Critical spatial clusters for vaccine preventable diseases

Jose Cadena
Biocomplexity Institute
Blacksburg, VA
jcadena@vt.edu

Achla Marathe
Biocomplexity Institute
Blacksburg, VA
amarathe@vt.edu

Anil Vullikanti
Biocomplexity Institute
Blacksburg, VA
vsakumar@vt.edu

ABSTRACT

Despite high vaccination rates for infectious diseases, such as measles, there have been several big disease outbreaks in recent years. This is, in part, due to misinformation about vaccinations in certain sub-populations, and their spatial clustering. Identifying potential clusters, which can result in big outbreaks in the event of reduced vaccination rate, is an important public health challenge. We develop a natural notion of criticality of such clusters, which extends the problems of influence maximization to connectivity constraints. We develop efficient approximation algorithms for finding critical clusters by exploiting the structural properties of the problem in contact networks. We apply our methods to find critical clusters in the state of Minnesota, with significantly higher criticality than those obtained by heuristics used in public health.

KEYWORDS

Criticality, submodularity optimization, epidemic spread

ACM Reference Format:
Jose Cadena, Achla Marathe, and Anil Vullikanti. 2018. Critical spatial clusters for vaccine preventable diseases. In *Proceedings of ACM KDD Conference (KDD'18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Many childhood diseases, such as measles and Pertussis, are easily preventable by vaccination. Therefore, it is worrisome that fairly large outbreaks of such diseases have occurred in recent years, such as the measles outbreaks in California in 2015 and in Minnesota in 2017—this is despite high vaccination coverage in the US, e.g., ∼ 95% for MMR, the measles vaccine. One of the reasons is the emergence of undervaccinated geographical clusters [17], often driven by misperceptions about side effects of vaccines [4]. The typical response by public health agencies is to monitor these clusters, run active information campaigns, and engage community leaders. However, such interventions are very expensive and time consuming. Another issue is that public health departments might not be aware of all such clusters, especially in the early stages. As a policy design question, public health agencies are interested in discovering which regions are "critical" spatial clusters, where a reduction in vaccination rate could cause a big outbreak. Current practices involve broad outreach efforts to communities considered at risk, which might not be efficient if some communities are not so critical. Formalizing the notion of critical clusters can help public health agencies focus their limited resources in areas where impact can be maximized. Our contributions are summarized below.

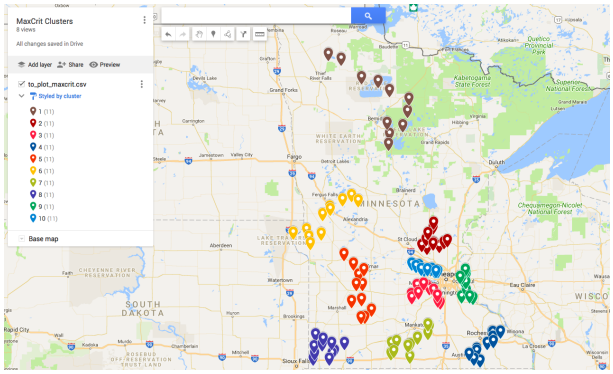

Figure 1: Critical sets in Minnesota discovered using our methods. These are contiguous regions that lead to large outbreaks of measles if not properly vaccinated.

1. Formalizing critical clusters. We define the *criticality* of a subpopulation S as the expected number of *additional* infections that would occur if the individuals in S are not properly immunized. Our focus is on subpopulations located within a bounded spatial cluster. We have different criticality objectives, MaxCrit and ECrit, which capture two distinct public health policy questions: is the source of the infection within the cluster or outside? (Section 4).

Table 1: A summary of our proposed methods

	MaxCrit problem	**ECrit** problem
Motivating policy question	Maximum criticality for source in cluster	Maximum criticality for a fixed source
Characteristics of optimal solutions (Section 3.3.1)	Less connected	Centrally located
Structural property (Section 3.3.2)	Submodular, but not locally modular	Submodular and locally modular
Approximation guarantee	$\Omega(1/k^{1/3})$-approximation (Theorem 3)	$\Omega(1/\log k)$-approximation (Theorem 2)
Empirical observations for MN	Much higher criticality, Most critical cluster is in a rural region	Lower criticality than for MaxCrit, well connected

2. Efficient algorithms for the MaxCrit and ECrit problems. We show that MaxCrit and ECrit are both NP-hard; then, we focus on efficient approximation algorithms for these two problems. Our algorithms exploit structural properties of the objective function

KDD'18, August 19-23, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

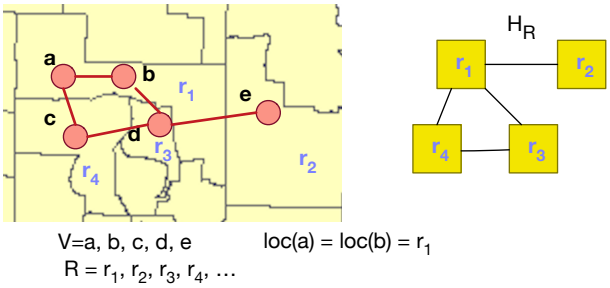

Figure 2: Definitions and notation used in our paper. The 5 red circle nodes (a, b, c, d, e) form a social contact network. Each node resides in a block group r_i, and these block groups from the auxiliary graph $H_{\mathcal{R}}$, where an edge represents that the block groups are neighbors on the map.

and the small world structure of contact networks. MaxCrit and ECrit are instances of submodular function maximization with connectivity constraints, a very challenging problem in combinatorial optimization. However, we show that ECrit has an *approximately modular* structure, which we use to derive a good approximation bound. For the MaxCrit objective, we use the spatial structure to obtain a good approximation factor. Table 1 summarizes these bounds.

3. Application. We evaluate our algorithms on a network model for the state of Minnesota. The sets we discover have very high criticality compared to heuristics commonly considered in the public health community. The critical clusters computed using our algorithms (shown in Figure 1) have meaningful demographic properties from a public health perspective: they typically involve people with lower than average income levels and age (Section 5).

4. Connections with the Influence Maximization problem. We show that criticality is related to the classical problem of influence maximization, but with one very significant difference: *the set of influencers has to form a connected spatial cluster.* As a result, the standard greedy algorithm cannot be used for finding critical clusters. We are not aware of any efficient algorithms with good approximation bounds. The closest is an $\Omega(1/\sqrt{k})$-approximation, which can be obtained by using the algorithm of [15] for submodular function maximization with a connectivity constraint. However, we are able to obtain significantly improved bounds by exploiting the structure of our problems, as summarized in Table 1.

