

Data-Driven Near-Optimal Regional Epidemic Surveillance

Rares-Mihail Neagu*

rneagu@uiowa.edu

University of Iowa

Iowa City, Iowa, USA

Jeffrey Keithley

jeffrey-keithley@uiowa.edu

University of Iowa

Iowa City, Iowa, USA

Xin Zan

xin-zan@uiowa.edu

University of Iowa

Iowa City, Iowa, USA

Sriram Pemmaraju

sriram-pemmaraju@uiowa.edu

University of Iowa

Iowa City, Iowa, USA

Bijaya Adhikari*

bijaya-adhikari@uiowa.edu

University of Iowa

Iowa City, Iowa, USA

Abstract

In the early stages of an epidemic outbreak, when the incidence data is limited, policy makers actively try to establish the scale and scope of the outbreak. Two of the most pertinent questions they seek to answer are *i)* has the outbreak started in a particular region (e.g. a country) and *ii)* how many sub-regions (e.g. a state or a county) have been affected thus far? We study these problems as a data-driven knowledge extraction task, i.e., given mobility and epidemiological data, we aim to extract actionable knowledge about which regions should receive additional surveillance resources so that an outbreak is detected early and its spatial footprint is accurately characterized. To this end, we formulate two discrete-lattice optimization objectives for epidemic surveillance, 1) maximizing the probability of detecting at least one infected sub-region and 2) maximizing the expected number of infected sub-regions detected. Both objectives turn out to be NP-hard. However, we prove that they are both submodular, which paves the way for efficient epidemic surveillance from limited data. We use this structure to derive curvature bounds for the objective. We also relate the bound with the parameters of the metapopulation disease dynamics. We propose a scalable algorithm, LATTICE-ADAPTED MODIFIED GREEDY (LAMGREEDY), that uses the curvature, and marginal-gain machinery to establish constant factor approximation guarantee. Extensive experiments on real-world mobility networks for COVID-19 and Influenza across three U.S. states show that LAMGREEDY consistently outperforms natural and the state-of-the-art baselines under multiple cost models. The empirical results validate the prevalence and curvature estimates, and demonstrate that the proposed tools extract high-quality surveillance knowledge from mobility and epidemic data under realistic budget constraints.

CCS Concepts

• **Applied computing** → **Health informatics**; • **Theory of computation** → **Approximation algorithms analysis**; • **Mathematics of computing** → **Discrete optimization**; • **Computing methodologies** → *Modeling and simulation*.

*Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '26, Rome, Italy*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2539-5/2026/11

<https://doi.org/10.1145/3799682.3841066>

Keywords

epidemic surveillance; lattice optimization; submodular optimization; metapopulation models

ACM Reference Format:

Rares-Mihail Neagu, Jeffrey Keithley, Xin Zan, Sriram Pemmaraju, and Bijaya Adhikari. 2026. Data-Driven Near-Optimal Regional Epidemic Surveillance. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 07–11, 2026, Rome, Italy. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3799682.3841066>

1 Introduction

Just in the last two decades, a surprisingly high number of large-scale epidemic outbreaks have occurred including the 2009 H1N1 epidemic, 2016 West-africa Ebola epidemic and the COVID-19 pandemic [15, 26, 28]. As we speak, localized Hantavirus [44] and Ebola [43] outbreaks are taking place and could potentially expand to a larger scale outbreak resembling COVID-19 pandemic. Past outbreaks wreaked havoc due to inadequate surveillance and responses during the early stages [12]. While blanket surveillance methods, such as serological surveys, are available to collect large scale data for epidemic response, they are far too expensive to deploy effectively [17].

Key questions epidemiologists and decision makers have during an early stage of the pandemic are *1) how can we develop the best-value epidemic surveillance methods that enable us to detect an outbreak with a high success rate?* and *2) how to accurately estimate the spread (e.g., number of sub-regions with positive prevalence) of the outbreak?* Both of these questions have to be answered within a limited budget as public health officials are constrained by available resources. Previous work on epidemic surveillance has not satisfactorily answered these questions. Heuristic approaches are usually simple and rely on intuition, however they do not provide any type of performance guarantee [8, 31]. Conversely, metrics-based approaches, which are common in computational literature, treat the problem as a submodular optimization problem over sets, which means that resource allocations are uniform across selected regions [22]. However, real-world deployment requires deciding not just where to test, but also how many resources to allocate to each region to maximize detection likelihood. This necessitates a shift from binary set selection to the more granular framework of integer lattice optimization.

To overcome these constraints, we formulate epidemic surveillance as an optimization problem over a integer lattice [32, 33], and apply this approach to a disease-agnostic representation of a

metapopulation model. While lattice optimization has been used in the field of epidemiology to a limited degree (e.g. healthcare associated infection monitoring [1]), it has not been used in this context. Here, we first formulate the two questions mentioned above as lattice optimization problems. Through theoretical analysis, we demonstrate their inherent computational difficulty, and show how each of these two problems have specific mathematical properties that allow us to develop near-optimal solutions. Specifically, we prove that the objectives of both problems are submodular, investigate the curvature associated with each objective, and derive approximate relationships between the curvature and disease parameters [10, 37, 47]. We provide a scalable lattice optimization approach which provides constant factor approximation guarantee. We also demonstrate that even better approximation guarantee is theoretically possible for an exponential algorithm [14, 33].

Our contributions are summarized as follows.

- First, we formulate two regional surveillance problems as novel lattice optimization objectives, proving that while they are NP-hard, they retain the property of submodularity [32].
- Second, we derive theoretical bounds for the curvature of these functions and explicitly connect these bounds to the epidemiological parameters of the meta-population disease model [13, 39].
- Third, we introduce two optimization algorithms; notably, one of these algorithms provides the first-ever $(1 - \kappa/e - \epsilon)$ approximation guarantee for lattice optimization problems under knapsack constraints [36, 47].
- Finally, we validate our approach through extensive experiments simulating COVID-19 and Influenza outbreaks across three U.S. states, demonstrating the practical efficacy of our methods [15, 26, 38].

2 Related Work

Our work is connected to three lines of research: submodular optimization, epidemic modeling, and surveillance resource allocation. **Lattice Submodularity and Optimization:** Submodularity is a standard tool for designing approximation algorithms in combinatorial optimization [10]. A well-known example in data mining is the CELF-style sensor placement formulation of outbreak detection [22]. Most of this literature treats a solution as a set. This is restrictive for our setting, because selecting a county once does not say how many tests or surveillance visits should be placed there. Work by Soma and Yoshida extends diminishing returns to integer lattices [32, 33], and Ene and Nguyen give a reduction from lattice problems to set functions [14]. We use these tools, but our emphasis is on the surveillance instance that is created from epidemic and mobility data, and on the scalable greedy algorithm that acts on that instance. Curvature-dependent guarantees [47] are useful in this paper mainly because they let us certify the behavior of the practical algorithm.

Epidemic Burden and Modeling: Healthcare-acquired infections and respiratory epidemics both create substantial health and economic burdens [7, 34]. Prior work has studied pathogens such as MRSA [11, 19] and *C. difficile* [6, 9, 35, 46], as well as larger-scale respiratory outbreaks. These systems are commonly modeled with

differential equations [24, 25, 40] or detailed agent-based simulations [3, 16, 41, 45]. In our setting, the metapopulation SEIR model provides the bridge between data and optimization: mobility flows and disease parameters are converted into county-level prevalence estimates, and those estimates become the inputs to the surveillance objective. We use the next-generation matrix framework [13, 39] and analytical SEIR approximations [27, 28] for this step.

Surveillance and Outbreak Detection: Outbreak detection has been studied using threshold rules, scan statistics, simulation-based evaluation, and data mining methods [21, 23]. Other approaches use social network sensors [8, 31], epidemiological network models [29], vaccination and defense optimization [5, 48, 49], and epidemic reconstruction [30]. The closest line of work to ours uses submodular optimization for surveillance and monitoring [1, 18]. We build on this style of problem formulation, but move from set selection to integer allocations. This lets the algorithm decide both where to monitor and how much resource to allocate to each region.

3 Preliminaries

In the following, we denote the set of non-negative integers by $\mathbb{Z}_+$ and the set of n first natural numbers by $[n]$.

Metapopulation SEIR Dynamics: A metapopulation model is an extension of classical compartmental epidemiological models (like SIR and SEIR) into a set of $k \in \mathbb{Z}_+$ geographically distinct subpopulations, each of which might be experiencing a different phase of the outbreak. We denote the size of each subpopulation $i \in [k]$ as N_i . The SEIR metapopulation model is parameterized by β (transmission rate), σ (progression rate from exposed to infectious i.e., $1/\sigma$ = incubation period), and γ (recovery rate).

