Information Theoretic Optimal Surveillance for Epidemic Prevalence in Networks

Ritwick Mishra[1,2], **Abhijin Adiga**[1], **Madhav Marathe**[1, 2], **S. S. Ravi**[1], **Ravi Tandon**[3], **Anil Vullikanti**[1,2]

[1] Biocomplexity Institute, University of Virginia
[2] Department of Computer Science, University of Virginia
[3] Department of Electrical and Computer Engineering, University of Arizona
{mbc7bu, abhijin, marathe, ssravi,vsakumar}@virginia.edu, tandonr@arizona.edu

Abstract

Estimating the true prevalence of an epidemic outbreak is a key public health problem. This is challenging because surveillance is usually resource intensive and biased. In the network setting, prior work on cost sensitive disease surveillance has focused on choosing a subset of individuals (or nodes) to minimize objectives such as probability of outbreak detection. Such methods do not give insights into the outbreak size distribution which, despite being complex and multi-modal, is very useful in public health planning.

We introduce TESTPREV, a problem of choosing a subset of nodes which maximizes the mutual information with disease prevalence, which directly provides information about the outbreak size distribution. We show that, under the independent cascade (IC) model, solutions computed by all prior disease surveillance approaches are highly sub-optimal for TESTPREV in general. We also show that TESTPREV is hard to even approximate. While this mutual information objective is computationally challenging for general networks, we show that it can be computed efficiently for various network classes. We present a greedy strategy, called GREEDYMI, that uses estimates of mutual information from cascade simulations and thus can be applied on any network and disease model. We find that GREEDYMI does better than natural baselines in terms of maximizing the mutual information as well as reducing the expected variance in outbreak size, under the IC model.

1 Introduction

Effective surveillance in the context of disease outbreaks and contagion processes (e.g., modeled as SIR type processes on networks (Kempe, Kleinberg, and Tardos 2003)) involves monitoring or testing a subset of individuals (referred to henceforth as nodes) to determine characteristics of an outbreak; see e.g., (Leskovec et al. 2007; Christakis and Fowler 2010; Shao et al. 2018; Tsui et al. 2024; Bai et al. 2017; Heavey et al. 2022). This is also referred to as the problem of choosing "sensors" (Leskovec et al. 2007). This is particularly important from a public health perspective because testing is very expensive, and resources are limited. In networked models of disease spread, a lot of prior work on surveillance has been on choosing optimal sensor sets, so that monitoring them allows a good estimation of metrics such as probability of infections, delay in detecting the outbreak (e.g., (Leskovec et al. 2007; Heavey et al. 2022)) and time of peak (e.g., (Christakis and Fowler 2010; Shao et al. 2018)). However, none of the prior methods provides information about the disease outbreak at a distribution level.

Information theoretic strategies are a natural way to capture distribution level information, and such approaches have been developed in other settings, such as placing sensors for environmental monitoring. Using a Gaussian process based modeling approach, Caselton and Zidek (1984) proposed the use of *mutual information*[1] as the optimization criterion. Specifically, for a sensor set $A \subset V$, their goal is to maximize the mutual information $I(X_A; X_{V\setminus A})$, where X_S denotes the state vector for a subset $S \subset V$ of nodes. Krause, Singh, and Guestrin (2008) showed that this objective (referred to henceforth as the CZK objective) leads to sensor placements that are most informative regarding locations without sensors, compared to other objectives.

Krause et al. (2008) showed that finding a placement of sensors that maximizes this objective is NP-hard; they also showed that $I(X_A; X_{V\setminus A})$ is a submodular function of A, and thus can be approximated well using a greedy strategy. Such an information theoretic perspective has not been studied for surveillance of disease outbreaks, and is our focus here. Specifically, we study the following: how should we select a subset A of nodes on a network which gives the most information about the outbreak size distribution, denoted by $P(Z)$, from the results obtained by testing only the nodes in A? This is particularly useful in the case of disease outbreaks because they often exhibit threshold effects, e.g., (Marathe and Vullikanti 2013; Pastor-Satorras et al. 2015), and a distribution level understanding gives better insights into the risk of having large outbreaks. Prior surveillance methods only optimize specific epidemic metrics, such as the probability of detection or peak. Their solutions cannot provide such insights about the outbreak, as we formally show.

