Critical spatial clusters for vaccine preventable diseases

Jose Cadena[1], Achla Marathe[2,4], and Anil Vullikanti[3,4]

[1] Lawrence Livermore National Laboratory, Livermore CA, USA
[2] Department of Public Health Sciences, University of Virginia
[3] Department of Computer Science, University of Virginia
[4] Network Systems Science and Advanced Computing Division, Biocomplexity Institute, University of Virginia, Charlottesville VA, USA

Abstract. The standard public health intervention for controlling the spread of highly contagious diseases, such as measles, is to vaccinate a large fraction of the population. However, it has been shown that in some parts of the United States, even though the average vaccination rate is high, geographical clusters of undervaccinated populations are emerging. Given that public health resources for response are limited, identifying and rank-ordering *critical* clusters can help prioritize and allocate scarce resources for surveillance and quick intervention.
We quantify the criticality of a cluster as the additional number of infections caused if the immunization rate in a cluster reduces. This notion of criticality has not been studied before, and, based on clusters identified in prior research, we show that the current underimmunization rate in the cluster, and its criticality are not correlated. We apply our methods to a population model for the state of Minnesota, where we find undervaccinated clusters with significantly higher criticality than those obtained by other natural heuristics.

1 Introduction

Many highly contagious childhood diseases, such as measles, can be prevented by vaccination. Thus, it is worrisome that large disease outbreaks have occurred in recent years, such as the measles outbreaks in the Pacific Northwest in 2019, in New York City in 2018, and in Minnesota in 2017—this is despite high vaccination coverage in the US—e.g., $\sim 95\%$ for MMR, the measles vaccine.

One of the reasons for the emergence of underimmunized geographical clusters, such as in California [15] and Minnesota [5], is misperceptions about the side effects of vaccines [2]. The typical response by public health agencies is to monitor clusters where immunization rates are falling, run active information campaigns, and engage community leaders.

Analyzing public school immunization records, Cadena et al. [5] identify six clusters in Minnesota that are statistically significant in terms of lower immunization rates relative to the statewide level. However, implementing public health interventions in all these clusters would be costly and time-consuming for public

health agencies, which motivates the following question: ***which of these clusters pose the most risk, and should be prioritized for treatment?*** A similar question was raised by Metcalf et al. [18], who stated that *"[t]here is also a need to understand under what conditions such clusters become at risk for epidemic spread, and the risk they pose to surrounding groups where vaccine coverage may be high."* It is useful to consider not only clusters in which the rates are presently low, but also the clusters that would pose a risk if fewer people within them were vaccinated. We develop a method to address these important public health policy questions. Our contributions are summarized below.

1. Formalizing criticality. We formalize the notion of *criticality* of a subset $S \subseteq V$ in a social contact network $G = (V, E)$, as the *expected number of additional infections* that would occur if the immunization rate within S is "low" compared to the statewide rate. Extending this notion, we introduce the MaxCrit problem: find a cluster S, which is (1) contiguous in space and (2) has the maximum criticality. The spatial proximity is motivated by the structure of clusters identified in [15, 2, 5], which are small and connected—this is desirable from a public health response perspective, since interventions involve field work. Spatial clustering can also help identify common risk factors, such as vulnerable communities and neighborhoods [3]. We estimate the criticality of a given cluster using a detailed agent-based simulation of the spread of measles in a population. However, solving the MaxCrit problem turns out to be a computationally challenging, and we design a greedy algorithm ApproxMaxCrit for this problem.

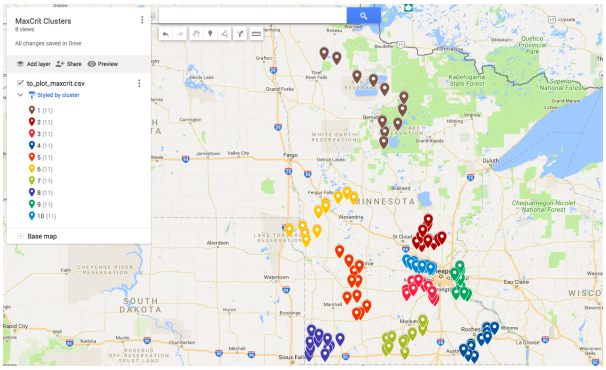

Fig. 1. Critical sets in Minnesota discovered using our methods. These are contiguous regions that lead to large simulated measles outbreaks if left undervaccinated.