We describe the mobility between regions using a flux matrix $\mathbf{W} = [w_{ij}]$, where w_{ij} represents daily flow of individuals into region i from region j . Now, let S_i , E_i , I_i , and R_i denote the number of susceptible, exposed, infected, and recovered individuals in region i . Then, the dynamics that characterizes the epidemics in a metapopulation SEIR model is given by the following set of differential equations:

$$\begin{aligned} \frac{dS_i}{dt} &= -\beta \frac{S_i I_i}{N_i} + \sum_{j=1}^K (w_{ij} S_j - w_{ji} S_i), \\ \frac{dI_i}{dt} &= \sigma E_i - \gamma I_i + \sum_{j=1}^K (w_{ij} I_j - w_{ji} I_i), \\ \frac{dE_i}{dt} &= \beta \frac{S_i I_i}{N_i} - \sigma E_i + \sum_{j=1}^K (w_{ij} E_j - w_{ji} E_i), \\ \frac{dR_i}{dt} &= \gamma I_i + \sum_{j=1}^K (w_{ij} R_j - w_{ji} R_i). \end{aligned}$$

Lattice Submodularity: Our goal here is to design algorithms to allocate epidemic surveillance resources to different subpopulations in a metapopulation model. To do so, we take advantage of the powerful lattice submodular optimization machinery. Let allocation vector $\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{Z}_+^K$ denote the amount of resources allocated to each subpopulation. The allocation vector $\mathbf{x}$ is an integer lattice as each $x_i \in \mathbb{Z}_+$. Given two lattices $\mathbf{x}$ and $\mathbf{y}$, we say $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ for all i . Using standard notation, we denote the i -th standard basis vector by $\mathbf{e}_i$ and a vector of all ones by $\mathbf{1}$.

Let us consider a real valued lattice function $h : \mathbb{Z}_+^K \rightarrow \mathbb{R}$ (which, in our case, quantifies the ‘benefit’ gained due to an allocation vector). h is said to be a DR-Submodular function if it exhibits the *diminishing return* property, where each additional allocation returns lesser and lesser benefit. Specifically, the DR-Submodularity is defined as follows:

DEFINITION 1 (DR-SUBMODULARITY [32]). A function $h : \mathbb{Z}_+^K \rightarrow \mathbb{R}$ is DR-submodular if for all $i \in [K]$ and all $\mathbf{x} \leq \mathbf{y}$ in $\mathbb{Z}_+^K$,

$$h(\mathbf{x} + \mathbf{e}_i) - h(\mathbf{x}) \geq h(\mathbf{y} + \mathbf{e}_i) - h(\mathbf{y}).$$

DR-submodularity captures the intuitive property that marginal gains decrease as allocations increase. For lattices, there is also a weaker notion of submodularity referred to as Lattice Submodularity, which is defined as follows:

DEFINITION 2 (LATTICE SUBMODULARITY [32]). A function $h : \mathbb{Z}_+^K \rightarrow \mathbb{R}$ is lattice submodular if for all $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_+^K$,

$$h(\mathbf{x}) + h(\mathbf{y}) \geq h(\mathbf{x} \vee \mathbf{y}) + h(\mathbf{x} \wedge \mathbf{y}),$$

where $\vee$ and $\wedge$ denote coordinatewise maximum and minimum, respectively.

We have the following relationship between the two notions of submodularities for lattice functions.

LEMMA 3 ([32]). Every DR-submodular function on $\mathbb{Z}_+^K$ is lattice submodular.

This hierarchy is important as DR-submodularity is easier to verify while lattice submodularity is required by many approximation algorithms.

4 Problem Formulation

We formulate two resource allocation problems that arise naturally in the early stage of an outbreak: *detecting at least one infected region* and *detecting as many infected regions as possible*. The two objectives use the same data, but answer different surveillance questions. The first is useful when the main goal is to confirm that an outbreak exists. The second is useful when the response requires knowing where the infections are.

Setup and Notation: Consider a system of $k \in \mathbb{Z}_+$ regions, such as counties in a state, indexed by $i \in [K]$. Region i has population size N_i . We focus on an early outbreak snapshot in which some regions may already have infectious individuals. Let I_i be the number of infectious individuals in region i , and let $p_i = I_i/N_i \in [0, 1]$ denote the infection prevalence. Regions with no infection have $p_i = 0$.

Let $\mathbf{x} = (x_1, \dots, x_K) \in \mathbb{Z}_+^K$ be the allocation vector, where x_i is the number of surveillance resources assigned to region i . For example, x_i may be the number of tests, sentinel samples, or sampling visits allocated to the region. Each unit in region i has cost $c_i > 0$, and the allocation must satisfy the knapsack constraint

$$\sum_{i=1}^K c_i x_i \leq B,$$

where B is the total surveillance budget. The different choices of c_i used in the experiments are described in Section 6.

4.1 Detection Likelihood Maximization

The first objective maximizes the probability of detecting the outbreak in at least one region. Allocating x_i resources to region i corresponds to drawing x_i independent samples from a population with infection probability p_i . Let $D_i \in \{0, 1\}$ be the indicator that at least one infection is detected in region i . Then

$$\Pr(D_i = 1 \mid \mathbf{x}) = 1 - (1 - p_i)^{x_i}.$$

This detection model is standard in surveillance optimization [1]. Since the surveillance window is short relative to epidemic time scales, and samples are collected independently after the prevalence snapshot is fixed, we treat the detection events as conditionally independent. This gives the following optimization problem.

DETECTION LIKELIHOOD MAXIMIZATION (DLM): Given a system with $k \in \mathbb{Z}_+$ subpopulations experiencing an early stage of an outbreak as described by meta-population model $\mathcal{M}$, per unit cost c_i of allocating surveillance resources for each subpopulation i , and a budget constraint B , find an allocation vector $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_k^*)$, such that

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} 1 - \prod_{i=1}^K (1 - p_i)^{x_i}$$

and $\sum_{i=1}^K c_i x_i \leq B$

DLM is an early-warning objective. It is appropriate when a single confirmed detection is enough to trigger additional response.

4.2 Expected Number of Infected Counties Detected

The second objective asks for more spatial information. Instead of only detecting the outbreak somewhere, we want to maximize the expected number of infected regions that are found by surveillance. Using the same detection variables as above, the objective is $\sum_i \Pr(D_i = 1 \mid \mathbf{x})$.

EXPECTED NUMBER OF INFECTED COUNTIES DETECTED (ENCD): Given a system with $k \in \mathbb{Z}_+$ subpopulations experiencing an early stage of an outbreak as described by meta-population model $\mathcal{M}$, per unit cost c_i of allocating surveillance resources for each subpopulation i , and a budget constraint B , find an allocation vector $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_k^*)$, such that

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} \sum_{i=1}^K \Pr(D_i = 1 \mid \mathbf{x})$$

and $\sum_{i=1}^K c_i x_i \leq B$

ENCD is useful when the public health response depends on knowing the spatial extent of the outbreak. For example, county-specific interventions require more than one positive signal. Unfortunately, both problems DLM and ENCD are computationally hard.

THEOREM 4. Problems DLM and ENCD are NP-hard.

We omit the proof for brevity. The NP-hardness of the problem rules out exact optimization as a potential solution and motivates approximation algorithms that exploit the structure of the surveillance objective.

5 Approach

We now develop approximation-algorithms to solve the problems mentioned above. The main observation behind our approach is simple: once the mobility and disease data give us regional prevalences p_i , both DLM and ENCD become monotone objectives on an integer lattice. The marginal value of an additional test decreases as more tests are placed. We first use this property to obtain curvature bounds. We then use the same objects—the lattice-to-set reduction, marginal gains, costs, and curvature estimates—to develop our proposed algorithm LAMGREEDY.

5.1 Diminishing returns in our objectives

Let $q_i := 1 - p_i$. We denote the DLM and ENCD objectives by

$$f(\mathbf{x}) = 1 - \prod_{i=1}^K q_i^{x_i}, \quad g(\mathbf{x}) = \sum_{i=1}^K (1 - q_i^{x_i}).$$

Both objectives have diminishing returns on the integer lattice.