2 PRELIMINARIES

Let V denote a population, and let $G = (V, E)$ be a contact graph on which a disease can spread. That is, a person $v \in V$ (referred to as a "node", henceforth) can propagate the disease to its neighbors. In the social contact network datasets that we consider (Section 5.1.1), each person v is associated with a geographical location, denoted by $\text{loc}(v)$; we will consider such locations at the resolution of census block groups. Let $\mathcal{R}$ denote the geographical region where the nodes V are located—for example, the state of Minnesota—and let $\mathcal{R} = \{r_1, \ldots, r_N\}$ be a decomposition of $\mathcal{R}$ into census block groups. For a block group $r_i \in \mathcal{R}$, let $V(r_i)$ denote the set of nodes v with $\text{loc}(v) \in r_i$. For a subset of block groups $R \subset \mathcal{R}$, let $V(R) = \cup_{r_i \in R} V(r_i)$ be the set of nodes located within it. We consider a graph $H_{\mathcal{R}} = (\mathcal{R}, E_{\mathcal{R}})$ on the set $\mathcal{R}$ of block groups, where two block groups are connected if they are geographically contiguous, i.e., they are adjacent on a map. These definitions are illustrated in Figure 2. Let $\text{Conn}(\mathcal{R})$ denote subsets $R \subset \mathcal{R}$ that are spatially connected in the block group graph $H_{\mathcal{R}}$.

Table 2: Summary of the notation used in the paper

Notation	Description
$G = (V, E)$	Contact graph on set V of individuals
$\text{loc}(v)$	Geographical location of node v
$\mathcal{R} = \{r_1, \ldots, r_N\}$	Geographical region where the nodes V are located—e.g., Minnesota—partitioned into block groups r_i
$V(r_i)$	Set of nodes of G with $\text{loc}(v) = r_i$
$V(R)$	$\cup_{r_i \in R} V(r_i)$
$H_{\mathcal{R}} = (\mathcal{R}, E_{\mathcal{R}})$	Network on $\mathcal{R}$, with adjacent block groups connected by an edge. Sometimes referred to as the "Auxiliary network"
$\text{Conn}(\mathcal{R})$	Set of $R \subset \mathcal{R}$ which are spatially connected in $H_{\mathcal{R}}$
S, E, I, R	States in the disease model
γ	Average region-wide vaccination rate
$\mathbf{x}$	Vaccination vector, with x_i denoting the probability that node i is vaccinated
$\mathbf{x}^R$	Vaccination vector, with nodes $V(R)$ undervaccinated, where $R \in \text{Conn}(\mathcal{R})$
Src_A, Src	Denotes the event that the source of the infection is from a set $A \subset \mathcal{R}$. Src is used when $A = \mathcal{R}$
$\#\text{inf}(\mathbf{x}, \text{Src}_A)$	Expected number of infections for vaccination vector $\mathbf{x}$ and source being Src_A
$\text{crit}(R, \mathbf{x}, \text{Src}_A)$	Criticality of $R \in \text{Conn}(\mathcal{R})$: expected number of additional infections that occur if R is not vaccinated
$\text{MaxCrit}(R)$, $\text{MaxCrit}(R, \mathbf{x})$	The objective value of MaxCrit for region $R \in \text{Conn}(\mathcal{R})$ in an instance $(G, H_{\mathcal{R}}, k)$
$\text{ECrit}(R)$, $\text{ECrit}(R, \mathbf{x})$	The objective value of ECrit for region $R \in \text{Conn}(\mathcal{R})$ in an instance $(G, H_{\mathcal{R}}, k)$

We will use an SEIR model for diseases like measles [3], where a node is in one of *four* states: Susceptible (S), Exposed (E), Infected (I) and Recovered/Removed (R). Measles is highly contagious, and an infected node v spreads the disease to each susceptible unvaccinated neighbor $u \in N(v)$ with high probability. Sometimes, we assume a transmission probability of 1, but all our results extend to the more general case. If v is vaccinated, it does not get infected. We assume the vaccine has 100% efficacy, which is not true in practice, but this is not crucial for our methodology.

Let γ denote the average region-wide vaccination rate—~ 0.97 in Minnesota. Let $\mathbf{x}$ be a *vaccination vector*: $x_i \in [0, 1]$ denotes the probability that node i is vaccinated (so $x_i = \gamma$, by default). Let Src_A denote the source of the infection: this could be one or a small number of nodes from a region $A \subset \mathcal{R}$, which initially get infected (e.g., by contact outside $\mathcal{R}$). We will drop the subscript if $A = \mathcal{R}$. Let $\#\text{inf}(\mathbf{x}, \text{Src}_A)$ denote the expected number of infections for the intervention $\mathbf{x}$, when the initial infection is at Src_A. When the initial conditions are clear from the context, we denote this by $\#\text{inf}(\mathbf{x})$.

2.1 Criticality and Problem Formulations

For a vaccination vector $\mathbf{x}$, let $\mathbf{x}^S$ denote the corresponding intervention where a subset $S \subset V$ of nodes is undervaccinated, and the remaining nodes are vaccinated with the same probability as in $\mathbf{x}$; that is $\mathbf{x}^S_i = \mathbf{x}_i$ for $i \notin S$ and $\mathbf{x}^S_i = \gamma'$ for $i \in S$, where γ' is much lower than γ, the region-wide vaccination rate. For simplicity, we sometimes consider $\gamma' = 0$.

We define the **criticality** of a set $S \subset V$ as $\text{crit}(S, \mathbf{x}, \text{Src}_A) = \#\text{inf}(\mathbf{x}^S, \text{Src}_A) - \#\text{inf}(\mathbf{x}, \text{Src}_A)$, which is the *expected number of additional infections that occur if S is not vaccinated* (with respect to

any specific initial conditions Src_A). Our focus is on finding spatial clusters of high criticality. Specifically, we will focus on $S = V(R)$ for a connected region $R \in \text{Conn}(\mathcal{R})$. We denote this by

$$\text{crit}(R, \mathbf{x}, \text{Src}_A) = \#\text{inf}(\mathbf{x}^R, \text{Src}_A) - \#\text{inf}(\mathbf{x}, \text{Src}_A),$$

which is the expected number of extra infections that might be caused if the nodes in the connected region R are under-vaccinated.