Our contributions. 1. We introduce a novel information theoretic criterion for optimizing surveillance for prevalence estimation. Motivated by (Caselton and Zidek 1984; Krause, Singh, and Guestrin 2008), our goal is to choose a subset

[1] Definitions of information theoretic concepts used in this paper can be found in standard texts such as Cover and Thomas (1991); MacKay (2003).

A which maximizes the mutual information $I(X_A; Z)$ with the prevalence (more generally, a weighted prevalence); we refer to this as the TESTPREV problem. We show that, in general, subsets that optimize other epidemic objectives (e.g., detection probability), do not maximize the mutual information; they can be, in fact, arbitrarily away from the optimal mutual information.

2. We show that TESTPREV is NP-hard, even to approximate within a $\Theta(\log n)$ factor, where n is the number of nodes. We also show that there are significant differences between the TESTPREV and CZK objectives. First, optimal solutions for the CZK objective can have arbitrarily low mutual information with the prevalence. Second, the CZK objective is computationally much simpler— it is submodular, and can be approximated within a $(1 - 1/e)$ factor by a greedy algorithm, while the TESTPREV objective cannot be approximated to within a $\Theta(\log n)$ factor, unless **P = NP**.

3. In general, computing the mutual information is computationally challenging since it involves an exponential sum. We show that for trees and one-hop disease models, the value of mutual information can be computed efficiently. For a simple path network, we derive a closed form expression for the optimal solution to TESTPREV. For the general setting, we present a greedy strategy, called GREEDYMI, based on estimating the mutual information from cascade samples.

4. We evaluate our method through simulations over both synthetic networks and a real contact network between patients and providers in a hospital ICU, under various disease probability regimes. We compare the surveillance sets found by our method with natural baselines to show its efficacy in finding robust solutions. We find that GREEDYMI outperforms the considered baselines and for budget as low as 2%, gives solutions with more than 60% reduction in variance about prevalence in many networks.

5. We analyze the solution sets in terms of their structural and dynamical properties. Our analysis reveals that GREEDYMI achieves a favorable balance between relevance (i.e., high information content) and redundancy (i.e., low marginal information gain), outperforming baseline methods in this trade-off.

For space reasons, proofs of some results are omitted. The full paper containing all omitted proofs is available on arXiv.

2 Problem Definition

Disease model. We consider the simplest SIR-type disease model on a network $G = (V, E)$, the Independent Cascade (IC) model. All nodes in the network except the Infected seed (s) start as Susceptible (S). At each time step, an Infected (I) node u can activate an adjacent Susceptible node v with probability $\lambda_{(u,v)}$, after which, it is Removed (R) from the process. The process continues until there are no more new infections. In a d-hop IC model, the spread is limited to a maximum of d-hops from the seed(s). Thus, the IC model is denoted by $IC(\boldsymbol{\lambda}, d)$, where the vector $\boldsymbol{\lambda}$ is composed of all the edge-wise disease probabilities λ_e and d is the number of hops; when there are no hop restrictions, we denote the model by $IC(\boldsymbol{\lambda}, \infty)$. In the homogeneous setting, where $\lambda_e = \lambda$, we use the notation $IC(\lambda, d)$ to denote such a model.

We will use upper-case letters for random variables and lower-case letters for deterministic variables; for example, X takes value x in case of scalars or $\mathbf{x}$ for vectors. We use a random variable $X_v = 1$ to indicate that node $v \in V$ is infected. The *weighted disease prevalence* (sometimes referred to as prevalence for simplicity), $Z = \sum_{i \in V} w_i X_i$, is a random variable representing the weighted sum of infections, where w_i denotes a non-negative weight associated with node $i \in V$. Let $Z_A = \sum_{i \in A} w_i X_i$ denote the weighted sum of the state variables of nodes in A, while $Z_A^- = \sum_{i \in V \setminus A} w_i X_i$ be the weighted sum of the state variables of the *remaining* nodes in the network.