THEOREM 5 (DR-SUBMODULARITY OF DETECTION LIKELIHOOD). *The DLM objective $f(\mathbf{x})$ is DR-submodular on $\mathbb{Z}_+^K$.*

PROOF. For coordinate i , the marginal gain of one additional unit is

$$\Delta_i f(\mathbf{x}) = f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) = p_i \prod_{k=1}^K q_k^{x_k}.$$

If $\mathbf{x} \leq \mathbf{y}$, then $q_k^{x_k} \geq q_k^{y_k}$ for all k , since $q_k \in (0, 1]$. Therefore $\prod_k q_k^{x_k} \geq \prod_k q_k^{y_k}$ and $\Delta_i f(\mathbf{x}) \geq \Delta_i f(\mathbf{y})$. $\square$

THEOREM 6 (DR-SUBMODULARITY OF EXPECTED DETECTIONS). *The ENCD objective $g(\mathbf{x})$ is DR-submodular on $\mathbb{Z}_+^K$.*

PROOF. The coordinate marginal is

$$\Delta_i g(\mathbf{x}) = g(\mathbf{x} + \mathbf{e}_i) - g(\mathbf{x}) = p_i q_i^{x_i}.$$

For $\mathbf{x} \leq \mathbf{y}$, we have $x_i \leq y_i$, and hence $q_i^{x_i} \geq q_i^{y_i}$. Thus $\Delta_i g(\mathbf{x}) \geq \Delta_i g(\mathbf{y})$. $\square$

The proofs above are short, but they are the key to the rest of the paper. They show that the quantities computed from the data— p_i , q_i , marginal gain, and marginal gain per unit cost—are exactly the right quantities for a greedy allocation rule.

5.2 Curvature bounds from epidemic prevalence

Curvature measures how much the marginal gains can decay. A smaller curvature leads to a better approximation guarantee than the standard worst-case bound for monotone submodular maximization.

DEFINITION 7 (TOTAL CURVATURE [32]). *For a monotone DR-submodular function $h : \mathbb{Z}_+^K \rightarrow \mathbb{R}_+$ defined on the integer lattice bounded by $\mathbf{B1}$, the total curvature is*

$$\kappa := 1 - \min_{i \in [K]} \frac{h(\mathbf{e}_i \mid \mathbf{B1} - \mathbf{e}_i)}{h(\mathbf{e}_i)},$$

where $h(a \mid b) = h(a + b) - h(b)$.

Let $b_i = \lfloor B/c_i \rfloor$ be the maximum number of resources that can be placed in region i alone. For our two objectives, the box curvatures can be written as

$$\kappa_g^{\text{box}} = 1 - \min_i q_i^{b_i-1}, \quad \kappa_f^{\text{box}} = 1 - \min_i \prod_{k=1}^K q_k^{b_k-1\{k=i\}}. \quad (1)$$

For ENCD, the worst marginal of coordinate i occurs after $b_i - 1$ previous allocations to the same region. For DLM, the marginal also depends on allocations in other regions, leading to the product term. These expressions show why the maximum prevalence controls the worst-case curvature.

PROPOSITION 8 (UPPER BOUND FOR κ_f AND κ_g). *For $c_{\min} = \min_i c_i$ and $p_{\max} = \max_i p_i$, we have:*

$$\kappa_f, \kappa_g \leq 1 - (1 - p_{\max})^{B/c_{\min}}.$$

PROOF. Since $1 - p_i \geq 1 - p_{\max}$ and $b_i \leq B/c_{\min}$ for every i , every marginal-ratio term in Equation (1) is lower bounded by $(1 - p_{\max})^{B/c_{\min}}$, up to integer rounding. Substituting this lower bound into $\kappa = 1 - \min_i$ (marginal ratio) gives the claim. $\square$

The proposition reduces the curvature calculation to estimating $p_{\max}$. In our setting this is a data extraction step: mobility flows and disease parameters are summarized into a worst-region prevalence estimate. Let

$$\mathbf{M} := \text{diag} \left(\sum_j w_{j1}, \dots, \sum_j w_{jK} \right) - \mathbf{W}$$

be the mobility operator. Applying the next-generation matrix construction to the mobility-coupled SEIR system gives

$$R^* := \rho \left(\beta(\gamma \mathbf{I}_K + \mathbf{M})^{-1} \sigma(\sigma \mathbf{I}_K + \mathbf{M})^{-1} \right). \quad (2)$$

To see where this expression comes from, linearize the exposed and infected compartments around the disease-free equilibrium and write

$$\mathbf{z} = (E_1, \dots, E_K, I_1, \dots, I_K)^\top.$$

New infections and transfers are represented by

$$\mathbf{F} = \begin{bmatrix} 0 & \beta \mathbf{I}_K \\ 0 & 0 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \sigma \mathbf{I}_K + \mathbf{M} & 0 \\ -\sigma \mathbf{I}_K & \gamma \mathbf{I}_K + \mathbf{M} \end{bmatrix}.$$

The nonzero block of $\mathbf{FV}^{-1}$ is

$$\mathbf{K}_{EE} = \beta \sigma (\gamma \mathbf{I}_K + \mathbf{M})^{-1} (\sigma \mathbf{I}_K + \mathbf{M})^{-1},$$

and $R^* = \rho(\mathbf{K}_{EE})$. If there is no mobility, this reduces to the usual β/γ . With mobility, the dominant eigenvector identifies where early infection mass is likely to concentrate. These regional signals are then used by both the curvature calculation and LAMGREEDY.

THEOREM 9 (AVERAGE PEAK PREVALENCE). *The average prevalence at epidemic peak satisfies*

$$\bar{p} = \frac{4\sigma \sqrt{\Delta(R^*)}}{(\sigma + \gamma + \sqrt{\Delta(R^*)})^2} \left(1 - \frac{1}{R^*} - \frac{1}{R^*} \ln R^* \right), \quad (3)$$

where $\Delta(R^*) = (\sigma - \gamma)^2 + 4R^* \sigma \gamma$.

For allocation, the worst regional prevalence is more informative than the average, because it determines the largest initial detection probabilities. We therefore adjust the average approximation using the dominant spatial mode of the next-generation matrix.

THEOREM 10 (MAXIMUM REGIONAL PEAK PREVALENCE). *Let $\mathbf{v}$ be the dominant eigenvector of $\beta\sigma(\gamma\mathbf{I}_K + \mathbf{M})^{-1}(\sigma\mathbf{I}_K + \mathbf{M})^{-1}$ with $\|\mathbf{v}\|_\infty = 1$, and define the condition number $\kappa(\mathbf{v}) = \max_i v_i / \min_i v_i$. Let $R_{\max}^* := R^* \cdot \kappa(\mathbf{v})$. Then the worst-affected region has prevalence:*

$$\hat{p}_{\max} \approx \frac{4\sigma\sqrt{\Delta(R_{\max}^*)}}{(\sigma + \gamma + \sqrt{\Delta(R_{\max}^*)})^2} \left(1 - \frac{1}{R_{\max}^*} - \frac{1}{R_{\max}^*} \ln R_{\max}^*\right) \quad (4)$$

where $\Delta(R_{\max}^*) = (\sigma - \gamma)^2 + 4R_{\max}^*\sigma\gamma$.

This gives a fast way to estimate $p_{\max}$ from the mobility network and disease parameters. In Section 6, we check this approximation against SEIR simulations before using it to interpret the greedy guarantees.

5.3 Our approach: LATTICE-ADAPTED MODIFIED GREEDY

The previous two subsections establish the structure that an allocation algorithm should exploit: the surveillance objectives are monotone DR-submodular on an integer lattice, and their marginal decay can be summarized by curvature. The remaining question is how to turn this structure into a scalable algorithm with an approximation guarantee. There is no known curvature-dependent approximation algorithm that applies directly to our lattice knapsack formulation. Curvature-dependent guarantees are available, but only for monotone submodular *set* functions under knapsack constraints. In particular, modified greedy algorithms for set functions repeatedly select the feasible item with largest marginal gain per unit cost and then compare that greedy solution with the best single feasible item [20, 42].

We therefore need one additional step before we can use these set-function tools: we must represent a lattice allocation as a set. To this end, we leverage Ene–Nguyen reduction [14] to get a power set representation of lattice allocation. Specifically, for each region i , let $b_i = \lfloor B/c_i \rfloor$ be the largest number of resources that could be assigned to that region alone. The reduction replaces the integer coordinate $x_i \in \{0, \dots, b_i\}$ by a small collection of binary increment elements. Selecting an element means adding a fixed block of resources to region i ; selecting a subset of these elements gives a lattice vector. Thus the power set of the constructed elements becomes a compact representation of the original feasible lattice allocations.