We focus on finding "small" connected regions, since this can lead to an actionable policy for public health agencies. We model this by adding a constraint $|R| \leq k$, where k is a parameter that can be tuned based on the available resources of a public health agency.

We propose two problems that model two different kinds of initial conditions of interest from a public health perspective. The first problem models the following question: *for any specific initial condition (e.g., Src denotes kids in an elementary school), what is the most critical set?*

Problem 1 (k-ECrit$(G, H_{\mathcal{R}}, k)$). *Given an instance $(G, H_{\mathcal{R}}, k)$, find a connected region $R \in Conn(\mathcal{R})$ of size at most k that maximizes criticality:*

$$R = argmax_{R' \in Conn(\mathcal{R}), |R'| \leq k} crit(R', \mathbf{x}, Src)$$

For convenience, we will sometimes also use ECrit$(R, \mathbf{x}, \text{Src})$ or ECrit(R) to denote crit$(R, \mathbf{x}, \text{Src})$, the objective value of ECrit for region R in an instance $(G, H_{\mathcal{R}}, k)$. The second problem models the following question: *what is the most critical cluster if the infection source is the worst possible, which will happen if the infection starts within the undervaccinated cluster itself?* This is formalized as

Problem 2 (k-MaxCrit$(G, H_{\mathcal{R}}, k)$). *Given an instance $(G, H_{\mathcal{R}}, k)$, find a connected region $R \in Conn(\mathcal{R})$ of size at most k that maximizes criticality over all choices of source:*

$$R = argmax_{R' \in Conn(\mathcal{R}), |R'| \leq k, Src_{R'}} crit(R', \mathbf{x}, Src'_R)$$

In other words, the k-MaxCrit problem involves maximizing over all possible choices of the sources $\text{Src}_{R'}$ in the cluster R'. As before, we will use MaxCrit$(R, \mathbf{x}, \text{Src})$ or MaxCrit(R) to denote the objective value of an instance of the problem.

3 KEY PROPERTIES OF CRITICALITY

We start by describing connections between the proposed MaxCrit and ECrit objectives, and influence maximization, which will have implications on the computational complexity. We also prove structural properties of these objectives, later used in our algorithms.

3.1 Complexity and connections with Influence Maximization

In the Influence Maximization (InfMax) problem [12], we are given a directed graph $G = (V, E)$ and edge weights $p(u, v) \in [0, 1]$ indicating the probability that node u influences node v. The *Independent Cascade* model is a special case of the SEIR model, where each node is infectious for exactly one time step. The goal is to find a set $S \subset V$ of k *seed nodes* to infect, such that the expected number of influenced nodes or *spread*, $\sigma(S)$, is maximized. There has been a lot of work on the InfMax problem since its introduction by [12]. An instance of InfMax consists of a single contact graph G, whereas instances of the MaxCrit and ECrit problems consist of the contact graph G, a partition of the nodes of G into regions, $\mathcal{R}$, and an auxiliary graph $H_{\mathcal{R}}$ that captures connectivity among $\mathcal{R}$.

3.1.1 NP-hardness. InfMax can be reduced to MaxCrit and ECrit, which implies their NP-hardness. The proof is by constructing a suitable auxiliary graph $H_{\mathcal{R}}$, a vaccination vector $\mathbf{x}$, and a source Src. This is summarized in Theorem 1, whose proof is presented in the Appendix in the full version of this paper [1].

Theorem 1. *MaxCrit and ECrit are NP-hard.*

3.1.2 Impact of connectivity requirement. The connectivity constraint has a strong effect on the solution of MaxCrit and ECrit. In particular, a solution computed for InfMax using the greedy algorithm of [12] can be arbitrarily suboptimal for the problems we propose. Informally, this follows from the property of InfMax that it is better to choose the set of seeds to be located far apart, so that their combined influence is maximized.

Observation 1. *There exists a family of instances $(G, H_{\mathcal{R}}, k)$ for which the optimum solution S^* to MaxCrit satisfies MaxCrit$(S^*) = O(\frac{1}{k}\text{InfMax}(\hat{S}))$, where $\hat{S}$ is the optimum solution to the InfMax version for this instance, without any connectivity requirements.*

3.2 Submodularity of MaxCrit and ECrit

A set function $f : 2^V \rightarrow \mathbb{R}$ is said to be *submodular* if it satisfies the diminishing returns property: for any $T \subset S \subset V$ and $x \in V \setminus S$, we have that $f(T \cup x) - f(T) \geq f(S \cup x) - f(S)$. We have the following result:

Lemma 3.1. *MaxCrit and ECrit are submodular.*

The proof is presented in the full version, but the argument is similar to the submodularity proof for the InfMax problem.

3.3 Differences between MaxCrit and ECrit

While these two problems model related public health problems, they have some significant differences, both in terms of the structure of the optimum solutions and a locality property, which is useful in designing efficient algorithms.

3.3.1 Difference in the structure of optimal solutions. The solution structure for both problems can be very different in the worst case. Consider a contact graph G split into regions $\mathcal{R} = \{r_1, \ldots, r_N\}$. For each r_i, we assume $V(r_i)$ has n nodes. The auxiliary graph $H_{\mathcal{R}}$ consists of two disjoint sets: graph H_1 induced by $r_1, \ldots, r_{N-k'}$, and graph H_2 induced by $r_{N-k'+1}, \ldots, r_N$. For each $i \leq N - k'$, the graph $G[V(r_i)]$ is a connected component. The graph H_1 forms a chain, with $V(r_1)$ having an edge to $V(r_2)$, $V(r_2)$ having edges to $V(r_1)$ and $V(r_3)$, etc. We have Src to be a node $s \in V(r_{n'})$, where $n' = \lfloor (N - k')/2 \rfloor$. The graph $G[H_2]$ restricted to H_2 is fully connected, but it is disconnected from H_1; we also choose it so that $G[H_2]$ has more nodes than H_1. Then, an optimum solution to the ECrit problem with Src will be the cluster of k block groups centered at $r_{n'}$, with criticality of $O(kn + 2n/\gamma)$, by considering a percolation process on a chain. If we choose $k' > k + 2/\gamma$, the optimal solution to the MaxCrit problem on this instance will be a cluster of k block groups from H_2, since the fully connected structure in H_2 will lead to a larger outbreak.

