(i)}\cdot i^{2a(i)}\Big).$$

Proof. We say that a vertex v is in *class* i if $w_v = i$. Fix a vertex v_0. We upper bound how many different paths of length k can start from v_0. A potential such path $P = (u_0, u_1, \ldots, u_k)$ can be constructed as follows:

1. Pick $\mathbf{a} = (a(w_{min}), a(w_{min}+1), \ldots, a(w_{max}))$ from $S(w_{max}, k)$. This will give us the selection of how many vertices from each degree class we should pick, such that in total we have chosen k vertices for P.

2. For the chosen $\mathbf{a}$, pick $a(i)$ vertices from each degree class i, where $w_{min} \leq i \leq w_{max}$. This is possible only if $a(i) \leq n_i$ for each class i. However, since we are only computing an upper bound, we will assume that such a selection is always possible. Notice that there may be some additional double counting, because we may end up choosing v_0 again. Nonetheless, as we are only concerned with an upper bound, we will permit such "unecessary" cases in our counting.

3. Choose the positions of the chosen k vertices among the indices $\{1, 2, \ldots, k\}$ of the path P to be constructed ($k!$ possibilities).

Overall, based on the 3 cases above, the total number of paths starting from v_0 can be at most:

$$k! \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \binom{n_i}{a(i)} \tag{7}$$

Let d_0 be the class of u_0. Suppose we complete the three steps above and let $(d_0, d_1, \ldots, d_k)$ be the ordered degree sequence obtained for the vertices in P, when the chosen vector was $\mathbf{a}$. The probability of such a path materializing in the edge selection phase is

$$\prod_{i=0}^{k-1} \left(\frac{d_i d_{i+1}}{\sum_v w_v} \right) = \frac{2^k}{m^k} \prod_{i=w_{min}}^{w_{max}} i^{2a(i)} \tag{8}$$

Combining (7) and (8), the expected number of paths of length k starting at u_0 is at most

$$\left(\frac{2^k k!}{m^k} \right) \cdot \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \left(\binom{n_i}{a(i)} \cdot i^{2a(i)} \right) \tag{9}$$

Summing this bound over all possible starting vertices results in the claim of the Lemma. □

4.1 A Positive Result When $\beta > 3$

We begin this section with a couple of important technical lemmas, and then move on to our final result regarding the expected number of paths in a randomly drawn graph.

Lemma 4.2. *Let $S(D,k)$ be as in Lemma 4.1, and let $N(D,k) \doteq \sum_{\mathbf{a}\in S(D,k)} \prod_{i=w_{min}}^{D} \frac{1}{i^{c_1 a(i)} \cdot a(i)!}$ for some constant value $c_1 > 1$. Then*

$$N(D,k) \leq \frac{1}{k!} \cdot \prod_{i=w_{min}+1}^{D} \left(1 + \frac{1}{i^{c_1}} \right)^k$$

Lemma 4.3. *Suppose $\beta = 2 + c_1$ for some constant $c_1 > 1$. Then, for all k, $\ell_k \leq \mathrm{poly}(n, 2^k)$.*

Proof. Before we proceed to our main arguments, we make some useful observations and give a bit more notation. At first, using (6) and the assumption that $w_{min} = O(1)$, we see that $\frac{n}{m} = O(1)$. Furthermore, because $n_i = \Theta(\frac{n}{i^\beta})$ for every $i \in [w_{min}, w_{max}]$, let λ be a universal constant such that $n_i \leq \frac{\lambda n}{i^\beta}$ for every i. Using Lemma 4.1 we upper bound ℓ_k as follows:

$$n \left(\frac{2^k k!}{m^k} \right) \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \left(\frac{n_i^{a(i)}}{a(i)!} i^{2a(i)} \right) \leq$$

$$n \left(\frac{2^k k!}{m^k} \right) \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \left(\frac{(\lambda n / i^{2+c_1})^{a(i)}}{a(i)!} i^{2a(i)} \right) \leq$$

$$nk! \left(\frac{\lambda n}{m} \right)^k \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \left(\frac{(1/i^{2+c_1})^{a(i)}}{a(i)!} i^{2a(i)} \right) =$$