2. Application. We study the phenomenon of criticality on a detailed population and contact network model for the state of Minnesota. We compute the criticalities of the significant underimmunized clusters reported in [5]. Quite surprisingly, we find that: (1) the cluster with the lowest vaccination rate among these is not the most critical, and (2) the criticality of the cluster computed using our algorithm is more than 10 times that of any of the clusters identified by [5]. We solve the MaxCrit problem and find clusters with very high criticality, compared to heuristics commonly considered for public health interventions. Our algorithm also achieves over 25% higher criticality for the objective compared

to all the baselines. Our methods can also combine social and demographic data for these clusters, available from the US Census, so they can be characterized, which may further guide targeted interventions. The critical clusters shown in Figure 1 involve people with lower than average income and age (Section 4).

Finally, due to lack of publicly available high-resolution, geo-located outbreak data, there is no easy way to validate our results, but we note that *one of the clusters we found to be critical lies in the Minneapolis metropolitan area where a large measles outbreak occurred in 2017 [10].*

Social impact. Our method for finding critical sets, applied to detailed population and contact network models, provides an operational tool for public health agencies to prioritize their limited surveillance and public outreach resources towards the most critical clusters. Our results imply that it is important to not only identify the undervaccinated clusters as in [5], but also determine which among them will likely cause an outbreak or an epidemic.

2 Preliminaries

2.1 Disease spread on a social contact network

Let V denote a population, and let $G = (V, E)$ be a contact graph on which a disease can spread. A person or node $v \in V$ can propagate the disease to its neighbors. There is an edge between two people if they come into close proximity during a typical day. Additionally, each person v is associated with a geographical location—i.e., their place of residence—denoted by $\text{loc}(v)$; we will consider such locations at the resolution of census block groups. Let $\mathcal{R}$ denote the geographical area 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 $r_i \in \mathcal{R}$, we use $V(r_i)$ to denote the set of nodes associated with location r_i; that is, those with $\text{loc}(v) \in r_i$. Analogously, for a set of block groups or *region* $R \subset \mathcal{R}$, let $V(R) = \cup_{r_i \in R} V(r_i)$ be the set of nodes located within R. We consider a graph $H_{\mathcal{R}} = (\mathcal{R}, E_{\mathcal{R}})$ on the set of block groups, where two block groups are connected if they are geographically contiguous, i.e., they are adjacent on a map. In particular, we are interested in *connected* subgraphs of $H_{\mathcal{R}}$. We use $\text{Conn}(\mathcal{R})$ to denote all the subsets $R \subset H_{\mathcal{R}}$ that are spatially connected. These definitions are illustrated in Figure 2.

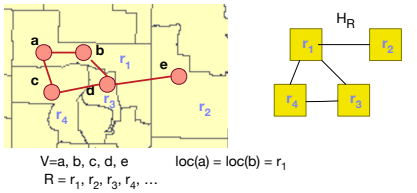

Fig. 2. Notation example. The 5 circle nodes (a–e) form a social contact network. Each node resides in a block group r_i, and these block groups form the block group graph $H_{\mathcal{R}}$, where an edge represents that the block groups are adjacent on the map.

For $u, v \in \mathcal{R}$, let $\texttt{dist}_{H_{\mathcal{R}}}(u, v)$ denote the distance between u and v in the graph $H_{\mathcal{R}}$, which is equal to the length of the shortest path between them. The ball centered at v, with radius ℓ is defined as $B_{H_{\mathcal{R}}}(v, \ell) = \{u : \texttt{dist}(u, v) \leq \ell\}$,

which is the set of all nodes within distance ℓ of v. When the graph is clear from the context, we drop it from the subscript in the notation for $B(\cdot)$ and $\texttt{dist}(\cdot)$.

Disease model. We use an SEIR model for diseases like measles [1], where a node is in one of four states: Susceptible (S), Exposed (E), Infected (I), and Recovered/Removed (R). Measles is highly contagious; an infected node spreads the disease to each susceptible neighbor with high probability. In our simulations, we assume a transmission probability of 1, but our methods extend to the more general case. If a node is vaccinated, it does not get infected. We assume 100% vaccine efficacy, but this assumption is not crucial for our methodology.

Let γ denote the average region-wide vaccination rate—around 0.97 in Minnesota. Let $\mathbf{x}$ be a *vaccination* or *intervention* vector: $x_i \in [0, 1]$ denotes the probability that node (i.e., person) i is vaccinated (so $x_i = \gamma$, by default). Let Src_A denote the source of the infection or *initial conditions* of the disease process: this could be one or a small number of nodes from a region $A \subset \mathcal{R}$, which initially get infected. We use $\#\mathrm{inf}(\mathbf{x}, \mathrm{Src}_A)$ to denote the expected number of infections given an intervention $\mathbf{x}$ and initial conditions Src_A. When Src_A is clear from the context, we simply use $\#\mathrm{inf}(\mathbf{x})$.