LEMMA 11 (ENE-NGUYEN ENCODING [14]). *For each region i , there is a decomposition $b_i = \sum_{j=1}^{t_i} a_{ij}$ with $t_i \leq 2 \log b_i + 1$ such that every integer allocation $0 \leq q \leq b_i$ can be represented as the sum of a subset of $\{a_{i1}, \dots, a_{it_i}\}$. Here, a_{ij} at a given j has values given by $a_i = 2^{i-2}$.*

PROOF SKETCH. Use the binary representation of b_i and include powers of two as increment elements, splitting the largest power if needed to cover all values up to b_i . This gives only logarithmically many increments per coordinate. $\square$

Let E' be the resulting ground set, with elements (i, j) corresponding to increment size a_{ij} in region i . For a subset $S \subseteq E'$, define the induced allocation

$$x_i(S) = \sum_{j:(i,j) \in S} a_{ij},$$

and define the reduced set function $\tilde{h}(S) = h(\mathbf{x}(S))$. Next, we demonstrate why this representation is not merely a data structure: it transfers the DR-submodular lattice objective into the set-function setting where modified greedy guarantees apply.

THEOREM 12 (REDUCTION PRESERVES STRUCTURE). *For $h \in \{f, g\}$, the set function $\tilde{h}(S) = h(\mathbf{x}(S))$ is monotone submodular, has curvature at most that of h , and has the same optimum value under the induced knapsack costs.*

PROOF. Let $S \subseteq T \subseteq E'$ and let $e = (i, j) \notin T$. The encoding is monotone, so $\mathbf{x}(S) \leq \mathbf{x}(T)$. DR-submodularity of h gives

$$h(\mathbf{x}(S) + a_{ij}\mathbf{e}_i) - h(\mathbf{x}(S)) \geq h(\mathbf{x}(T) + a_{ij}\mathbf{e}_i) - h(\mathbf{x}(T)),$$

which is the diminishing returns inequality for $\tilde{h}$. Monotonicity follows because adding an increment only increases one coordinate.

For curvature, each reduced marginal is a lattice marginal along a feasible coordinate increment. The smallest reduced marginal ratio is therefore no smaller than the smallest marginal ratio defining the lattice curvature, so $\kappa_{\tilde{h}} \leq \kappa_h$. Finally, Lemma 11 ensures that every feasible integer allocation is represented by a subset of E' and every subset maps to a feasible allocation with the same objective value and induced cost. Hence the optimal values coincide. $\square$

We can now put the pieces together. First, the prevalence estimates define $h \in \{f, g\}$ and the closed-form marginal gains. Second, DR-submodularity lets the Ene–Nguyen encoding produce a monotone submodular set function $\tilde{h}$. Third, Wolsey’s modified greedy rule runs on the encoded increments using marginal gain per cost. Finally, the chosen set maps back to the original lattice allocation. We call the resulting algorithm LAMGREEDY, and it is the algorithm used in all empirical results.

Algorithm 1 LAMGREEDY

Input: Objective $h \in \{f, g\}$, costs $\mathbf{c}$, budget B

Output: Allocation $\hat{\mathbf{x}} \in \mathbb{Z}_+^K$

- 1: Compute $b_i = \lfloor B/c_i \rfloor$ for each region i .
 - 2: Construct the Ene–Nguyen increment ground set $E' = \{(i, j) : j \leq t_i\}$ and the induced set function $\tilde{h}(S) = h(\mathbf{x}(S))$.
 - 3: Run modified greedy on $\tilde{h}$: iteratively add the feasible element $e \in E' \setminus S$ maximizing $[\tilde{h}(S \cup \{e\}) - \tilde{h}(S)]/c(e)$.
 - 4: Let S_g be the density-greedy set and let S_1 be the best single feasible element.
 - 5: Choose $S^* = \arg \max\{\tilde{h}(S_g), \tilde{h}(S_1)\}$ and map it back to $\hat{\mathbf{x}} = \mathbf{x}(S^*)$.
 - 6: **return** $\hat{\mathbf{x}}$
-

LAMGREEDY is natural for surveillance. It can place several resources in a high-risk county while the marginal gain stays high, and it can move to a lower-cost county when the density becomes better there. That is the operational tradeoff between repeated sampling in one place and broader spatial coverage.

Marginal gains used in the implementation. The implementation does not need to estimate the objective by simulation while it is optimizing. Once the prevalence vector is fixed, the marginal gains are available in closed form. Suppose the reduced element

corresponds to adding a units to region i . If the current allocation is $\mathbf{x}$, then for DLM,

$$f(\mathbf{x} + a\mathbf{e}_i) - f(\mathbf{x}) = \left(\prod_{k=1}^K q_k^{x_k} \right) (1 - q_i^a),$$

and for ENCD,

$$g(\mathbf{x} + a\mathbf{e}_i) - g(\mathbf{x}) = q_i^{x_i} (1 - q_i^a).$$

The cost of the increment is ac_i , so LAMGREEDY ranks candidates by these gains divided by ac_i . The reduction is therefore not only a proof device. It supplies a compact list of feasible increments for every county, and the formulas above evaluate each candidate directly from the extracted prevalence and cost data. We maintain the product term for DLM and the coordinate terms for ENCD, so each greedy step updates only the quantities affected by the selected increment. No epidemic trajectory is simulated inside the greedy loop. Simulation and spectral calculation build the surveillance instance, and LAMGREEDY then solves the deterministic allocation problem that instance induces. This keeps the method scalable and makes every selected unit auditable through its explicit marginal value and cost.

THEOREM 13 (CURVATURE-DEPENDENT GUARANTEE FOR LAMGREEDY). *Let $\kappa_{\tilde{h}} \in (0, 1]$ be the curvature of the reduced set function $\tilde{h}$ induced by $h \in \{f, g\}$. Algorithm 1 returns an allocation $\hat{\mathbf{x}}$ satisfying*

$$h(\hat{\mathbf{x}}) \geq \frac{1}{\kappa_{\tilde{h}}} (1 - e^{-\kappa_{\tilde{h}} z}) h(\mathbf{x}^*), \quad (5)$$

where $\mathbf{x}^*$ is an optimal lattice allocation and $z \in [0, 1]$ is the unique root of

$$\frac{1}{\kappa_{\tilde{h}}} (1 - e^{-\kappa_{\tilde{h}} z}) = \frac{1 - z}{1 + \kappa_{\tilde{h}} (1 - z)}.$$

Moreover, Theorem 12 gives $\kappa_{\tilde{h}} \leq \kappa_h$, so the prevalence-based curvature bound from Proposition 8 provides a computable conservative guarantee for the original lattice problem.

PROOF SKETCH. The Ene–Nguyen encoding converts the original lattice instance into the set-function instance $\tilde{h}$ without changing the optimum value. Algorithm 1 is exactly the modified greedy algorithm applied to $\tilde{h}$ under the induced knapsack costs. The curvature-dependent modified-greedy guarantee for monotone submodular set functions therefore applies to $\tilde{h}$ [20, 42]. Mapping the returned set back through $\mathbf{x}(S)$ preserves objective value, which gives the stated guarantee for h . $\square$

5.4 Online performance certification

Worst-case ratios are conservative. After LAMGREEDY returns an allocation, we compute an instance-specific upper bound using the same marginal gains used by the algorithm [1]. Let $\hat{\mathbf{x}}$ be the LAMGREEDY solution. For each region i , define $\sigma_i := \lfloor B/c_i \rfloor$, $\Delta_i := h(\hat{\mathbf{x}} \vee \sigma_i \mathbf{e}_i) - h(\hat{\mathbf{x}})$, and $\delta_i := \Delta_i / (c_i \sigma_i)$.

THEOREM 14 (ONLINE BOUND). *Order regions such that $\delta_{s_1} \geq \delta_{s_2} \geq \dots \geq \delta_{s_{K^*}}$, and let K^* be the cutoff index where $\sum_{j=1}^{K^*-1} c_{s_j} \sigma_{s_j} \leq B < \sum_{j=1}^{K^*} c_{s_j} \sigma_{s_j}$. Then:*

$$h(\mathbf{x}^*) \leq h(\hat{\mathbf{x}}) + \sum_{j=1}^{K^*-1} \Delta_{s_j} + \frac{B - \theta}{c_{s_{K^*}} \sigma_{s_{K^*}}} \Delta_{s_{K^*}}, \quad (6)$$

where $\theta = \sum_{j=1}^{K^*-1} c_{s_j} \sigma_{s_j}$.

PROOF. Consider any feasible allocation $\mathbf{y}$. DR-submodularity implies that the value gained by moving from $\hat{\mathbf{x}}$ toward $\mathbf{y}$ is upper bounded by evaluating remaining one-coordinate improvements at $\hat{\mathbf{x}}$. For coordinate i , the largest such block improvement is

$$\Delta_i = h(\hat{\mathbf{x}} \vee \sigma_i \mathbf{e}_i) - h(\hat{\mathbf{x}}),$$

where $\sigma_i = \lfloor B/c_i \rfloor$. The cost-normalized value $\delta_i = \Delta_i / (c_i \sigma_i)$ is an upper bound on the residual density available from coordinate i .