$$nk! \left(\frac{\lambda n}{m} \right)^k \sum_{\mathbf{a}\in S(w_{max},k)} \prod_{i=w_{min}}^{w_{max}} \frac{1}{i^{c_1 a(i)} a(i)!} =$$

$$\mathrm{poly}(n, 2^k) \cdot k! \cdot N(w_{max}, k).$$

Using the bound on $N(D,k)$ from Lemma 4.2, we have

$$\begin{aligned} \ell_k &\leq \mathrm{poly}(n, 2^k) \cdot \prod_{i=w_{min}+1}^{\infty} \left(1 + \frac{1}{i^{c_1}} \right)^k \\ &\leq \mathrm{poly}(n, 2^k) \cdot \prod_{i=w_{min}+1}^{\infty} e^{\frac{k}{i^{c_1}}} \\ &= \mathrm{poly}(n, 2^k) \cdot e^{k \cdot \sum_{i > w_{min}} \frac{1}{i^{c_1}}} \\ &\leq \mathrm{poly}(n, 2^k) \end{aligned}$$

The last inequality follows because $\sum_{i > w_{min}} (1/i^{c_1}) = O(1)$ when $c_1 > 1$. □

Corollary 4.4. *Let G be a graph drawn from the Chung-Lu distribution with power law weights, with parameter $\beta = 2 + c_1$ for some constant $c_1 > 1$. Then there is a constant $c_0 > 0$ that depends only on c_1, such that the following holds: if the probability p_e of retaining edge e during the disease percolation process satisfies $p_e \leq c_0$ for any e, then the expected number Γ of paths in $G(\vec{p})$ is upper-bounded by* $\mathrm{poly}(n)$. *(This expectation is over randomness of types 1 and 3.)*

Proof. From Lemma 4.3, we have $\ell_k \leq \text{poly}(n, 2^k)$; let C be a constant such that $\ell_k \leq n^C 2^{Ck}$ for all k. We choose $c_0 = 2^{-C}$. Then, when $p_e \leq c_0$ for every e, the probability that a given path of length k survives in $G(\vec{p})$ is at most c_0^k. Therefore, the expected number of paths in $G(\vec{p})$ is

$$\Gamma \leq \sum_k \ell_k c_0^k \leq \sum_k n^C (c_0 2^C)^k \leq n^{C+1} \qquad \square$$

Combining Corollaries 4.4 and 3.2, we get a bicriteria approximation algorithm for MinInf-CL.

4.2 A Negative Result When $\beta < 3$

We now consider the case $\beta < 3$ and show an interesting contrast to Lemma 4.3.

Lemma 4.5. *When $\beta = 2 + c_0$ for some constant $c_0 < 1$, there may exist k with $\ell_k = \omega(poly(n, 2^k))$.*

Proof. In the proof of Lemma 4.1 we gave an upper bound for ℓ_k. However, the double counting or the unnecessary cases we involved in our counting can only account for low-order terms. In other words, we can assume that ℓ_k is:

$$\Theta\left(n\left(\frac{2^k k!}{m^k}\right) \sum_{\mathbf{a} \in S(w_{max}, k)} \prod_{i=w_{min}}^{w_{max}} \left(\binom{n_i}{a(i)} i^{2a(i)}\right)\right) \tag{10}$$

Consider the case where $w_{min} = 1$ and $w_{max} = k$, and just take the one sequence $\mathbf{a} = (1, 1, \ldots, 1)$. Furthermore, because $n_i = \Theta(\frac{n}{i^\beta})$ for every $i \in [w_{min}, w_{max}]$, let λ be a universal constant such that $n_i \geq \frac{\lambda n}{i^\beta}$ for every i. Since $a(i) = 1$ for all i here, the quantity inside the Θ notation in (10), which we denote by Q, can be lower-bounded as follows

$$\begin{aligned}
Q &\geq n \cdot \left(\frac{2^k k!}{m^k}\right) \cdot \prod_{i=w_{min}}^{w_{max}} \left(n_i \cdot i^2\right) \\
&\geq n \cdot \left(\frac{2^k k!}{m^k}\right) \cdot \prod_{i=1}^{k} \left((\lambda n / i^\beta) \cdot i^2\right) \\
&= n \cdot \left(\frac{2^k \lambda^k n^k k!}{m^k}\right) \cdot \prod_{i=1}^{k} i^{-c_0} \\
&= n \cdot \left(\frac{2^k \lambda^k n^k}{m^k}\right) \cdot (k!)^{1-c_0} \\
&= \text{poly}(n, 2^k) \cdot (k!)^{1-c_0}
\end{aligned}$$