3.3.2 Local modularity property. ECrit has a local modularity structure, which is motivated by [14]. Specifically, for a set $A_1 \cup A_2 \cup \ldots A_r$ of disjoint and roughly similar sized clusters, ECrit$(A_1 \cup A_2 \cup$

$\ldots A_r) \geq c \cdot \sum_i \text{ECrit}(A_i)$, for a constant $c < 1$. This is different in form from the notion of (r, δ)-local function of [14]. A function $F(\cdot)$ is (r, δ)-local if $F(A_1 \cup A_2) \geq F(A_1) + \delta F(A_2)$ for two sets A_1 and A_2, which are distance r away. It is not clear that ECrit satisfies such property, but the specific kind of property it satisfies is sufficient for using the subsequent technique of [14]. In contrast, we show that the MaxCrit is not locally modular.

We assume our contact graph is a "small world" network, following the model of [13] in which nodes have local connections to nearby nodes and a small number of long range connections. A node u has a long range connection to node v with probability proportional to $\frac{1}{d_{uv}^{\alpha}}$, where α is the "power law exponent", typically $\alpha > 2$. We will consider a set of clusters $A_1, \ldots, A_r$, where A_i has size n_i. We assume all clusters have roughly similar size, so $n_i \leq n_j \beta$ for a constant β. We assume the clusters are small, specifically $n_i \leq \sqrt{n}$. For the analysis below, we assume the disease is highly contagious and there is enough local connectivity within each cluster. Therefore, if nodes in A_i are not vaccinated, and some node $v \in A_i$ gets infected (from outside the cluster), the entire cluster will get infected. We also assume the clusters A_i are fairly localized, so that we can consider d_{ij} to be the distance from the centroid of A_i to that of A_j. For simplicity of the analysis, we will assume that in the small world network model for H, the probability that a node u in A_i connects to a node v in A_j is proportional to $\frac{1}{d_{ij}^{\alpha}}$. The following Lemma—proven in the full version—shows the local modularity of ECrit.

Lemma 3.2. *Let $A_1, \ldots, A_r$ be disjoint clusters, with the model and notation as described above. Then,*

$$\textit{ECrit}(A_1 \cup A_1 \cup \ldots A_r) \geq \frac{1}{1 + 3\gamma\beta/(\alpha - 1)} \Big(\textit{ECrit}(A_1) + \ldots + \textit{ECrit}(A_r) \Big)$$

In contrast, MaxCrit does not satisfy the property from Lemma 3.2. Consider a setting where each block group induces a clique, which is disjoint from all other block groups. Then, $\text{MaxCrit}(A_1) \propto \max_{b \in A_1} |V(b)|$ is proportional to the largest block group in the set A_1. Similarly, we have $\text{MaxCrit}(A_2) \propto \max_{b \in A_2} |V(b)|$. For disjoint sets A_1, A_2, we have

$$\text{MaxCrit}(A_1 \cup A_2) = \max_{b \in A_1 \cup A_2} |V(b)| < \text{MaxCrit}(A_1) + \text{MaxCrit}(A_2)$$

4 PROPOSED METHODS

4.1 Algorithm ApproxECrit

Our algorithm ApproxECrit uses the locality property from Lemma 3.2 and builds on the approach of Krause et al. [14] and Borgs et al. [5]. Algorithm 1 gives a pseudocode description, and we give the intuitive ideas below.

(1) **Padded decompositions.** This is a partition of the graph $H_{\mathcal{R}}$ into clusters $C_1, \ldots C_\ell$, each of diameter at most $12r$. If a node v and all nodes at distance at most r of v are in the same cluster, we say that v is r-padded. After clustering, all the nodes that are not r-padded are removed; this occurs with probability 1/2 for each node, and the best solution S of size k after removal has objective value $F(S) \geq \frac{1}{2} F(S^*)$, where S^* is the optimal subgraph of size k.

(2) **Greedy solution in the clusters.** The purpose of the padded decomposition was to partition the graph into small clusters where we can ignore the connectivity cost [14]. For each cluster, we now run the greedy algorithm for submodularity maximization to obtain an ordering of the nodes; the first j nodes in this ordering are approximately the most informative nodes in the cluster. We implement Algorithm 2, a modified version of the algorithm in [5] to account for the fact that we want a graph that is connected in the auxiliary graph, but with the epidemic process occurring in the social contact network. The greedy algorithm degrades the quality of the optimal solution by a factor of at most $(1 - 1/\epsilon)$.

(3) **Running Quota Steiner Tree on $\mathcal{R}$.** Finally, we compute wt_r for each $r \in \mathcal{R}$, and then compute a quota Steiner tree T' of size k, which maximizes $\sum_{r \in T'} \text{wt}_r$. The subroutine k-MaxST uses the fixed parameter algorithm of [6] to find an optimal solution, as described in Section 4.2.1.

Algorithm 1 ApproxECrit(G, $H_{\mathcal{R}}$, k, Src).

1: Partition $H_{\mathcal{R}}$ into clusters $C_1, \ldots C_\ell$, each of diameter at most $12r$, using the method of [14] (referred to as a *Padded decomposition*)
2: For each cluster $C_i = \{r_{i1}, \ldots, r_{ij}\}$, let $(r_{ia_1}, r_{ia_2}, \ldots, r_{ia_j}) =$ Greedy(C_i, j, Src), be an ordering of block groups obtained by running Greedy without connectivity constraints
3: Construct a connected graph G' on the nodes $\{r_{ia_1} : i = 1, \ldots, \ell\}$ with an edge (r_{ia_1}, r_{ja_1}) having weight equal to the shortest path length in $H_{\mathcal{R}}$. Run the Budgeted Steiner Tree algorithm of [11] to find a tree T with k nodes and maximum total criticality
4: **for** $r \in H_{\mathcal{R}}$ **do**
5: Let $\text{wt}_r = \text{crit}(r)$
6: **end for**
7: Let $T' = k - \text{MaxST}(H_{\mathcal{R}}, \text{wt}, k)$ using the algorithm of [6]
8: return $\max\{\text{ECrit}(T), \text{ECrit}(T')\}$

Algorithm 2 Greedy(C, j, Src).