Relax the residual problem to a fractional knapsack instance whose items are these coordinate blocks. This relaxation upper bounds any feasible integral allocation, including the optimum. The fractional knapsack solution sorts blocks by decreasing density and takes full blocks until the cutoff, plus a fraction of the cutoff block. This gives the expression in the theorem after adding $h(\hat{\mathbf{x}})$. $\square$

This certificate is useful in the experiments because it is computed from the same data products used by LAMGREEDY: prevalence estimates, costs, and residual marginal gains. It tells us how close the returned allocation is to optimal for the particular surveillance instance.

5.5 A theoretical benchmark via continuous greedy

We next state the strongest theoretical algorithm that follows from the above structure. This algorithm is useful as a benchmark and for the approximation result, but it is not the algorithm used in the experiments as it is prohibitively expensive.

Algorithm 2 Continuous Greedy for Surveillance Allocation

Input: Objective $h \in \{f, g\}$, costs $\mathbf{c}$, budget B , precision ϵ

Output: Allocation $\hat{\mathbf{x}} \in \mathbb{Z}_+^K$ with $\|\hat{\mathbf{x}}\|_c \leq B$

- 1: Compute bound on curvature κ using Proposition 8
 - 2: Use the increment encoding from Subsection 5.3 to construct $\tilde{h} : 2^E \rightarrow \mathbb{R}_+$, where $|E| = O(K \log B)$
 - 3: Apply Yoshida’s algorithm [47] to $\tilde{h}$ with knapsack constraint
 - 4: Convert solution $S \subseteq E$ back to lattice allocation $\hat{\mathbf{x}}$
 - 5: **return** $\hat{\mathbf{x}}$
-

The continuous-greedy benchmark uses the same Ene–Nguyen encoding introduced in Subsection 5.3. Theorem 12 ensures that the reduced function is a monotone submodular set function with no larger curvature and the same optimal value as the original lattice instance. Therefore Yoshida’s curvature-dependent set-function algorithm can be applied to the reduced instance and then mapped back to a surveillance allocation.

THEOREM 15 (APPROXIMATION GUARANTEE). *For both objectives $h \in \{f, g\}$ with curvature κ under knapsack constraint $\sum_i c_i x_i \leq B$, Algorithm 2 returns an allocation $\hat{\mathbf{x}}$ such that*

$$h(\hat{\mathbf{x}}) \geq \left(1 - \frac{\kappa}{e} - \epsilon\right) h(\mathbf{x}^*), \quad (7)$$

where $\mathbf{x}^*$ is the optimal solution. Its running time is

$$O\left(K^5 \log^5 B + K^4 \log^4 B \operatorname{polylog}(K) (1/\epsilon)^{\operatorname{poly}(1/\epsilon)}\right).$$

PROOF SKETCH. The reduction constructs a set function $\tilde{h}$ on $O(K \log B)$ elements that preserves feasibility, objective value, monotonicity, submodularity, and curvature up to the bound in Theorem 12. Applying Yoshida’s curvature-dependent continuous-greedy algorithm to $\tilde{h}$ and mapping the solution back to the lattice gives the stated guarantee and running time. $\square$

6 Experiments

We now evaluate our algorithm, LAMGREEDY, against baselines in various tasks. The goals of the experiments are: (i) to compare LAMGREEDY against natural baselines on surveillance instances built from mobility data, (ii) to check if the prevalence approximation used in the curvature analysis is accurate, and (iii) to compare the LAMGREEDY solutions with the online and offline bounds. All experiments were conducted on an AMD EPYC 7763 machine with 2TB memory and 8 NVIDIA A40 GPUs, each with 48GB memory. The code and data are available at <https://github.com/Mordecai2001/near-optimal-surveillance>.

Datasets and Setup: We construct county-level mobility networks for Iowa, Georgia, and Texas from two heterogeneous sources: *FRED* and *Facebook*. *FRED* (Framework for Reconstructing Epidemiological Dynamics) provides synthetic population and commuting data that is useful for influenza modeling [38]. *Facebook* provides social connectedness and co-location probability data released by Meta [2]. On each network, we simulate both COVID-19, representing a high-transmissibility setting [15], and Influenza, representing a moderate-transmissibility setting [26].

Baselines: We compare against five baselines. NAIVE INTEGER GREEDY repeatedly selects the feasible unit with the largest marginal gain until the budget is exhausted, but does not use the modified-greedy comparison between the greedy prefix and the best single item. SET GREEDY selects each region at most once using marginal gain. LOG-POPULATION HEURISTIC allocates resources according to $\log(\text{pop}_i)$. CENTRALITY HEURISTIC allocates according to eigenvector centrality in the mobility network. RANDOM BASELINE selects feasible regions uniformly at random until the budget is exhausted.

Costs: We use three cost models. In the LOG-POPULATION COST model, $c_i = \log(1 + N_i) / \log(1 + N_{\max})$. Larger counties are more expensive, but the cost grows sublinearly to reflect economies of scale in surveillance infrastructure [4]. In the DISTANCE-BASED COST model, we treat the mobility network as an effective-distance graph with edge weight $w_{uv} = 1 / (M_{uv} + \epsilon)$ and set $c_i = 1 + \beta d(h, i) / d_{\max}$, where h is the most populous hub and $\beta = 2.0$. Thus counties that are well connected by mobility flow are cheaper to reach than isolated counties. In the CENTRALITY-BASED COST model, we use betweenness centrality $C_B(i)$ and set $c_i = \max(0.5, 3.0 - 2\delta C_B(i))$ with $\delta = 2.0$. This captures the assumption that central locations already have transportation or laboratory infrastructure. These three models test whether the algorithm is robust to population, distance, and network-infrastructure costs.

From data to optimization instances: Each experiment follows the same pipeline. First, we convert county-level mobility flows into a directed weighted graph and normalize it to obtain the movement operator in the SEIR model. Second, the disease parameters and mobility graph give a prevalence vector, either through SEIR simulation or through the analytical approximation in Section 5.2.

Third, the prevalence vector p_i and the chosen cost vector c_i define the DLM and ENCD objectives. Finally, all methods are run on the same (p_i, c_i, B) instance. This keeps the comparison focused on allocation quality, rather than differences in the epidemiological inputs.

Evaluation protocol: For every state, disease, mobility source, objective, cost model, and budget, we record the final objective value returned by each method, which is the detection probability for DLM and the expected number of detected counties for ENCD. We sweep budgets from scarce-resource settings, where good allocation is most important, to larger budgets, where detection values begin to saturate. For the prevalence approximation, we compare $\hat{p}_{\max}$ with the empirical maximum prevalence from repeated SEIR simulations and report relative error.

Implementation details: All methods are evaluated on the same deterministic objective values after the prevalence vector has been extracted. This means that the optimization algorithms are not allowed to use different random epidemic realizations. For integer methods, a candidate move is an additional allocation unit or reduced-lattice increment, with cost equal to the number of units times the county cost. For set greedy, a county can be selected only once. This is a deliberate baseline that shows the loss from forcing a set representation. For LAMGREEDY and naive integer greedy, repeated allocations are allowed until the budget is exhausted. Ties are broken deterministically. This setup is close to what a public-health planner would see, namely one estimated risk snapshot, one cost model, and a fixed budget.

This separation also tells us where error originates. The validation plots and tables ask whether the analytical prevalence estimate is close to SEIR simulation, and in our experiments it is accurate enough that the dominant differences in objective value come from the allocation strategy, especially under heterogeneous costs.

E 1. Performance on Simulated Outbreaks in Real Networks:

Figure 1 shows representative Iowa results for both mobility sources. The top row corresponds to COVID-19 and the bottom row to Influenza. Within each disease, the Facebook panel and the FRED panel come from the same disease parameters, cost models, objectives, and budget grid, so the figure isolates the effect of changing the mobility data source while the optimization task stays fixed. All four panels solve both DLM and ENCD under the reported cost models. LAMGREEDY consistently gives larger objective values than the baselines, and the improvement is most visible when costs are heterogeneous.