Because $c_0 < 1$, we have that $(k!)^{1-c_0}$ grows faster than $\text{poly}(2^k)$. Hence, $Q = \omega(\text{poly}(n, 2^k))$ and consequently $\ell_k = \omega(\text{poly}(n, 2^k))$. $\square$

Using reasoning similar to that used in Corollary 4.4, we see that Lemma 4.5 implies that under this stochastic regime, our proof approach cannot provide meaningful results for MinInf-CL. Thus we see a *phase transition* for the expected number of paths of any length k: from at least $(k!)^{\Omega(1)}$ to $\text{poly}(n, 2^k)$ at $\beta = 3$. It is an open question as to what happens when $\beta = 3$.

5 A DETERMINISTIC ROUNDING SAA APPROACH

In this section we revisit the SAA approach of Section 3, and instead of a randomized rounding, we apply a simple deterministic rounding scheme. The advantage of the latter is that the success probability of the algorithm no longer relies on the value Γ. However, this comes at the expense of much worse bicriteria factors.

Once again we are going to sample $N = \frac{3n}{\epsilon^2} \log\left(n^2 \cdot 2^{m+1}\right)$ graphs $G_j = (V, E_j)$ from $G(\vec{p})$, and then construct LP (1)-(4). Let (x, y) be the optimal fractional solution of the LP. In this case, our returned solution will be $F_0 = \{e \in E : x_e \geq \frac{1}{4n^{2/3}}\}$.

Theorem 5.1. *With high probability, the set F_0 is an $(O(n^{2/3}), O(n^{2/3}))$-approximation for* MinInfEdge.

6 CONCLUSIONS AND FUTURE WORK

Despite the fundamental nature of MinInfEdge and MinInfNode, their computational complexity remained open for the $p < 1$ setting. A number of heuristics have been proposed, and rigorous algorithms are only known for very special random graphs. We present the first rigorous approximation results for these problems for certain classes of instances; however, even these turn out to be quite challenging, and require adapting the cut sparsification and sample-average approximation techniques in a nontrivial manner.

Our work raises several interesting questions. First, it would be interesting to extend the result based on Karger's cut sparsification technique to the non-uniform probability setting. Second, it would be interesting to extend our work to other realistic random models of social-contact networks, and to also identify what reasonable assumptions on *deterministic* network models would guarantee efficient solutions. Third, to capture a wider variety of realistic scenarios, it would be beneficial improving the constant upper bound on p in Lemma 4.4. Finally, it is of interest to see if our approximation guarantees and running times can be improved.

Acknowledgements

Michael Dinitz was supported by NSF award CCF-1909111. Aravind Srinivasan was supported in part by NSF awards CCF-1422569, CCF-1749864, and CCF-1918749, as well as research awards from Adobe, Amazon, and Google. Leonidas Tsepenekas was supported in part by NSF awards CCF-1749864 and CCF-1918749, and by research awards from Amazon and Google. Anil Vullikanti's work was partially supported by NSF awards IIS-1931628, CCF-1918656, and IIS-1955797, NIH award R01GM109718, and CDC U01CK000589. The contents are those of the author and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government.

References

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Bollobás, B. and Riordan, O. (2004). Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509.

Chung, F. R. K. and Lu, L. (2006). The volume of the giant component of a random graph with given expected degrees. *SIAM J. Discret. Math.*, 20(2):395–411.

Cohen, R., Havlin, S., and Ben-Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.*, 91:247901.

Eames, K. T., Read, J. M., and Edmunds, W. J. (2009). Epidemic prediction and control in weighted networks. *Epidemics*, 1(1):70 – 76.

Eubank, S., V. S. Anil Kumar, Marathe, M. V., Srinivasan, A., and Wang, N. (2006). Structure of Social Contact Networks and Their Impact on Epidemics. In *Discrete Methods in Epidemiology*, volume 70, pages 179–200. American Math. Soc., Providence, RI.

Eubank, S. G., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., and Wang, N. (2004). Structural and algorithmic aspects of massive social networks. In Munro, J. I., editor, *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004*, pages 718–727. SIAM.