Let $h(p)$ denote the binary entropy function defined by $h(p) = -p \log p - (1-p) \log(1-p)$, $p \in [0, 1]$. All the logarithms in this paper use base 2. Note that $h(p)$ represents the entropy $H(Q)$ of a random variable Q which takes on one of two values with probabilities p and $1 - p$ respectively (Cover and Thomas 1991). For two random variables P and Q, $H(P|Q)$ represents the conditional entropy of P given Q. We use the following lemma regarding conditional entropy repeatedly (proof in the supplement).

Lemma 1. *$H(Z|X_A) = H(Z_A^-|X_A)$. When the variables are all independent, $H(Z|X_A) = H(Z_A^-)$.*

The prevalence mutual information criterion. Given a set of nodes A, we define a function $\mathcal{M} : 2^V \to \mathbb{R}_{\geq 0}$ as the mutual information between X_A and prevalence Z.

$$\mathcal{M}(A) = I(X_A; Z) = \sum_{z=0}^{n} \sum_{x \in \mathcal{X}_A} p(x, z) \log \frac{p(x, z)}{p(x)p(z)} \quad (1)$$

Here, $p(x, z)$ is the joint probability mass function of X_A and Z, while $p(x)$ and $p(z)$ are their respective marginal probabilities. $\mathcal{X}_A$ represents the support of X_A. $\mathcal{M}(\cdot)$ quantifies the effect of knowing the states of a set of nodes on the prevalence distribution, which serves as our optimization criterion. In a limited budget setting, we want to query a limited subset of nodes whose effect on the cascade size distribution is the greatest among all such sets.

The TESTPREV problem. Given a network $G = (V, E)$, disease parameters $\boldsymbol{\lambda}, d$, weights w_i and costs $c_i, i \in V$, and budget k, our goal is to find a set of nodes with the maximum mutual information, i.e., $A^* \in \operatorname{argmax}_{A \subset V, \sum_{i \in A} c_i \leq k} \mathcal{M}(A)$. The weights can be used to model the relative importance of subgroups within the population.

Since $I(X_A; Z) = H(Z) - H(Z|X_A)$ (Cover and Thomas 1991), maximizing $I(X_A; Z)$ is equivalent to minimizing $H(Z|X_A)$. In some of our results, we will consider uniform weights and costs, i.e., $w_i = 1, c_i = 1$ for all $i \in V$.

Caselton-Zidek-Krause (CZK) objective. This objective denoted by $\mathcal{K}(A)$, is defined by $\mathcal{K}(A) = I(X_A; X_{V \setminus A}) = H(X_{V \setminus A}) - H(X_{V \setminus A}|X_A)$. It chooses a subset A which maximizes the reduction in conditional entropy over the rest of the nodes. This was first proposed by (Caselton and Zidek 1984) and studied extensively by (Krause, Singh, and

Guestrin 2008) for spatial processes; it has not been used for disease spread.

3 Related Work

There is considerable prior work on epidemic surveillance with diverse objectives such as to detect outbreaks (Leskovec et al. 2007; Bai et al. 2017), determine outbreak characteristics (Christakis and Fowler 2010; Shao et al. 2018). Our surveillance problem is new and is generally open in the context of epidemic processes over networks. As mentioned earlier, similar mutual information criteria have been used for problems such as observation selection (Krause and Guestrin 2007, 2012), sensor placement (Krause, Singh, and Guestrin 2008; Caselton and Zidek 1984) and feature selection (Peng, Long, and Ding 2005; Brown et al. 2012). (Tsui et al. 2024) consider an active learning framework for node subset selection, where the test feedback is used to inform the next choice.

The problem of placing sensors for detecting outbreaks has been considered in (Leskovec et al. 2007; Adhikari et al. 2019; Heavey et al. 2022). Our goal instead is to find nodes which give the most information about the prevalence. We adapt the algorithm on trees in Burkholz and Quackenbush (2021) to compute the conditional prevalence given a set of observations. The problem of estimating entropy has a long history (Paninski 2003) with more recent works (Valiant and Valiant 2011; Wu and Yang 2016) focusing on advanced estimators to achieve a near-optimal asymptotic sample complexity. Here, we use a simple plug-in estimator for implementability and low computational overhead. Wang and Ding (2019) speed-up the estimation of empirical entropy by taking a random subsample of the dataset. Here, our goal is to estimate the true entropy of the prevalence from samples.