For DLM, the advantage is strongest at small and moderate budgets, where each additional unit must be placed carefully before detection probability saturates. For ENCD, the advantage persists over a wider budget range because the objective continues to reward spatial coverage of multiple infected counties. The FRED panels also show that the pattern is not an artifact of the Facebook mobility graph: changing the mobility source shifts the relative difficulty of counties and changes the detailed budget curves, but the density-based lattice allocation still dominates the set-based and heuristic baselines. This is important because FRED and Facebook encode different mobility signals, with FRED emphasizing synthetic commuting and population movement and Facebook reflecting social connectedness and co-location data. Our additional

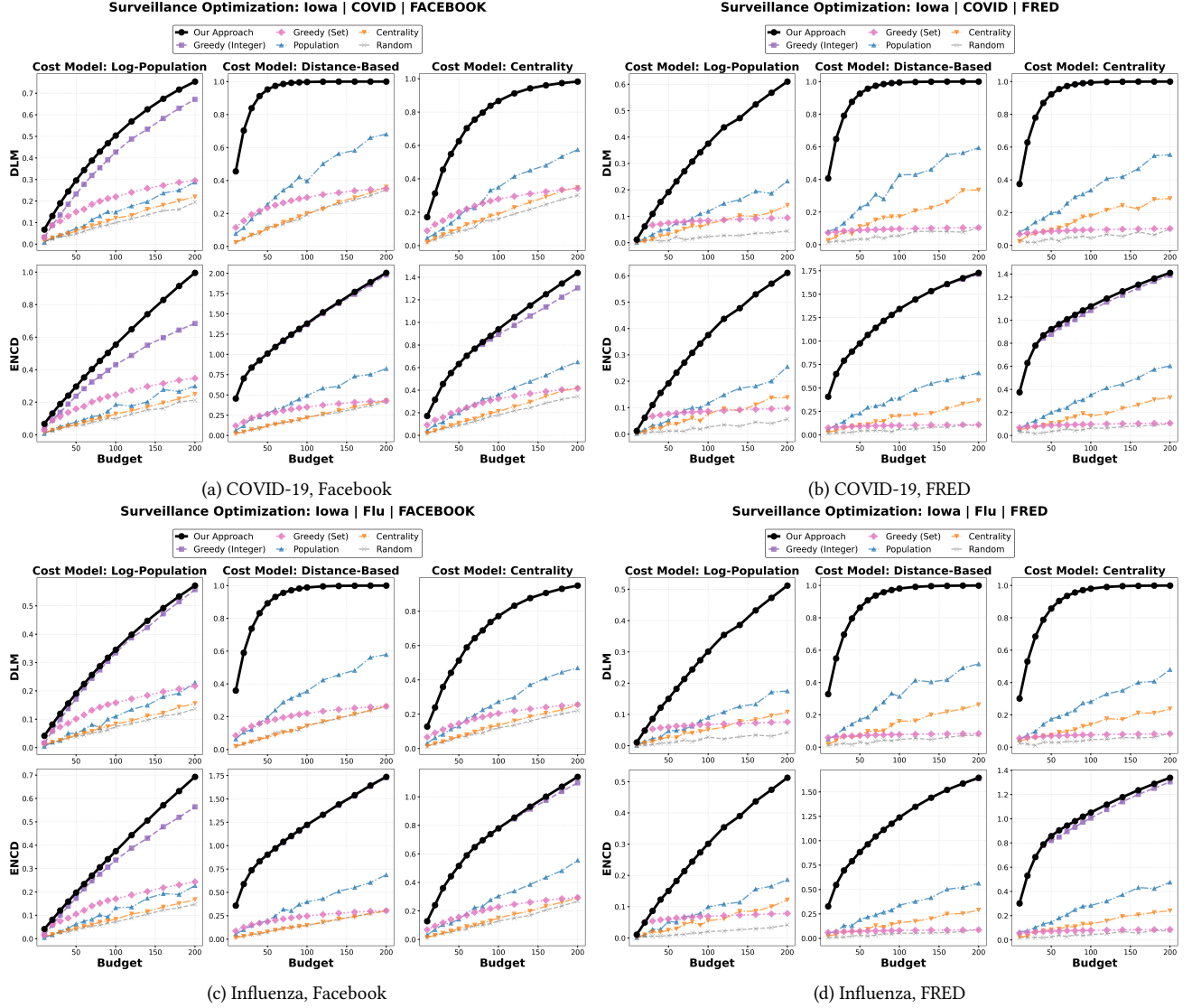


Figure 1: Performance comparison of LAMGREEDY and baselines as a function of surveillance budget in Iowa. The top row shows COVID-19 and the bottom row shows Influenza; the left column uses the Facebook mobility graph and the right column uses the FRED mobility graph.

experiments show that the same pattern appears on the Georgia and Texas networks.

E 2. Performance with Respect to Bounds: We next compare LAMGREEDY with the online and offline upper bounds. We use a synthetic instance with 10 regions, heterogeneous costs, capacities, and prevalences so that the bounds can be inspected directly. The online upper bound comes from Theorem 14 and adds to the LAMGREEDY value a fractional-knapsack relaxation over the remaining marginal gains. The offline upper bound is obtained from the worst-case approximation ratio in Theorem 13. Figure 2 shows that the LAMGREEDY solution is close to the online upper bound across budgets, while the offline bound is looser. This is expected

as the offline bound is worst-case, whereas the online bound uses the marginal gains of the actual instance.

E 3. Validation of approximated $p_{\max}$: The curvature analysis uses an analytical estimate $\hat{p}_{\max}$ of the largest regional prevalence. We therefore check whether this estimate matches simulations. We vary the reproduction number R_0 , the mobility topology, and the budget-to-cost ratio. For each setting, we compare $\hat{p}_{\max}$ with the maximum prevalence observed in stochastic SEIR simulations.

Table 1 shows that the approximation tracks the simulation values closely, with errors below 6% in the tested settings. We also validate the approximation on a real mobility network. Figure 3 uses the Iowa Facebook graph and shows that the analytical curve follows the SEIR simulation points over a range of reproduction

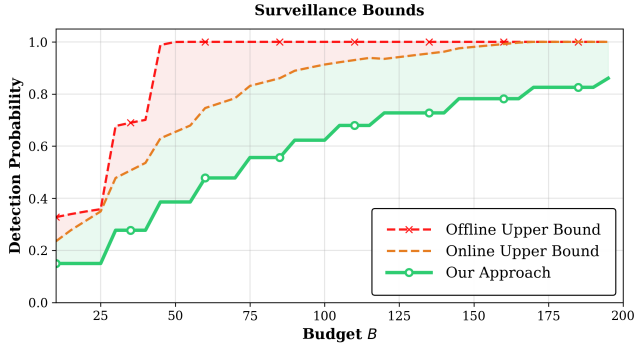


Figure 2: Comparing LAMGREEDY solution to online and offline theoretical upper bounds.

Table 1: Validation of theoretical predictions against SEIR simulations. Errors are relative to simulation means.

R_0	Topo	B/c	Max Prev $p_{\max}$	
			Approx	Sim (Err)
2.0	Mixed	2.0	0.0933	0.0942 (1.0%)
2.5	Hub	2.5	0.1391	0.1439 (3.3%)
3.0	Ring	3.0	0.1753	0.1815 (3.4%)
3.5	Hub	3.5	0.2036	0.2165 (5.9%)
4.0	Mixed	4.0	0.2259	0.2398 (5.8%)

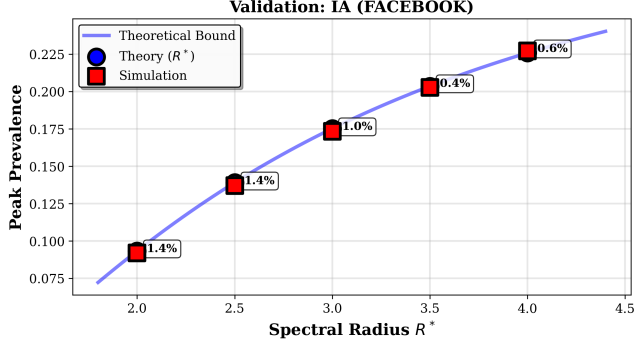


Figure 3: Validation of theoretical prevalence predictions against SEIR metapopulation simulations on real data using the Facebook mobility graph on IA.

numbers. This gives us confidence that the prevalence estimate is accurate enough to support the curvature and greedy-bound calculations.

E 4. Curvature Bounds: We finally examine the curvature values used to interpret LAMGREEDY. Once $\hat{p}_{\max}$ is available, the bounds κ_f and κ_g can be computed without running a large number of additional simulations. Table 2 reports representative settings. For small and moderate values of B/c , the curvature is noticeably below one, and the corresponding greedy guarantee is tighter than the standard $(1 - 1/e)$ bound. This is the regime that is most relevant to

Table 2: Curvature bounds and greedy guarantees. Curvature bounds κ_g, κ_f are computed from validated prevalence values.