1: $\mathcal{S} = \phi$, $L = cj|E| \log |V|$, for a constant c
2: $\ell = 0$
3: **while** $\ell < L$ **do**
4: Pick random subgraph G' of G with (1) edges sampled based on disease transmission probability, (2) node $v \in V(C)$ sampled with probability $1 - \gamma'$, (3) $v \notin V(C)$ sampled with probability $1 - \gamma$
5: $\ell = \ell + |E(G')|$
6: Let C_i be the set of components reachable from Src in G'
7: $\mathcal{S} = \mathcal{S} \cup \{C_i\}$
8: **end while**
9: Initialize $X = \phi$
10: For each $r \in C$, define $deg(r, \mathcal{S})$ to be the number of sets $C_i \in \mathcal{S}$ that contain some node in $V(r)$
11: **for** $i = 1$ to j **do**
12: Append $r = \text{argmax}_{r'} deg(r', \mathcal{S})$ to X
13: Remove all sets C_i hit by $V(r)$ from $\mathcal{S}$ and update all $deg(\cdot)$
14: **end for**

Theorem 2. *Let S^* denote an optimal solution to an instance of the ECrit$(G, H_{\mathcal{R}}, k, \mathbf{x}, Src_A)$ problem. Let S be the cluster returned by ApproxECrit. If $H_{\mathcal{R}}$ forms a small world network, and the sizes of all block groups in $\mathcal{R}$ are within a constant factor of each other, then S has $O(k)$ nodes and ECrit$(S, \mathbf{x}, Src_A) \geq \Omega(\frac{1}{1+3c'\gamma\beta/(\alpha-1)})$ECrit$(S^*, \mathbf{x}, Src_A)$, where γ, β and α are as defined in Lemma 3.2. The worst case running time of ApproxECrit is $O(|\mathcal{R}||E|k(2e)^k)$.*

4.2 Algorithm ApproxMaxCrit

Algorithm 3 ApproxMaxCrit(G, $H_{\mathcal{R}}$, k).

1: **for** $r \in H_{\mathcal{R}}$ **do**
2: Let C_r be the set of block groups within distance $B = O(k^{2/3})$ of r in $H_{\mathcal{R}}$. Construct graph $H_{\mathcal{R}}[C]$ induced by the block groups in C
3: Run Greedy(C, B) with the following modification: the source in the sampling step is picked from $V(C)$ randomly in each iteration. Let $r_1, r_2, \ldots, r_B$ be the block groups which are picked
4: Construct a minimum Steiner tree T_r of $r_1, \ldots, r_B$
5: **end for**
6: **for** $r \in H_{\mathcal{R}}$ **do**
7: Let $\text{wt}_r = \text{crit}(r)$
8: **end for**
9: Let $T' = k - \text{MaxST}(H_{\mathcal{R}}, \text{wt}, k)$ using the algorithm of [6]
10: return $\max\{\max_r \text{MaxCrit}(T_r), \text{MaxCrit}(T')\}$

Algorithm ApproxMaxCrit uses ideas from [15], who consider the problem of connected submodular function maximization. Theorem 3 gives a significantly better approximation bound with better running time than [15] by exploiting the spatial properties of our problem. As in the case of ApproxECrit, we also consider a quota Steiner tree and take the best of the two solutions.

Theorem 3. *For an instance $(G, H_{\mathcal{R}}, k)$, let $\hat{S}$ be the solution returned by ApproxMaxCrit. Let S^* be the optimum solution for this instance. If the aspect ratio of the bounding box containing $\mathcal{R}$ and each block group is constant, MaxCrit$(\hat{S}) \geq \Omega(\frac{1}{k^{1/3}})$MaxCrit$(S^*)$. The worst case running time is $O(|\mathcal{R}|k^{2/3} + |\mathcal{R}||E|k(2e)^k)$.*

Proof. (Sketch) For simplicity, assume each block group is a square; the arguments extend easily with a constant factor increase in the approximation bounds, since the aspect ratios are constant. Our proof is in two parts: (1) for any r, the Steiner tree T_r has at most k nodes, (2) there is a set of $O(k^{1/3})$ trees $T_{r'_1}, \ldots, T_{r'_s}$, such that they together cover S^*. We first argue that the theorem follows from these two statements. Statement (1) above implies that each T_r is a feasible solution to k-MaxCrit, since T_{r_i} is a connected subgraph in $H_{\mathcal{R}}$. Statement (2) implies $\sum_i \text{MaxCrit}(T_{r'_i}) \geq \text{MaxCrit}(S^*)$, by submodularity. Thus, there exists a node r_i such that $\text{MaxCrit}(T_{r'_i}) \geq \Omega(1/k^{1/3})\text{MaxCrit}(S^*)$, and the theorem follows.

We now prove statement (1). We consider any node r in $H_{\mathcal{R}}$. First, observe that a set of $O(k^{2/3})$ square subgraphs, each of side $O(k^{1/3})$ covers $H_{\mathcal{R}}$; let these be $y_1, \ldots, y_s$. Next, there exists a tree T' of length $O(k^{2/3} \cdot k^{1/3}) = O(k)$ that connects the centers of all the squares y_i. Then, T' can be augmented with additional paths to connect all the nodes $r_1, \ldots, r_B$, with only a constant factor increase in the number of nodes. This follows because each r_i is within some square y_j of size $O(k^{1/3}) \times O(k^{1/3})$, so that it can be connected to T' with a path of length at most $O(k^{1/3})$. Since $B = O(k^{2/3})$, tree T_r connects all the r_j's with a total length of $O(k)$.

Finally, we prove statement (2). Consider a tree T^* spanning S^*. We find the trees $T_{r'_1}, \ldots$ above in an iterative manner. First, pick a leaf r'_1 of T^*, and remove from T^* all the block groups which are within distance $k^{2/3}$ of r'_1, and repeat the process on the residual tree. Each such tree $T_{r'_1}$ covers at least $\Omega(k^{2/3})$ nodes of T^*. Therefore, $O(k^{1/3})$ trees computed in this manner cover T^*. □

4.2.1 Subroutine k-MaxST for the quota Steiner tree problem. Both algorithms ApproxECrit and ApproxMaxCrit involve solving an instance of the quota Steiner tree problem: given a graph $H_{\mathcal{R}}$, a weight wt_r for each $r \in \mathcal{R}$, and a parameter k, the objective is to compute a tree T' in $H_{\mathcal{R}}$ with at most k nodes, such that $\sum_{r \in T'} \text{wt}_r$ is maximized. There are constant factor approximations for this problem [21]. Here, we adapt the randomized fixed-parameter tractable algorithm of Cadena et al. [6] for Prize-Collecting Steiner Tree, which gives an optimal solution with high probability. The algorithm relies in the seminal color-coding technique of Alon et al. [2]. Naively, one could find a solution to k-MaxST by exhaustively checking all the possible $\binom{n}{k}$ subgraphs of k nodes in time $O(n^k)$. The algorithm does a random k-coloring of the nodes of $H_{\mathcal{R}}$, and it only considers maximum weight trees of each size that are "colorful"—this means all the nodes have distinct colors. It can be shown that such colorful solutions can be computed using a dynamic program. Further, the optimal solution is colorful with probability $k!/k^k$, which is large enough for the algorithm to work. Thus, the color coding technique allows us to reduce the search space to $O((2e)^k)$, keeping the computation feasible.