4 Analytical Results

Complexity results for TESTPREV. In establishing our complexity results, we use the fact (from Section 2) that maximizing $I(X_A; Z)$ is equivalent to minimizing the conditional entropy $H(Z|X_A)$. Thus, we will consider the following (equivalent) version of TESTPREV: given a directed contact network $G = (V, E)$, with a non-negative weight w_i and cost c_i for each node $v_i \in V$, a probability λ_e for each directed edge $e \in E$, a non-negative weight bound $k \leq \sum_{i=1}^{n} c_i$ and a non-negative rational value R, is there a subset $A \subseteq V$ with weight $\leq k$ such that $H(Z|X_A) \leq R$? The following result establishes the computational intractability of TESTPREV.

Theorem 1. *(a)* TESTPREV *is NP-hard even when the propagation is restricted to one hop. (b) The problem remains NP-hard even if the constraint on the weight of A can be violated by a factor $(1-\epsilon)\log n$, for any $\epsilon < 1$, where $n = |V|$.*

The proof is by a reduction from the Minimum Set Cover (MSC) problem (details in the supplement).

TESTPREV vs. the CZK objective. We show here that there is a significant difference TESTPREV and the CZK objective, both in terms of the objective value (they can differ arbitrarily), and structure of optimal solutions (the TESTPREV objective is not always submodular, unlike CZK).

Observation 2. *There exist instances in which optimizing the CZK criterion $\mathcal{K}(A) = I(X_A; X_{V\setminus A})$ does not optimize for mutual information with prevalence $\mathcal{M}(A) = I(X_A; Z)$.*

We also show that $\mathcal{M}(\cdot)$ has a different structure from the CZK objective.

Observation 3. *Given a set of nodes V whose states are mutually independent random variables, the function $\mathcal{M}(A) = I(Z_V; X_A)$ is supermodular in $A \subseteq V$.*

Observation 4. *There exist instances in which the function $\mathcal{M}$ is submodular.*

TESTPREV vs. optimizing surveillance objectives. We show that prior work on optimizing epidemic objectives, e.g., detection probability, does not give good solutions to TESTPREV.

Observation 5. *There exist instances in which node selection for maximizing detection likelihood can lead to solutions with objective value for* TESTPREV *less than the optimal by $\Theta(n)$.*

All these observations are discussed in detail in the supplement.

5 Our Approach

Overview. Since the TESTPREV objective is computationally very challenging, we first study it for special classes of networks and disease models. We show that it can be computed efficiently for certain special networks and disease models. Finally, we develop a sampling-based greedy strategy for finding solutions to TESTPREV for general networks and disease models.

5.1 1-Hop Disease Spread

Consider a scenario where we are given a set of infectious individuals and we would like to know which among their immediate circle of contacts should be prioritized for testing. We model this as a directed bipartite network $G = (U \cup W, E)$ where U is the set of infected nodes and W is the set of susceptible nodes which are neighbors of U. Figure 1 has an example. The disease spreads for one timestep from U to W. We are given disease probabilities $\boldsymbol{\lambda} = \{\lambda_{ji} : (j, i) \in E\}$. The probability p_i of a node $i \in W$ becoming infected is given by $p_i = 1 - \prod_{(j,i)\in E}(1-\lambda_{ji})$. The goal of the TESTPREV problem is to find a subset of nodes A which maximizes $I(Z; X_A) = H(Z) - H(Z|X_A)$, which is equivalent to minimizing $H(Z|A)$. Due to independence of $\{X_i\}_{i\in W}$, $H(Z|X_A) = H(Z_A^-)$ by Lemma 1. Thus, choosing the best set of nodes minimizes the entropy of the sum of the remaining node states in W.