Ganesh, A., Massoulie, L., and Towsley, D. (2005). The effect of network topology on the spread of epidemics. *Proceedings of INFOCOM*.

Grötschel, M., Lovász, L., and Schrijver, A. (1988). *Geometric Algorithms and Combinatorial Optimization*. Springer, Berlin, Heidelberg.

Hayrapetyan, A., Kempe, D., Pál, M., and Svitkina, Z. (2005a). Unbalanced graph cuts. In *Proceedings of the 13th Annual European Conference on Algorithms*, ESA'05, page 191–202, Berlin, Heidelberg. Springer-Verlag.

Hayrapetyan, A., Kempe, D., Pál, M., and Svitkina, Z. (2005b). Unbalanced graph cuts. In *ESA*, pages 191–202.

Karger, D. R. (1999). Random sampling in cut, flow, and network design problems. *Mathematics of Operations Research*, pages 383–413.

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502.

Marathe, M. and Vullikanti, A. (2013). Computational epidemiology. *Communications of the ACM*, 56(7):88–96.

Miller, J. C. and Hyman, J. M. (2007). Effective vaccination strategies for realistic social networks. pages 780–785.

Ogura, M. and Preciado, V. M. (2017). *Optimal Containment of Epidemics in Temporal and Adaptive Networks*, pages 241–266. Springer Singapore, Singapore.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925.

Preciado, V. M., Zargham, M., Enyioha, C., Jadbabaie, A., and Pappas, G. J. (2013). Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. In *IEEE Conference on Decision and Control*. IEEE.

Preciado, V. M., Zargham, M., Enyioha, C., Jadbabaie, A., and Pappas, G. J. (2014a). Optimal resource allocation for network protection against spreading processes. In *IEEE Transactions on Control of Network Systems*, pages 99 – 108. IEEE.

Preciado, V. M., Zargham, M., and Sun, D. (2014b). A convex framework to control spreading processes in directed networks. In *Annual Conference on Information Sciences and Systems (CISS)*. IEEE.

Ruszczynski, A. P. and Shapiro, A. (2003). *Stochastic programming*, volume 10. Elsevier, Amsterdam.

Saha, S., Adiga, A., Prakash, B. A., and Vullikanti, A. (2015). Approximation algorithms for reducing the spectral radius to control epidemic spread. In *SIAM SDM*.

Sambaturu, P., Adhikari, B., Prakash, B. A., Venkatramanan, S., and Vullikanti, A. (2020). Designing effective and practical interventions to contain epidemics. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1187–1195.

Shapiro, A. (2003). Monte Carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425.

Svitkina, Z. and Tardos, É. (2004). Min-max multiway cut. In Jansen, K., Khanna, S., Rolim, J. D. P., and Ron, D., editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 207–218, Berlin, Heidelberg. Springer Berlin Heidelberg.

Swamy, C. and Shmoys, D. B. (2012). Sampling-based approximation algorithms for multistage stochastic optimization. *SIAM Journal on Computing*, 41(4):975–1004.

Yang, Y., McKhann, A., Chen, S., Harling, G., and Onnela, J.-P. (2019). Efficient vaccination strategies for epidemic control using network information. *Epidemics*, 27:115 – 122.

Supplementary Material: Controlling Epidemic Spread using Probabilistic Diffusion Models on Networks

A MISSING PROOFS

Proof of Lemma 3.1. For a fixed set F, the quantities $h(G_j, F)$ are independent. Further, let $X_j = \frac{h(G_j,F)}{n} \in [0,1]$ and $X = \sum_{j=1}^{N} X_j$. Note that $X = \frac{N}{n} h(G, F)$. Using the Chernoff bound of Lemma C.1 yields:

$$\Pr\Big[X \notin \big[(1-\epsilon)\mathbb{E}[X], (1+\epsilon)\mathbb{E}[X]\big]\Big] \leq 2e^{-\frac{\epsilon^2 \mathbb{E}[X]}{3}} = 2e^{-\frac{\epsilon^2 N \cdot \mathbb{E}[h(G,F)]}{3n}} \leq \frac{1}{n^2 2^m}$$