2.2 Criticality

For a vaccination vector $\mathbf{x}$, let $\mathbf{x}^S$ denote the corresponding intervention where a subset $S \subset V$ of nodes is undervaccinated. 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. Without loss of generality, we consider $\gamma' = 0$ for mathematical convenience.

We define the **criticality** of a set $S \subset V$ as the *expected number of additional infections that occur if S is not vaccinated*, with respect to some initial condition Src_A. Since we are interested in finding spatial clusters of high criticality, we focus on $S = V(R)$ for a connected region $R \in \mathrm{Conn}(\mathcal{R})$. Then, we define the criticality of a region as

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

which is the expected number of extra infections if nodes in the region R are undervaccinated. In order to simplify the notation, we will drop $\mathbf{x}$ and Src from the inputs to $\mathrm{crit}(\cdot)$, whenever it is clear from the context.

2.3 Problem Formulation

Modeling considerations. In practice, public health interventions involve intensive field work, and they are most effective when focused within small, localized geographical regions. Therefore, we aim to find regions that have high criticality *and* are small in size. In modeling terms, this can be accomplished by adding a size parameter k, which can be tuned based on the available public health resources. Given the discussion above, we pose the task of finding spatial clusters of high criticality as the following optimization problem.

Problem 1 (MaxCrit$(G, H_{\mathcal{R}}, k)$). Given an instance $(G, H_{\mathcal{R}}, k)$, find a connected region $R \in \text{Conn}(\mathcal{R})$ of size at most k that maximizes criticality over all choices of source:

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

In words, the MaxCrit problem involves maximizing over *all* possible choices of the sources $\text{Src}_{R'}$ in the cluster R'. From a public health perspective, our problem models the following question: *what is the most critical cluster of size k if the disease starts within the undervaccinated cluster itself?* An obvious question is how should the parameter k be chosen. This can depend on a number of factors, such as availability of medical resources, jurisdiction constraints, social and ethical considerations, location of under-served communities etc. [7].

3 Our approach

MaxCrit is closely related to the Influence Maximization problem [12]. The influence function is known to be submodular—informally, this means that the function has a diminishing returns property, as a result of which, a greedy algorithm gives a good approximation. We show that the crit function is also submodular by following the approach of Kempe et al. [12]. However, a crucial difference in our case is that the decision space is restricted to sets S that are connected. We design algorithm ApproxMaxCrit (as discussed in [4]), by adapting the technique of Kuo et al. [13], who give an $\Omega(1/\sqrt{k})$ approximation algorithm to find a connected subset of size k that maximizes a submodular function.

4 Experimental results

Our experiments focus on the following questions:

1. **Relationship between criticality and underimmunization.** Is the criticality of a cluster directly correlated with its underimmunization rate?
2. **Finding critical clusters.** Can we find highly critical regions with our methods? How do they compare to standard public health heuristics?
3. **Characteristics of critical clusters.** What are the demographic properties of critical clusters? Where are they located?

Dataset and disease model. Simulation of an infectious disease epidemic that spreads through physical proximity requires social contact networks in which an edge represents physical contact between two people. Such networks cannot be constructed easily because of the difficulty in tracking contacts for a large set of people throughout the day. This has been recognized as a significant challenge in the public health community, and multiple methods have been developed to construct realistic contact network models by integrating diverse public datasets (e.g., US Census, land use, and activity surveys) and commercial data (e.g., from Dun & Bradstreet on location profiles). We use agent-based models developed by the approach of [8]; see also [17, 9] for network models developed by other

public health groups. Multiple such network models were evaluated in a study by the Institute of Medicine [11].

Here, we focus on the population of Minnesota (MN) with $5,048,920$ individuals in total, aggregated into 4,082 census block groups from the 2010 U.S. census. We consider an SEIR stochastic model for measles, as described in Section 2. The criticality of a region R of block groups is assessed by leaving every individual inside R unvaccinated; everybody else in the population is vaccinated with probability 0.97, which is the statewide vaccination rate. We also use the underimmunized clusters in MN computed by [5] in our analysis here.

Baseline Methods We compare our algorithms with two heuristics used in public health 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 $H_{\mathcal{R}}$.