R_0	Net Type	Bud B/c	$\hat{p}$ (val)	$p_{\max}$ (val)	κ_f, κ_g bounds	Ratio $(1 - \frac{\kappa}{e})$
2.0	Mixed	10	0.094	0.094	0.62	77%
2.0	Mixed	30	0.094	0.094	0.95	65%
2.5	Hub	10	0.144	0.144	0.78	72%
2.5	Hub	20	0.144	0.144	0.96	65%
3.0	Ring	10	0.181	0.182	0.85	69%
3.0	Ring	30	0.181	0.182	1.00	63%
3.5	Hub	10	0.216	0.217	0.89	67%

Note: The ratio column reports the curvature-dependent guarantee obtained from the listed curvature bound.

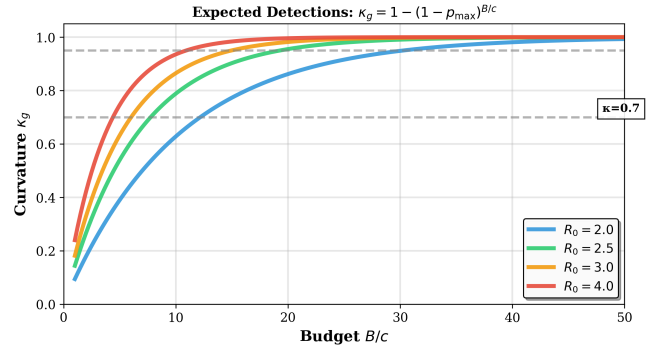


Figure 4: Curvature bound κ_g for ENCD as a function of budget.

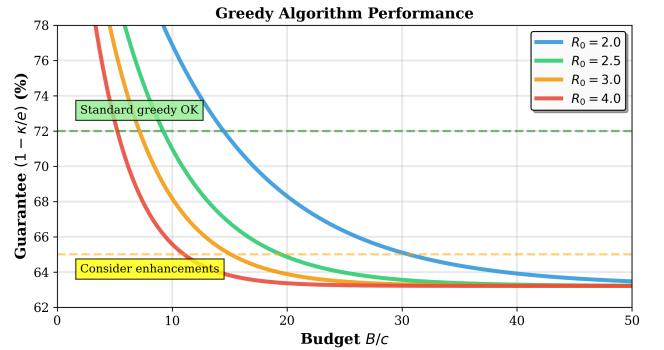


Figure 5: Curvature-dependent greedy guarantees as a function of surveillance budget.

early-stage surveillance, where resources are limited and the first allocations matter most.

Figure 4 shows the same effect as the budget grows. Curvature approaches one as the best detection opportunities are saturated. This does not mean that the algorithm fails; it means that the problem becomes closer to the standard worst case. Figure 5 shows that the guarantee smoothly approaches the classical threshold rather than collapsing.

7 Discussion and Robustness

The main empirical finding is that the lattice formulation matters. Across Iowa, Georgia, and Texas, for COVID-19 and Influenza, on Facebook and FRED mobility networks, and under all three cost models, LAMGREEDY matches or improves over the strongest integer-greedy baseline and clearly outperforms one-shot set greedy and the population, centrality, and random heuristics. The gap is largest when costs are heterogeneous, because the algorithm must decide whether to add another unit to a high-prevalence region or move to a cheaper region with slightly lower prevalence. That is the decision captured by the marginal-gain-per-cost rule, and selecting a county once cannot express it. The integer lattice also makes the allocation easier to audit, because each selected increment carries a marginal value, a cost, and a contribution to the online certificate.

The theoretical quantities are useful in practice as well. When budgets are small, a few regions have high marginal value and the curvature falls below its worst-case value, giving stronger guarantees. As the budget grows, the best opportunities are used up and the guarantee falls back toward the classical submodular bound. The online bound stays more informative because it is computed after LAMGREEDY has seen the actual instance and reflects the residual marginal gains of the returned allocation.

7.1 Modeling assumptions and deployment

Our workflow separates the epidemiological layer from the optimization layer. Mobility data, disease parameters, and any available surveillance measurements produce a prevalence vector p_i , and the optimizer then takes p_i , the cost vector c_i , and the budget B and returns an allocation. If wastewater measurements, syndromic reports, or hospital admissions update the prevalence estimates, only p_i changes and LAMGREEDY can be rerun. Updated mobility data or temporary changes in laboratory capacity are handled the same way, by recomputing the movement operator or the cost vector.

The independence assumption in the objectives is integral to our analysis, so any real-world allocation made as per our algorithm has to be independent of other allocations for our theoretical results to hold. Disease spread across counties is not assumed independent, because mobility-driven dependence is already reflected in the metapopulation SEIR model and in the prevalence estimates. The assumption is that, after the prevalence snapshot is fixed, random samples collected in different regions produce conditionally independent detection events. Deployment effects such as shared laboratories, delays, or reporting correlations can be handled by recalibrating the detection probabilities before the allocation step.

The current experiments use simple per-unit cost models to isolate the effects of population, mobility, and network centrality. Real deployments may add fixed start-up costs, staffing constraints, or equity requirements, which richer lattice increments or additional constraints can represent, though they require separate approximation analysis if the same certificates are desired. The present contribution is therefore a baseline data-to-allocation pipeline, in which prevalence extraction defines the objective, LAMGREEDY solves the lattice allocation, and the online bound certifies the returned solution.

8 Conclusion

We studied how to allocate limited surveillance resources across regions during the early stage of an epidemic. We formulated two objectives, maximizing the probability of detecting at least one infected region and maximizing the expected number of infected regions detected. Both objectives are NP-hard, and both are DR-submodular on the integer lattice. This structure is what connects epidemic data to approximation algorithms, since the prevalence estimates determine the objective, the curvature bounds, and the marginal gains used by the optimizer.

The main algorithmic contribution is LAMGREEDY, a lattice-adapted modified greedy algorithm for surveillance allocation. Continuous greedy provides a theoretical benchmark, but LAMGREEDY is the practical method. It reuses the lattice-to-set representation, the curvature estimates, and the marginal-gain quantities developed in the analysis, and it returns an allocation that can be inspected one increment at a time. On real mobility networks from Iowa, Georgia, and Texas, for both COVID-19 and Influenza, LAMGREEDY consistently outperforms natural baselines under several cost models. Mobility and disease data become prevalence estimates, those estimates define a lattice objective, LAMGREEDY produces a feasible allocation, and the online bound certifies it for early outbreak surveillance planning.

Acknowledgments

This project was supported in part by the National Science Foundation (NSF) Career Award IIS 2442159. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

GenAI Usage Disclosure

Generative AI tools were used only to edit and proofread text written by the authors, for grammar, wording, and clarity. They were not used to generate research ideas, technical content, or proofs. The authors reviewed and verified all edited text and take full responsibility for the content of this paper.

References

- [1] Bijaya Adhikari, Bryan Lewis, Anil Vullikanti, José Mauricio Jiménez, and B. Aditya Prakash. 2019. Fast and Near-Optimal Monitoring for Healthcare Acquired Infection Outbreaks. *PLOS Computational Biology* 15, 9 (2019), e1007284.
- [2] Michael Bailey, Ruiqi Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. Social Connectedness: Measurements, Determinants, and Effects. *Journal of Economic Perspectives* 32, 3 (2018), 259–280. doi:10.1257/jep.32.3.259
- [3] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe. 2008. EpiSimdemic: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*. 1–12.
- [4] Luis MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. 2007. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences* 104, 17 (2007), 7301–7306.
- [5] L. Briesemeister, P. Lincoln, and P. Porras. 2003. Epidemic profiles and defense of scale-free networks. In *Proceedings of the 2003 ACM Workshop on Rapid Malcode*. 67–75.
- [6] R. Carrico et al. 2013. Guide to preventing Clostridium difficile infections. *APIC* 16 (2013).
- [7] Alessandro Cassini et al. 2016. Burden of six healthcare-associated infections on European population health: estimating incidence-based disability-adjusted life years through a population prevalence-based modelling study. *PLoS Medicine* 13, 10 (2016), e1002150.