To get the last inequality we use the definition of N, and the fact that $\mathbb{E}[h(G, F)] \geq 1$ (since there is always at least one infection, namely the node s). Finally, since $X = \frac{N}{n} h(G, F)$, we also have:

$$\Pr\Big[h(G,F) \notin \big[(1-\epsilon)\mathbb{E}[h(G,F)], (1+\epsilon)\mathbb{E}[h(G,F)]\big]\Big] \leq \frac{1}{n^2 2^m}$$

Because the number of possible subsets F is 2^m, a union bound over them concludes the proof. □

Proof of Corollary 3.2. When $\Gamma \leq \text{poly}(n)$, we set γ large enough such that $O(\frac{\Gamma}{\epsilon^2 n^\gamma}) = O(\frac{1}{n^2})$. Using Theorem 1.1 we then have:

$$\begin{aligned}\mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s)] &= \mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s) | \mathcal{A}] \Pr[\mathcal{A}] + \mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s) | \bar{\mathcal{A}}] \Pr[\bar{\mathcal{A}}] \\ &\leq (1 + O(\epsilon))OPT + nO(1/n^2) \leq (1 + O(\epsilon) + O(1/n))OPT\end{aligned}$$

To get the first inequality we use the simply upper bound of $\mathbb{E}[inf(V, E(\vec{p}) \setminus F_0, s) | \bar{\mathcal{A}}] \leq n$, and for the last one we use the fact that $1 \leq OPT$ (s is always infected). □

Proof of Lemma 4.2. We prove by induction on D the stronger statement (A), which directly implies the Lemma.

$$\forall k, \; N(D,k) \leq \frac{1}{k!} \cdot \prod_{i=w_{min}+1}^{D} \left(1 + \frac{1}{i^{c_1}}\right)^k \tag{A}$$

The base case $D = w_{min}$ is easy. For this notice that $S(w_{min}, k) = \{(k)\}$, and hence we have

$$N(w_{min}, k) = \frac{1}{w_{min}^{c_1 k} \cdot k!} \leq \frac{1}{k!}$$

The inequality above follows since $c_1 > 1$.

We complete the proof by strong induction. Suppose $D > w_{min}$. Elementary calculations and the definition of $N(\cdot, \cdot)$ reveal the following recurrence when $D > w_{min}$

$$N(D,k) = \sum_{j=0}^{k} \left(N(D-1, k-j) \cdot \frac{1}{D^{c_1 j} \cdot j!}\right) \tag{11}$$

Recurrence (11) and the induction hypothesis yield

$$\begin{aligned}
N(D,k) &\leq \sum_{j=0}^{k}\left(\frac{1}{D^{c_1 j}\cdot j!}\cdot\frac{1}{(k-j)!}\prod_{i=w_{min}+1}^{D-1}\left(1+\frac{1}{i^{c_1}}\right)^{k-j}\right)\\
&\leq \sum_{j=0}^{k}\left(\frac{1}{D^{c_1 j}\cdot j!}\cdot\frac{1}{(k-j)!}\prod_{i=w_{min}+1}^{D-1}\left(1+\frac{1}{i^{c_1}}\right)^{k}\right)\\
&= \left(\prod_{i=w_{min}+1}^{D-1}\left(1+\frac{1}{i^{c_1}}\right)^{k}\right)\sum_{j=0}^{k}\left(\frac{1}{D^{c_1 j}\cdot j!}\cdot\frac{1}{(k-j)!}\right)\\
&= \frac{1}{k!}\cdot\left(\prod_{i=w_{min}+1}^{D-1}\left(1+\frac{1}{i^{c_1}}\right)^{k}\right)\cdot\sum_{j=0}^{k}\left(\binom{k}{j}\cdot\frac{1}{D^{c_1 j}}\right)\\
&= \frac{1}{k!}\cdot\left(\prod_{i=w_{min}+1}^{D}\left(1+\frac{1}{i^{c_1}}\right)^{k}\right)
\end{aligned}$$

The last inequality above follows from the binomial sum $\sum_{j=0}^{k}\left(\binom{k}{j}\cdot\frac{1}{D^{c_1 j}}\right)=(1+1/D^{c_1})^k$. □

Proof of Theorem 5.1. Before we proceed with our analysis, let us recall some important notation from Section 3. For any fixed $F\subseteq E$, $h(G_j,F)=inf(V,E_j\setminus F,s)$ and $h(G,F)=\frac{1}{N}\sum_{j=1}^{N}h(G_j,F)$. Finally, F^* denotes the optimal edge set for the given instance of MinInfEdge, and $\hat{F}$ denotes the optimal integral solution of LP (1)-(4).