1. Relationship between criticality and underimmunization We compute the criticality of the four most significant underimmunized clusters in MN, as identified by authors in [5]. The clusters are numbered 1–4 based on their statistical significance with respect to underimmunization rates, so that cluster 1 is more significant than cluster 4. However, as shown in Figure 3 (left), it seems clear that *the outbreak size of cluster 4 is much higher than that of cluster 1*—the 95th percentile value for the number of infections in cluster 4 is almost four times that of cluster 1. The results show that criticality is not directly correlated with the level of underimmunization. Instead, network structure plays a more important role in determining the criticality.

2. Optimization power In Figure 3 (right), we show the criticality obtained by ApproxMaxCrit 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. It is also interesting that the population-based heuristic does not have monotonic improvement with k. 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 much larger outbreak. Overall, the Population heuristic has better performance among the baselines, and it even surpasses our algorithm for $k = 5$. However, ApproxMaxCrit exhibits notably better performance in general. The maximum improvement on

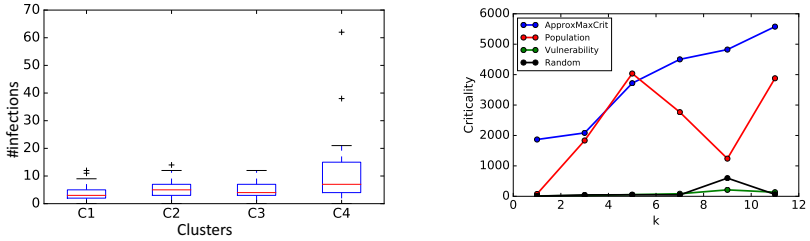

Fig. 3. Left: Distributions of the number of infections resulting from an outbreak starting in each of four underimmunized clusters in MN identified in [5]. Right: Comparison of algorithms for the MaxCrit problem as a function of the solution size k.

criticality occurs on the 9-node cluster, where our method finds a cluster that leads to 4 times more infections than the POPULATION baseline.

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. 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. 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.

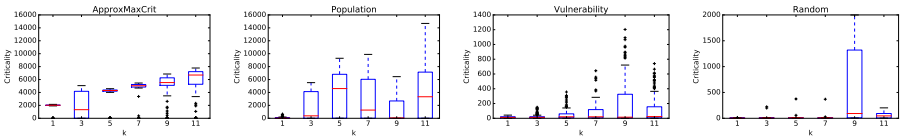

Fig. 4. Criticality scores over 100 runs of the disease model for each method evaluated

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, children are over-represented; 26.6% of the population are children in "School" age compared to the 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

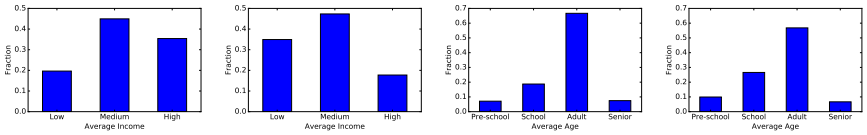

Fig. 5. Average income (top) and age (bottom) in the entire state (left) and in the cluster discovered by ApproxMaxCrit (right). There are more children in school age and lower income households in the discovered critical cluster.

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 1, 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. The second cluster has very similar criticality score, but in a more urban region.

Table 1. Population and criticality in the top 5 clusters found 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%

Finally, we repeat our experiments for MaxCrit on the Minneapolis area instead of the entire state. The most critical cluster covers Brooklyn Park, where measles outbreaks occurred in 2017 and 2019[5]. However, we emphasize the need for domain-expert analysis to better interpret and make use of these results.

5 Related Work

Mathematical models have played an important role in epidemiology for over a century [1]. Traditionally, epidemiological models have been differential equation models, which assume very simplistic mixing patterns of the underlying population. In the last decade, several research groups have developed agent-based methods using complex networks as a way to model more realistic mixing [8, 17, 9, 16]. Such methods have been used for policy analysis by local and national government agencies [11]. We use this paradigm in our work.

[5] https://tinyurl.com/y359zapv

All prior work on undervaccinated clusters has been restricted to identifying these clusters. For instance, [15] analyze health records of children in Northern California to identify significant clusters of underimmunization and vaccine refusal using spatial scan statistics. However, such methods are not directly useful for the question of identifying *critical* clusters, which is our focus. There is a large body of work related to outbreak detection in networks. [6] use the "friend of random people" effect to monitor a subset of people and infer characteristics of the epidemic curve for the entire population. [14] study early detection of different kinds of events—e.g., in 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.

6 Conclusions

Prior research has identified geographical clusters of undervaccinated populations in many states. However, the potential risk of causing large outbreaks from such clusters is not well understood, and actionable response requires a way to prioritize the threat from these undervaccinated clusters. Public health response (e.g., surveillance and field work) is very costly, and therefore, a method to quantify such risk is an important public health contribution.