Lemma 2. *For the 1-hop disease spread scenario, the conditional entropy $H(Z|X_A)$ can be computed exactly in polynomial time.*

Proof. For any subset $A \subseteq W$, Z_A^- follows a Poisson binomial distribution with parameters $\{p_i\}_{i\in W\setminus A}$. We can compute the probability distribution of Z_A^-, using Direct Convolution method (Biscarri, Zhao, and Brunner 2018) in

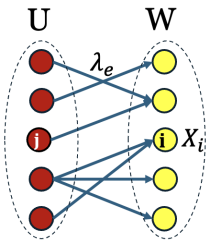

Figure 1: 1-hop example on a bipartite network

$O(|W|^2)$ time. This allows efficient computation of the entropy $H(Z_A^-) = H(Z|X_A)$. □

Greedy heuristic. Due to all the node states being independent, $\mathcal{M}(A) = I(Z; X_A)$ is supermodular from Observation 3. As $I(Z; X_A) = H(Z) - H(Z|X_A)$ and $H(Z)$ is constant, $H(Z|X_A) = H(Z_A^-)$ is submodular in A. Here, TESTPREV is minimizing a submodular function with cardinality constraints, which is generally NP-hard. Thus, in Algorithm 3 (in the supplement) we have a greedy heuristic which uses the exact computations of $H(Z|X_{A\cup\{i\}}) = H(Z^-_{A\cup\{i\}})$ (based on Lemma 2) for each node $i \in W$ in the greedy step.

Complexity. The greedy heuristic takes $O(k|W|^3)$ under the assumption that the greedy step of computing $H(Z|X_{A\cup\{i\}})$ for a node i takes $O(|W|^2)$ time.

5.2 Rooted Trees

Consider a tree network T_r of size n with a single source at the root r, under $IC(\boldsymbol{\lambda}, \infty)$. We want to find a subset $A \subseteq V$ such that $\mathcal{M}(A)$ is maximized, which is equivalent to minimizing $H(Z|X_A)$.

Computing $H(Z|X_A)$. Given a node set A, ENTROPYONTREE (Algorithm 4 in the supplement) provides an exact method to compute the conditional entropy $H(Z|X_A)$ of the prevalence on a rooted tree with the root being the source of infection. It uses the following subroutines (whose pseudocodes appear in the supplement):
1. FEASIBLE($T_r, \mathbf{x}$): filters out zero-probability infection status vectors $\mathbf{x}$.
2. CONTRACT(T_r, Γ_{rv}): contracts the path from r to v.
3. REMOVE(T_r, v): removes the node v from the tree.
4. MESSAGEPASSING($\widetilde{T}_r, \boldsymbol{\lambda}, v$): computes the unconditional prevalence distribution, $P(Z_{\widetilde{T}_r})$.

Lemma 3. *For any subset $A \subseteq T_r$,* ENTROPYONTREE *exactly computes the conditional entropy of the prevalence $H(Z|X_A)$.*

Proof sketch. We compute $P(X_A = \mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}_A$ using the probability of the live-edge paths from the (un)infected nodes. To compute $H(Z|X_A = \mathbf{x})$, we contract the live-edge paths and remove uninfected nodes to reduce the problem to computing unconditional prevalence distributions $P(Z_{\widetilde{T}})$. This is done using a message-passing algorithm from (Burkholz and Quackenbush 2021). Finally, we compute $H(Z|X_A) = \sum_{\mathbf{x}\in\mathcal{X}_A} P(X_A = \mathbf{x})H(Z|X_A = \mathbf{x})$. More details are in the supplement.

Greedy heuristic. We use a greedy heuristic in Algorithm 7 (in the supplement) which repeatedly adds nodes to the query set based on the exact computations of the conditional entropy of the prevalence using ENTROPYONTREE.

Complexity. MESSAGEPASSING has a complexity of $O(n^2)$ where n is the size of the tree. It is called a maximum of 2^k times by ENTROPYONTREE. The greedy heuristic calls ENTROPYONTREE $O(kn)$ times, resulting in a complexity of $O(k2^k n^3)$. The running time is polynomial when k is fixed.