- [8] N. A. Christakis and J. H. Fowler. 2010. Social network sensors for early detection of contagious outbreaks. *PLoS One* 5, 9 (2010), e12948.
- [9] S. H. Cohen et al. 2010. Clinical practice guidelines for *Clostridium difficile* infection in adults: 2010 update by SHEA and IDSA. *Infection Control & Hospital Epidemiology* 31, 5 (2010), 431–455.
- [10] Michele Conforti and Gérard Cornuéjols. 1984. Submodular Set Functions, Matroids and the Greedy Algorithm: Tight Worst-Case Bounds and Some Generalizations of the Rado-Edmonds Theorem. *Discrete Applied Mathematics* 7, 3 (1984), 251–274.
- [11] S. E. Cosgrove et al. 2005. The impact of methicillin resistance in *Staphylococcus aureus* bacteremia on patient outcomes: Mortality, length of stay and hospital charges. *Infection Control and Hospital Epidemiology* (2005), 166–174.
- [12] Jiaming Cui, Bijaya Adhikari, Arash Haddadan, A. S. M. Ahsan-Ul Haque, Jilles Vreeken, Anil Vullikanti, and B. Aditya Prakash. 2025. Accurately Estimating Unreported Infections using Information Theory. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics (SIAM), Alexandria, Virginia, U.S.
- [13] O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts. 2010. The Construction of Next-Generation Matrices for Compartmental Epidemic Models. *Journal of the Royal Society Interface* 7, 47 (2010), 873–885.
- [14] Alina Ene and Huy L. Nguyen. 2019. A Reduction for Optimizing Lattice Submodular Functions with Diminishing Returns. *arXiv preprint arXiv:1606.08362* (2019).
- [15] Neil M. Ferguson, Daniel Laydon, and et al. Nedjati-Gilani. 2020. *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*. Technical Report. Imperial College London. doi:10.25561/77482
- [16] M. E. Halloran et al. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences* 105, 12 (2008), 4639–4644.
- [17] Fiona P Havers, Carrie Reed, Travis Lim, Joel M Montgomery, John D Klena, Aron J Hall, Alicia M Fry, Deborah L Cannon, Cheng-Feng Chiang, Aridith Gibbons, et al. 2020. Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23–May 12, 2020. *JAMA internal medicine* 180, 12 (2020), 1576–1586.
- [18] Jeffrey Keithley, Akash Choudhuri, Bijaya Adhikari, and Sriram V. Pemmaraju. 2025. Analyzing Greedy Vaccine Allocation Algorithms for Metapopulation Disease Models. *PLOS Computational Biology* 21, 7 (2025). Forthcoming.
- [19] R. M. Klevens et al. 2007. Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. *JAMA* 298, 15 (2007), 1763–1771.
- [20] Max Klimm and Martin Knaack. 2024. Maximizing a Submodular Function with Bounded Curvature Under an Unknown Knapsack Constraint. *arXiv preprint arXiv:2209.09668v3* (2024).
- [21] B. Leclère, D. L. Buckeridge, P.-Y. Boëlle, P. Astagneau, and D. Lepelletier. 2017. Automated detection of hospital outbreaks: A systematic review of methods. *PLoS One* 12 (2017), e0176438.
- [22] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 420–429.
- [23] B. Lewis, S. Eubank, A. M. Abrams, and K. Kleinman. 2013. In silico surveillance: evaluating outbreak detection with simulation models. *BMC Medical Informatics and Decision Making* 13 (2013), 12.
- [24] E. Lofgren, S. Cole, D. Weber, D. Anderson, and R. Moehring. 2014. Hospital-acquired *Clostridium difficile* infections: Estimating all-cause mortality and length of stay. *Epidemiology* (2014), 570–575.
- [25] E. T. Lofgren, R. W. Moehring, D. J. Anderson, D. J. Weber, and N. H. Fefferman. 2013. A mathematical model to evaluate the routine use of fecal microbiota transplantation to prevent incident and recurrent *Clostridium difficile* infection. *Infection Control and Hospital Epidemiology* (2013).
- [26] Rafael T. Mikolajczyk and Markus Kretzschmar. 2009. Influenza—insights from mathematical modelling. *Deutsche medizinische Wochenschrift* (1946) 134, 47 (November 2009), 2385–2388. doi:10.1055/s-0029-1243843
- [27] Joel C. Miller. 2012. A Note on the Derivation of Epidemic Final Sizes. *Bulletin of Mathematical Biology* 74, 9 (2012), 2125–2141.
- [28] Nicola Piovella. 2020. Analytical Solution of SEIR Model Describing the Free Spread of the COVID-19 Pandemic. *Chaos, Solitons & Fractals* 140 (2020), 110243.
- [29] B. Y. Reis, I. S. Kohane, and K. D. Mandl. 2007. An epidemiological network model for disease outbreak detection. *PLoS Medicine* 4, 6 (2007), e210.
- [30] P. Rozenstein, A. Gionis, B. A. Prakash, and J. Vreeken. 2016. Reconstructing an epidemic over time. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1835–1844.
- [31] H. Shao et al. 2018. Forecasting the flu: designing social network sensors for epidemics. In *SIGKDD epiDAMIK Workshop*.
- [32] Tasuku Soma and Yuichi Yoshida. 2015. A Generalization of Submodular Cover via the Diminishing Return Property on the Integer Lattice. In *Advances in Neural Information Processing Systems (NeurIPS 2015)*, Vol. 28. https://proceedings.neurips.cc/paper_files/paper/2015/file/7bcd75ad237b8e02e301f4091fb6bc8-Paper.pdf
- [33] Tasuku Soma and Yuichi Yoshida. 2016. Maximizing Monotone Submodular Functions over the Integer Lattice. In *Proceedings of the 18th Conference on Integer Programming and Combinatorial Optimization (IPCO 2016)*. 325–336.
- [34] P. W. Stone. 2009. Economic burden of healthcare-associated infections: an American perspective. *Expert Review of Pharmacoeconomics & Outcomes Research* 9, 5 (2009), 417–422.
- [35] R. H. Sunenshine et al. 2006. *Clostridium difficile*-associated disease: new challenges from an established pathogen. *Cleveland Clinic Journal of Medicine* 73, 2 (2006), 187.
- [36] Maxim Sviridenko. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters* 32, 1 (2004), 41–43.
- [37] Maxim Sviridenko, Jan Vondrák, and Justin Ward. 2014. Optimal approximation for submodular and supermodular optimization with curvature. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. 1324–1341.
- [38] U.S. Census Bureau. 2020. LEHD Origin-Destination Employment Statistics (LODES). <https://lehd.ces.census.gov/data/lodes/> Longitudinal Employer-Household Dynamics Program. Accessed: 2025-01-18.
- [39] P. van den Driessche and James Watmough. 2002. Reproduction Numbers and Sub-threshold Endemic Equilibria for Compartmental Models of Disease Transmission. *Mathematical Biosciences* 180, 1-2 (2002), 29–48.
- [40] E. van Kleef, J. V. Robotham, M. Jit, S. R. Deeny, and W. J. Edmunds. 2013. Modelling the transmission of healthcare associated infections: a systematic review. *BMC Infectious Diseases* 13 (2013), e1001172.
- [41] K. V. Wendelsdorf et al. 2012. Enteric immunity simulator: a tool for in silico study of gastrointestinal infections. *IEEE Transactions on Nanobioscience* 11, 3 (2012), 273–288.
- [42] Laurence A. Wolsey. 1982. Maximising Real-Valued Submodular Functions: Primal and Dual Heuristics for Location Problems. *Mathematics of Operations Research* 7, 3 (1982), 410–425.
- [43] World Health Organization. 2026. Epidemic of Ebola Disease caused by Bundibugyo virus in the Democratic Republic of the Congo and Uganda determined a public health emergency of international concern. <https://www.who.int/news/item/17-05-2026-epidemic-of-ebola-disease-in-the-democratic-republic-of-the-congo-and-uganda-determined-a-public-health-emergency-of-international-concern> WHO news item, 17 May 2026.
- [44] World Health Organization. 2026. Hantavirus cluster linked to cruise ship travel, Multi-country. <https://www.who.int/emergencies/disease-outbreak-news/item/2026-DON601> Disease Outbreak News, 13 May 2026.
- [45] J.-S. Yeom et al. 2014. Overcoming the scalability challenges of epidemic simulations on Blue Waters. In *2014 IEEE 28th International Parallel and Distributed Processing Symposium*. 755–764.
- [46] J. Yoo and A. L. Lightner. 2010. *Clostridium difficile* infections: what every clinician should know. *The Permanente Journal* 14, 2 (2010), 35.
- [47] Yuichi Yoshida. 2019. Maximizing a Monotone Submodular Function with a Bounded Curvature under a Knapsack Constraint. *SIAM Journal on Discrete Mathematics* 33, 3 (2019), 1452–1471.
- [48] Y. Zhang, A. Adiga, S. Saha, A. Vullikanti, and B. A. Prakash. 2016. Near-optimal algorithms for controlling propagation at group scale on networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (2016), 3339–3352.
- [49] Y. Zhang and B. A. Prakash. 2014. DAVA: Distributing vaccines over networks under prior information. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. 46–54.