5 EXPERIMENTAL RESULTS

Our experiments focus on the following questions:

(1) **Finding critical clusters.** Can we find highly critical regions with our proposed methods? How do they compare to standard baselines used in public health? (Section 5.2)
(2) **Demographics.** What are the demographic properties of critical clusters? Where are they located? (Section 5.3)
(3) **MaxCrit vs. ECrit** What are the differences and similarities of the clusters discovered under the two proposed problems? (Section 5.4)

5.1 Experimental Setup

5.1.1 Dataset and disease model. A study of epidemics that spread through physical proximity requires social contact networks in which an edge represents an actual physical contact between two people at some location during the day. Such networks are not readily available and cannot be constructed easily because of the difficulty in tracking contacts for a large set of people. This has been recognized as a significant challenge in the public health community, and multiple methods have been developed to construct large scale realistic contact network models by integrating diverse public datasets (e.g., US Census, land use and activity surveys) and commercial data (e.g., from Dunn & BradStreet on location profiles). We use models developed by the approach of [8];[1] see also [9, 19] for network models developed by other public health groups.[2] Multiple such network models were evaluated in a study by the Institute of Medicine [10].

Here, we focus on a population for Minnesota with 5,048,920 individuals in total, which are aggregated into 4,082 census block groups from the 2010 U.S. census. We consider an SEIR type of stochastic model for measles, as described earlier in Section 2. For the MaxCrit formulation, the criticality of a cluster C of block groups

[1] See ndssl.vbi.vt.edu/synthetic-data/download for networks available for download.
[2] Models are available at http://www.epimodels.org/drupal/?q=node/70 and https://www.rti.org/impact/synthpop

is assessed by leaving every individual inside C unvaccinated; everybody else in the population is vaccinated with probability 0.97, which is the statewide vaccination rate. The source Src is picked as a set of three nodes in C. For the ECrit formulation, we focus on the Minneapolis metropolitan area, and pick Src to be a set of 100 children. As before, we assess the criticality of a cluster by leaving its inhabitants unvaccinated, with a 0.97 vaccination rate elsewhere.

5.1.2 Baseline Methods. We compare our algorithms with two heuristics used in epidemiology and a naive random baseline.

(1) Population. Find a cluster of size k with the largest total population. The motivation behind this heuristic is leaving as many people as possible unvaccinated.
(2) Vulnerability. The vulnerability of an individual is the probability that this person will get infected when the disease is left to propagate with no intervention—i.e., $x_v = 0$ for all nodes. This baseline finds a cluster of size k with as large total vulnerability as possible, thus prioritizing individuals who are most likely to get infected.
(3) Random. Find a connected cluster of size k by doing a random walk on the auxiliary graph.

5.2 Optimization power

In Figure 3, we show the criticality obtained by ApproxMaxCrit (top) and ApproxECrit (bottom) compared to the three baseline methods as a function of k. As expected, selecting subgraphs at random performs poorly and results in almost no additional infections compared to the initial disease conditions. Surprisingly, Vulnerability does not perform much better than random, especially on the MaxCrit objective. It is also interesting that the population-based heuristic does not have monotonic improvement with k. For the top plot, even though the subgraph of size 9 has 55,800 inhabitants, the smaller subgraph of size 5 with a population of 34,000 leads to a significantly larger outbreak. Overall, the population-based heuristic has better performance among the baselines, and it even surpasses our algorithm for $k = 5$ in MaxCrit. However, both ApproxMaxCrit and ApproxECrit exhibit notably better performance. For the ECrit objective, the 11-node cluster discovered using our method leads to 8 times more infections than the baselines.

Another important quantity is the probability of having a large outbreak. In Figure 4, we show the distribution of criticality values for each method over 100 simulations of the disease model. For the MaxCrit objective (top), we observe that even the largest outbreaks caused by Vulnerability and Random are much smaller than those of ApproxMaxCrit and the Population baseline. We also note that the population-based clusters have larger variance in criticality and can result in larger outbreaks than those from our algorithm. We observe a similar effect on the ECrit formulation (bottom), where the 9-node Population cluster has extreme cases with more infections than ApproxECrit. This suggests that if the goal for a public health department is to prevent the worst-case scenario, then intervening the most-populated areas is a good heuristic. However, in doing so, one could miss smaller regions that, on average, are likely to infect more people.

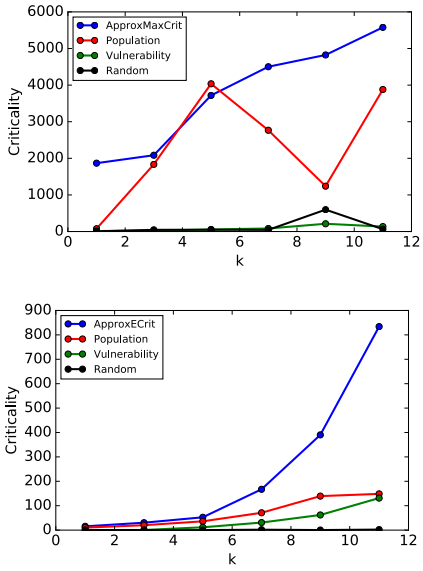

Figure 3: Comparison of algorithms for MaxCrit (top) and ECrit (bottom) as a function of the solution size k

5.3 Critical clusters and demographics

We compare the distribution of age and income in the cluster discovered by ApproxMaxCrit ($k = 11$) to that of the entire state. We aggregate household income into "Low" (below $25,000), "Medium" (between $25,000 and $75,000), and "High" (above $75,000). Ages are binned into "Pre-school" (below 5 years old), "School" (between 5 and 18 years old), "Adult" (between 18 and 70 years old), and "Senior" (above 70 years old). In Figure 5, we see the critical cluster has significantly more households of low income compared to the entire state—19.6% to 34.9%. Similarly, in the discovered cluster, children are over-represented. 26.6% of the population are children in "School" age compared to the national average of 18.7%.