5.3 Path Networks

We are given a path network of size $n+1$ with the node set being $V = \{0, 1, \ldots, n\}$. Assume a single source at node 0. The TESTPREV problem is to find a subset of nodes $\{i_j; i_j \in V\}$ which maximize the prevalence mutual information $\mathcal{M}$ with budget k. Equivalently, we can find the sequence of optimal separations, $\{g_1 = i_1, g_2 = i_2 - i_1, \ldots, g_k = i_k - i_{k-1}\}$, i.e., $\{g_j = i_j - i_{j-1}; j = 1, \ldots, k, g_j \geq 1\}$ where i_0 is defined to be 0. Consider the version of this problem where we remove the integrality constraint on the variables, i.e., we allow $g_j \in \mathbb{R}_{\geq 0}$. With this relaxation, we can obtain a closed form solution.

Theorem 6. *For budget k, a sufficiently long path, i.e., $n > -\log(k+1)/\log\lambda$, and $IC(\lambda, \infty)$ with homogeneous disease probability $\lambda \in (0, 1)$, the* TESTPREV*-optimal separation without integrality constraints is $\{g_j = \log(\frac{k+1-j}{k+2-j})/\log\lambda; j = 1, \ldots, k\}$.*

Proof sketch. $\mathcal{M}\{X_{i_j}\} = I(Z; \{X_{i_j}\}) = H(\{X_{i_j}\}) - H(\{X_{i_j}\}|Z) = H(\{X_{i_j}\})$ because once prevalence is given, all node states become deterministic in this path setting. $H(\{X_{i_j}\}) = h(\lambda^{g_1}) + \lambda^{g_1}h(\lambda^{g_2}) + \lambda^{g_1+g_2}h(\lambda^{g_3}) + \cdots + \lambda^{\sum_{m=1}^{k-1} g_m}h(\lambda^{g_k})$ by chain rule, where h is the binary entropy function. Using induction, we find the optimal separations in reverse-order via the first derivative test. The details are in the supplement.
Remark: Since the optimal separations $\{g_j\}$ in Lemma 6 may be fractional, our proposed simple heuristic in Algorithm 8 (in the supplement) checks the nearby integral solutions. It has a time complexity of $O(k)$.

5.4 Sampling-Based Methods

On general networks, we rely on sampling-based techniques for estimating the mutual information function $\mathcal{M}$ for any given query-set A. The sampling method first constructs a dataset D from large number of i.i.d. samples from the cascade distribution $IC(\boldsymbol{\lambda}, d)$. Then we estimate the conditional entropy of the prevalence from the observations in D, ignoring the probability of any infection vectors and sizes not seen in the dataset. The method is described in Algorithm 2.

Algorithm 1: GREEDYMI

Input: A contact network $G = (V, E)$, disease parameters $IC(\boldsymbol{\lambda}, d)$, budget k, number of cascade samples T.
Output: Query-set $A \subseteq V$.

$A \leftarrow \phi$
Sample T i.i.d. disease cascades from $IC(\boldsymbol{\lambda}, d)$
Construct a matrix D such that, $D_{i,j} = 1$ if j-th node is infected in the i-th cascade sample, 0 otherwise.
for j = 1 to k **do**
    for each $v \in V \setminus A$ **do**
        $\delta_v \leftarrow$ EMPIRICALENTROPY$(D, A \cup \{v\})$
    end for
    $v^* \leftarrow \text{argmin}_{v \in V \setminus A} \delta_v$
    $A \leftarrow A \cup \{v^*\}$
end for
return A

Algorithm 2: EMPIRICALENTROPY

Input: Dataset of cascade samples D, subset of nodes A
Output: Empirical conditional entropy of prevalence $H_D(Z|A)$

For each $\mathbf{x} \in \mathcal{X}_A$, compute the empirical probability $P_D(X_A = \mathbf{x})$, empirical conditional prevalence, $P_D(Z|X_A = \mathbf{x})$, thereby its entropy $H_D(Z|X_A = \mathbf{x})$.
Compute the empirical conditional entropy, $H_D(Z|A) = \sum_{\mathbf{x} \in \mathcal{X}_A} P_D(X_A = \mathbf{x}) H_D(Z|X_A = \mathbf{x})$
return $H_D(Z|A)$

Greedy heuristic. We provide a greedy heuristic which sequentially adds nodes to the query-set, maximizing the information gain in each step. Equivalently, the greedy step adds a node v' to current query-set A such that, $v' \in \text{argmin}_{v \in V \setminus A} H(Z|X_{A \cup \{v\}})$. GREEDYMI is described in Algorithm 1 and has the empirical entropy subroutine in Algorithm 2.