To begin with, by the definition of F_0 and constraint (3), we have

$$\sum_{e\in F_0}c_e\leq 4n^{2/3}\sum_{e\in F_0}c_e x_e\leq 4n^{2/3}B$$

Moving forward, note that by Lemma 3.1, we have $h(G,F^*)\leq(1+\epsilon)\mathbb{E}[h(G,F^*)]$ and $h(G,F_0)\geq(1-\epsilon)\mathbb{E}[h(G,F_0)]$ with probability at least $1-O(1/n^2)$. If we show that $h(G,F_0)\leq 2n^{2/3}h(G,\hat{F})$, then we are done. This is because:

$$\mathbb{E}[h(G,F_0)]\leq\frac{h(G,F_0)}{1-\epsilon}\leq\frac{2n^{2/3}}{1-\epsilon}h(G,\hat{F})\leq\frac{2n^{2/3}}{1-\epsilon}h(G,F^*)\leq\frac{2(1+\epsilon)n^{2/3}}{1-\epsilon}\mathbb{E}[h(G,F^*)]$$

At first, suppose $h(G,\hat{F})>n^{1/3}$. Since $h(G_j,F_0)\leq n\leq n^{2/3}h(G,\hat{F})$ for any j, $h(G,F_0)\leq 2n^{2/3}h(G,\hat{F})$ follows trivially through the definition of $h(G,F_0)$.

Next, suppose $h(G,\hat{F})\leq n^{1/3}$. This implies $\frac{1}{N}\sum_{j\in[N]}\sum_{v\in V}(1-y_{vj})\leq h(G,\hat{F})\leq n^{1/3}$, because the optimal LP-value is a lower bound for $h(G,\hat{F})$. Let now $A'=\{j\in[N]:\sum_{v\in V}(1-y_{jv})\leq n^{2/3}\}$ and $A''=[N]\setminus A'=\{j\in[N]:\sum_{v\in V}(1-y_{vj})>n^{2/3}\}$. The upper bound of the optimal fractional solution value then gives $|A''|\leq N/n^{1/3}$. Consider now any $j\in A'$, and let v be a node such that $1-y_{vj}\leq 1/2$. We will argue below that for any path $P\in\mathcal{P}(s,v,G_j)$, there exists an edge $e\in P$ such that $e\in F_0$. This means that if v is infected in $(V,E_j\setminus F_0)$, then $1-y_{vj}>1/2$, and so $h(G_j,F_0)\leq\sum_v 2(1-y_{vj})$. Hence,

$$\begin{aligned}
h(G,F_0)&=\frac{1}{N}\sum_{j\in A'}h(G_j,F_0)+\frac{1}{N}\sum_{j\in A''}h(G_j,F_0)\\
&\leq\frac{1}{N}\sum_{j\in A'}\sum_{v\in V}2(1-y_{vj})+\frac{n|A''|}{N}\\
&\leq\frac{1}{N}\sum_{j\in[N]}\sum_{v\in V}2(1-y_{vj})+n^{2/3}\\
&\leq\frac{1}{N}\sum_{j\in[N]}\sum_{v\in V}2(1-y_{vj})+n^{2/3}h(G,\hat{F})\\
&\leq(2+n^{2/3})h(G,\hat{F})\leq 2n^{2/3}h(G,\hat{F})
\end{aligned}$$

where the third inequality follows because $h(G_j, \hat{F}) \geq 1$, and thus $h(G, \hat{F}) \geq 1$.