We find critical clusters in different regions over Minnesota. Figure 1 shows the top 10 non-overlapping clusters discovered using ApproxMaxCrit. The most critical cluster—with over 5,000 infections—is located on the rural northern part of the state, spanning the Leech Lake and Red Lake reservations. We note that this cluster results in the largest spread despite having a relatively small population of 14,910 people, compared to clusters in urban regions. For example, the second most critical cluster—north of Minneapolis—has 48,889 inhabitants.

In addition to analyzing the most critical cluster, we look at the top-5 non-overlapping clusters discovered by ApproxMaxCrit. These correspond to different choices of root on the k-MaxST algorithm. In Table 3, we report the total population size, criticality, and percentage of infections to the total population of the cluster—i.e., criticality / population. Note that this latter number could be larger than 1, since there are infections outside the cluster. As we discussed before, the top region leads to a large spread (41% of its population size) despite having less inhabitants than the successive clusters. However, the second cluster follows right after, with virtually the same criticality score, but in a more urban region.

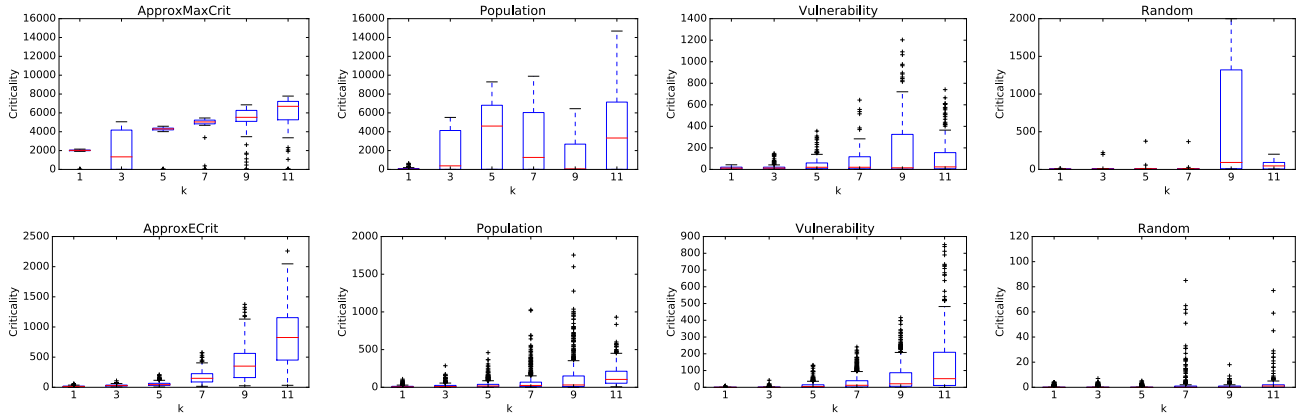

Figure 4: Criticality scores on the MaxCrit objective (top) and ECrit objective (bottom) over 100 runs of the simulation for each method evaluated

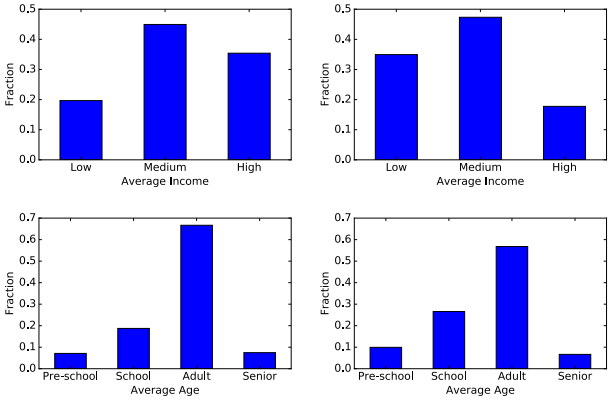

Figure 5: Average income (top) and age (bottom) in the entire state (left) and in the cluster discovered by ApproxMaxCrit. There are more children in school age and lower income households in the discovered critical cluster.

Table 3: Total population and criticality in the top 5 clusters discovered by ApproxMaxCrit

Rank	Population	Criticality	% population
1	14,910	6,138	41.2%
2	48,889	6,093	12.5%
3	23,391	1,388	5.9%
4	15,731	647	4.1%
5	9,936	372	4.7%

For ECrit, we focus on Minneapolis. In Figure 6, we show the most critical clusters for this region. The cluster that produces the largest spread covers the city of Brooklyn Park, which is a "majority-minority" suburb with a large immigrant population.[3] However, we emphasize the need for domain-expert analysis to better interpret and make use of these results. In Table 4, we report the population

[3]https://tinyurl.com/y97k7y2l

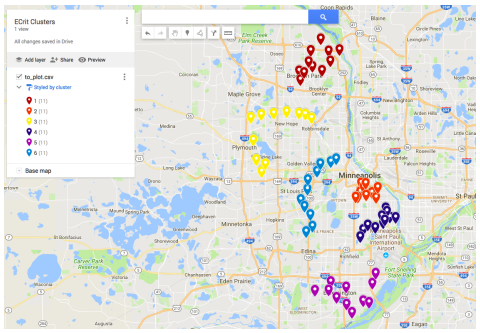

Figure 6: Critical clusters in Minneapolis on the ECrit objective with seeds being children of ages 10 and below.

Table 4: Total population and criticality in the top 5 clusters discovered by ApproxECrit

Rank	Population	Criticality	% population
1	22,273	858	3.9%
2	20,149	58	.3%
3	12,248	40	.3%
4	8,998	14	.2%
5	9,620	12	.1%

and criticality for the 5 most critical clusters. The difference in criticality between the first and second clusters is striking even though their population size is very similar.

5.4 MaxCrit and ECrit

In order to compare clusters from both formulations, we repeat our experiments for the MaxCrit objective on the Minneapolis area instead of the entire state. We find that the clusters discovered with both formulations overlap by a large margin. In Figure 7, we show the MaxCrit clusters in orange circles and the ECrit clusters in blue markers. Not only do the clusters cover the same parts of Minneapolis, but the criticality ranking is the same too. For instance,

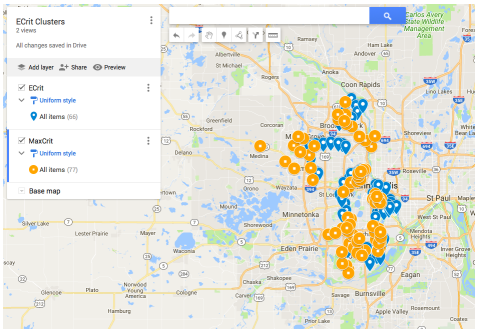

Figure 7: MaxCrit and ECrit clusters in Minneapolis. Both solutions find similar critical clusters.

the most critical cluster using MaxCrit covers Brooklyn Park, just as the ECrit cluster that we discussed in the previous section; this result holds even though the seeds for MaxCrit are chosen from the entire population, whereas we chose children only for ECrit.