This research makes several contributions: (i) we formalize the problem MaxCrit for finding critical clusters for highly contagious diseases that can be prevented by vaccination, and that will lead to large outbreaks if left unvaccinated; (ii) we combine a detailed agent-based model of Minnesota and its social contact network with a disease model to compute a realistic measure of clusters' criticality; (iii) we find clusters that have higher criticality than discovered by baseline methods; (iv) we characterize the clusters, and (v) we provide a way to prioritize intervention based on the availability of resources.

This research has a broader applicability than just the spread of measles and infectious diseases. Other societal problems that have a component of social connectedness and propagation potential e.g. depression, addiction, suicides etc. can also be studied with this methodology.

7 Acknowledgements

This study has been partially supported by the NIH grant 1R01GM109718, NSF BIG DATA Grant IIS-1633028, NSF DIBBS Grant ACI-1443054, DTRA subcontract/ARA S-D00189-15-TO-01-UVA. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-806042.

References

1. Anderson, R., May, R.: Infectious Diseases of Humans. Oxford University Press, Oxford (1991)

2. Atwell, J.E., Otterloo, J.V., Zipprich, J., Winter, K., Harriman, K., Salmon, D.A., Halsey, N.A., Omer, S.B.: Nonmedical vaccine exemptions and pertussis in california, 2010. Pediatrics (2013)
3. Azman, A.S., Lessler, J.: Reactive vaccination in the presence of disease hotspots. Proceedings of the Royal Society B: Biological Sciences **282**(1798), 20141341 (2015)
4. Cadena, J., Marathe, A., Vullikanti, A.: Finding spatial clusters susceptible to epidemic outbreaks due to undervaccination (extended abstract). In: Proc. AAMAS (2020)
5. Cadena, J., Falcone, D., Marathe, A., Vullikanti, A.: Discovery of under immunized spatial clusters using network scan statistics. BMC Medical Informatics and Decision Making **19**(1), 28 (2019)
6. Christakis, N., Fowler, J.: Social network sensors for early detection of contagious outbreaks. PloS one **5**(9), e12948 (2010)
7. Dummer, T.J.: Health geography: supporting public health policy and planning. Cmaj **178**(9), 1177–1180 (2008)
8. Eubank, S., Guclu, H., V. S. Anil Kumar, Marathe, M., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. Nature **429**, 180–184 (2004)
9. Ferguson, N., Cummings, D., Fraser, C., Cajka, J., Cooley, P., Burke, D.: Strategies for mitigating an influenza pandemic. NATURE-LONDON- **442**(7101), 448 (2006)
10. Hall, V., Banerjee, E., Kenyon, C., et al.: Measles outbreak—minnesota april-may 2017. MMWR Morb Mortal Wkly Rep pp. 713–717 (2017)
11. Halloran, M., Ferguson, N., Eubank, S., Longini, I., Cummings, D., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C., Burke, D., Cooley, P.: Modeling targeted layered containment of an influenza pandemic in the United States. In: PNAS. pp. 4639–4644 (March 10 2008), pMCID:PMC2290797
12. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD. pp. 137–146 (2003)
13. Kuo, T.W., Lin, K.C.J., Tsai, M.J.: Maximizing submodular set function with connectivity constraint: Theory and application to networks. IEEE/ACM Transactions on Networking **23**(2), 533–546 (2015)
14. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.S.: Cost-effective outbreak detection in networks. In: KDD. pp. 420–429 (2007)
15. Lieu, T.A., Ray, G.T., Klein, N.P., Chung, C., Kulldorff, M.: Geographic clusters in underimmunization and vaccine refusal. Pediatrics **135**(2), 280–289 (2015)
16. Liu, F., Enanoria, W., Zipprich, J., Blumberg, S., Harriman, K., Ackley, S.F., Wheaton, W.D., Allpress, J.L., Porco, T.C.: 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), 447 (May 2015)
17. Longini, I.M., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D.A., Halloran, E.M.: Containing pandemic influenza at the source. Science **309**(5737), 1083–1087 (August 2005)
18. Metcalf, C., Andreasen, V., Bjørnstad, O., Eames, K., Edmunds, W., Funk, S., Hollingsworth, T., Lessler, J., Viboud, C., Grenfell, B.: Seven challenges in modeling vaccine preventable diseases. Epidemics **10**(Supplement C), 11 – 15 (2015). https://doi.org/https://doi.org/10.1016/j.epidem.2014.08.004, http://www.sciencedirect.com/science/article/pii/S1755436514000395, challenges in Modelling Infectious DIsease Dynamics