Finally, we prove that for any $j \in A'$, and any v such that $1 - y_{vj} \leq 1/2$, it must be the case that for each $P \in \mathcal{P}(s, v, G_j)$ we have $P \cap F_0 \neq \emptyset$. Let $P = (v_0, v_1, \ldots, v_r)$ with $v_0 = s$ be such a path of $\mathcal{P}(s, v, G_j)$. First, suppose $|P| = r \leq 2n^{2/3}$. Then, constraint (2) yields $\sum_{e \in P} x_e \geq 1/2$, and hence there exists $e \in P$ with $x_e \geq 1/(2|P|) \geq 1/(4n^{2/3})$, which implies $e \in F_0$. Next, suppose $|P| > 2n^{2/3}$. Let $P' = (v_0, v_1, \ldots, v_k)$ be the prefix of P of length $k = 2n^{2/3}$. We will show that $P' \cap F_0 \neq \emptyset$, which implies $P \cap F_0 \neq \emptyset$. By definition of A', we have $\sum_{i=0}^{k}(1 - y_{v_i j}) \leq n^{2/3}$. Since $k = 2n^{2/3}$, there exists some $1 \leq \ell \leq k$ such that $1 - y_{v_\ell j} \leq 1/2$, or $y_{v_\ell j} \geq 1/2$. Constraint (2) applied for v_ℓ and the path $(v_0, v_1, \ldots, v_\ell)$ gives $\sum_{i=0}^{\ell-1} x_{(v_i, v_{i+1})} \geq y_{v_\ell j} \geq 1/2$, and thus there exists an edge $(v_i, v_{i+1}) \in P'$ with $x_{(v_i, v_{i+1})} \geq \frac{1}{2\ell} \geq \frac{1}{2k} \geq \frac{1}{4n^{2/3}}$, which means that $(v_i, v_{i+1}) \in F_0$. □

B THE MinInfNode PROBLEM

While the results based on Karger's technique (Section 2) do not easily extend to the MinInfNode problem, our SAA based results do, and we explain this here.

We make a few small changes to the linear program LP (1)-(4). At first, we use variable x_v as the indicator for removing (i.e., vaccinating) vertex v. Furthermore, each $P \in \mathcal{P}(s, v, G_j)$ will now contain the vertices of the path and not the edges. Everything else remains the same, and thus we get the following linear program, denoted by LP_{vacc}:

$$\min \frac{1}{N} \sum_j \sum_v (1 - y_{vj}) \text{ such that} \tag{12}$$

$$\sum_{v \in P} x_v \geq y_{vj}, \quad \forall j \ \forall P \in \mathcal{P}(s, v, G_j) \tag{13}$$

$$\sum_v c_v x_v \leq B \tag{14}$$

$$x_v, y_{vj} \in [0, 1], \quad \text{for all } j \in [N], v \in V \tag{15}$$

MinInfNode in the Chung-Lu model. Our rounding scheme now involves constructing a subset $F_0 \subseteq V$, by picking each $v \in V$ with probability

$$x'_v = \min\left\{\frac{(\gamma + 5)x_v \log n}{\epsilon}, 1\right\}$$

It is easy to verify that Theorem 1.1 and Corollary 3.2 hold in the case of vertex removal, by considering the quantity $inf(V, E \setminus \{(u, v) \in E : u \in F \text{ or } v \in F\}, s)$ instead of $inf(V, E \setminus F, s)$. Corollary 4.4 is unchanged for the vertex removal case, as well. Putting these together, we have the following result for the MinInfNode problem.

Corollary B.1. *The solution F_0 picked by the rounding scheme above is an $(O(\log n), O(1))$-approximation for the* MinInfNode *problem for graphs drawn from the Chung-Lu model with power law weights, with parameter $\beta = 2 + c_1$ for some constant $c_1 > 1$.*

MinInfNode in general graphs. The deterministic rounding of Section 5 holds if the rounding for edges is replaced by the same rounding for nodes, which gives us the following result.

Corollary B.2. *There is an $(O(n^{2/3}), O(n^{2/3}))$–approximation for the* MinInfNode *problem.*

C AUXILIARY LEMMAS

Lemma C.1. *(Chernoff, 1952) Let $X_1, X_2, \ldots, X_K$ be independent random variables with $X_k \in [0, 1]$ for every k. For $X = \sum_{k=1}^{K} X_k$ with $\mu = \mathbb{E}[X]$:*

- *For any $\delta > 0$, we have* $\Pr\left[X \notin [(1 - \delta)\mu, (1 + \delta)\mu]\right] \leq e^{\frac{-\mu\delta^2}{3}}$.
- *For any $R \geq 6\mu$, we have* $\Pr[X \geq R] \leq 2^{-R}$.