6 RELATED WORK

Traditionally, epidemiological models have been differential equation models, which assume very simplistic mixing patterns of the underlying population. In the last decade, a number of research groups have developed agent-based methods using complex network models as a way to handle these issues [8–10, 18, 19]. Such methods have been used for policy analysis by local and national government agencies [10]. Since data for large scale contact networks is not available, we use this paradigm in our work.

All prior work on undervaccinated clusters has been restricted to identification. For instance, Lieu et al. [17] analyze electronic health records among children in 13 counties in Northern California and identify various significant geographic clusters of underimmunization and vaccine refusal, using spatial scan statistics. However, such methods are not directly useful for the policy questions of identifying *critical* clusters, which is our focus here.

There has been a lot of work on different kinds of detection problems related to outbreaks in networks. For instance, Christakis and Fowler [7] use the "friend of random people" approach to monitor a subset of people and infer characteristics of the epicurve for the entire population. Leskovec et al. [16] study the problem of early detection of different kinds of events—e.g., in water networks or social networks. However, these approaches have been focused on either just detecting that some event (e.g., start of an infection) has occurred or the epidemic characteristics for the entire region. Instead, we are interested in finding regions that would lead to a big number of infections if left unvaccinated.

Our work is also related to submodular function maximization with connectivity constraints. This constraint makes the problem much harder than other constraints, such as cardinality or matroid constraints, which can be approximately optimized using a simple greedy procedure [20]. The most relevant work is by Kuo et al. [15], who proposed a $\Omega(1/\sqrt{k})$ approximation algorithm to this problem. We are able to obtain an improved $\Omega(1/k^{1/3})$ approximation for MaxCrit by exploiting the spatial structure in our problem. Finally, Krause et al. [14] propose an approximation algorithm for budgeted submodularity maximization on graphs based on exploiting local structure. Our algorithm for ECrit builds on this approach by exploiting a slightly different type of local modularity bound.

7 CONCLUSIONS

Our work is motivated by public health policy questions of quantifying potential risks of large outbreaks as a result of reducing vaccination rates in a cluster. We formalize two problems, ECrit and MaxCrit, for finding critical clusters for highly contagious diseases that can be prevented by vaccination. These two formulations have different properties and solution structure, and they capture two different policy questions. We show that these problems are variants of the classical influence maximization problem, with an additional connectivity requirement on an auxiliary network, and we design algorithms with rigorous approximation guarantees. Experimental results show that our formulations perform significantly better than heuristics from epidemiology. Such an approach can help public health agencies prioritize response to the challenges of reduced vaccination coverage.

REFERENCES

[1] 2018. Critical spatial clusters for vaccine-preventable diseases. https://tinyurl.com/yc2tw7u7. (2018).
[2] Noga Alon, Raphael Yuster, and Uri Zwick. 1995. Color-coding. *Journal of the ACM (JACM)* (1995).
[3] R.M. Anderson and R.M. May. 1991. *Infectious Diseases of Humans.* Oxford University Press, Oxford.
[4] Jessica E. Atwell et al. 2013. Nonmedical Vaccine Exemptions and Pertussis in California, 2010. *Pediatrics* (2013).
[5] Christian Borgs et al. 2014. Maximizing Social Influence in Nearly Optimal Time. In *Proc. SODA*. 946–957.
[6] Jose Cadena, Feng Chen, and Anil Vullikanti. 2017. Near-Optimal and Practical Algorithms for Graph Scan Statistics. In *SIAM Data Mining (SDM)*.
[7] N.A. Christakis and J.H. Fowler. 2010. Social Network Sensors for Early Detection of Contagious Outbreaks. *PloS one* 5, 9 (2010), e12948.
[8] S. Eubank et al. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429 (2004), 180–184. Issue 6988.
[9] N.M. Ferguson, D.A.T. Cummings, C. Fraser, J.C. Cajka, P.C. Cooley, and D.S. Burke. 2006. Strategies for mitigating an influenza pandemic. *NATURE-LONDON-* 442, 7101 (2006), 448.
[10] M. Halloran et al. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. In *PNAS*. 4639–4644. PMCID:PMC2290797.
[11] D. Johnson, M. Minkoff, and S. Phillips. 2000. The Prize Collecting Steiner Tree Problem: Theory and Practice. In *ACM SODA*.
[12] D. Kempe, J. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD*. 137–146.
[13] J. Kleinberg. 2000. The small world phenomenon: An algorithmic perspective. *Proceedings of STOC* (2000).
[14] Andreas Krause et al. 2006. Near-optimal sensor placements: Maximizing information while minimizing communication cost. In *International conference on Information processing in sensor networks*. ACM, 2–10.
[15] Tung-Wei Kuo, Kate Ching-Ju Lin, and Ming-Jer Tsai. 2015. Maximizing Submodular Set Function With Connectivity Constraint: Theory and Application to Networks. *IEEE/ACM Transactions on Networking* 23, 2 (2015), 533–546.
[16] Jure Leskovec et al. 2007. Cost-effective outbreak detection in networks. In *KDD*. 420–429.
[17] Tracy A Lieu et al. 2015. Geographic clusters in underimmunization and vaccine refusal. *Pediatrics* 135, 2 (2015), 280–289.
[18] F. Liu et al. 2015. The role of vaccination coverage, individual behaviors, and the public health response in the control of measles epidemics: an agent-based simulation for California. *BMC Public Health* 15, 1 (01 May 2015), 447.
[19] Ira M. Longini et al. 2005. Containing Pandemic Influenza at the Source. *Science* 309, 5737 (August 2005), 1083–1087.
[20] GL Nemhauser, LA Wolsey, and ML Fisher. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming* 14, 1 (1978), 265–294.
[21] Ramamurthy Ravi et al. 1996. Spanning trees—short or small. *SIAM Journal on Discrete Mathematics* 9, 2 (1996), 178–200.