Complexity analysis. The subroutine 2 can be implemented with a hashing-based grouping of the dataset D. On average, the time complexity of EMPIRICALENTROPY is $O(kn)$. The time complexity of sampling from $IC(\boldsymbol{\lambda}, d)$ is $O(n+m)$ where m is the number of edges. This leads to a complexity of $O(T(n+m) + k^2 n^2)$ for GREEDYMI, where T denotes the number of cascade samples in Algorithm 1.

Sample complexity. The number of samples required for consistent estimation depends on the joint alphabet size of $P(Z, X_A)$, which can be as large as $n2^k$, but is usually much smaller in practice due to infeasible state configurations. With our plug-in estimator, we have a sample complexity, $O(n2^k/\epsilon^2)$ where ϵ is the desired additive error.

6 Experiments

We conduct extensive experiments using disease simulations on various network topologies, seeding scenarios, and disease transmission probability regimes to investigate the following research questions:
1. How does GREEDYMI perform compared to baseline methods in identifying good node sets for disease surveillance?
2. What are the distinguishing topological and epidemiological properties of surveillance nodes selected by GREEDYMI versus baseline methods?
3. What is the minimum number of cascade samples required for TESTPREV to converge to a stable solution?

Network	Nodes	Edges	Clust. coeff.	Avg. Shortest Path Length
PowLaw	675.3	1118.8	0.052	4.1
ER	1000	24912.0	0.049	2.03
HospICU	879	3575	0.599	4.31

Table 1: Networks and their properties. For the synthetic networks, average values are reported across 10 replicates.

Datasets and Methods. We use both synthetic and real-world networks. See Table 1 for a summary.

1. PowLaw: We construct several power-law networks using the Chung-Lu random graph model (Chung and Lu 2002). The power-law exponent is set to $\gamma = 2.5$.
2. ER: We generate several Erdős-Rényi networks using the $G(n, q)$ model with $n = 1000, q = 0.05$.
3. HospICU: This is a contact network based on the co-location of patients and healthcare providers in the ICU of a large hospital (name withheld for anonymity), built using Electronic Health Records (EHR) collected between Jan 1, 2018 and Jan 8, 2018. This network is quite relevant to the considered surveillance problem in the context of hospital acquired infections (Heavey et al. 2022; Jang et al. 2021).

In the case of synthetic networks, we use 10 replicates for each graph family. Also, in each case, we use the largest connected component.

Disease scenarios. We evaluate our methods on a range of model parameters for $IC(\lambda, d)$. In each case, we assume a homogeneous disease probability setting with $\lambda \in \{0.1, 0.2\}$ for PowLaw and HospICU, and $\lambda \in \{0.05, 0.07\}$ for ER. We set the maximum number of hops $d \in \{2, 4\}$. We evaluate our method for a budget up to $k = 10$. These regimes are chosen so that the d-hop prevalence variance is high enough for surveillance to have an effect. Each simulation instance is initiated with a single seed node. We consider two seeding scenarios,
1. Known-source: The seed node is fixed for all cascades and chosen randomly from among the nodes. We consider 10 replicates. Accordingly, we have 10 sets of cascades.
2. Random-source: For each cascade, the seed node is picked uniformly at random. In our experiments, we sample 30,000 cascades for each disease scenario.

Baselines. We compare our method against these baselines, which have been used in prior work on surveillance based on epidemic metrics (Leskovec et al. 2007; Shao et al. 2018; Marathe and Vullikanti 2013):

1. DEGREE: The top-k nodes by degree form the surveillance set.
2. VULNERABLE: In the set of sampled cascades used to compute the conditional entropy, we find the top-k most frequently infected nodes. This quantity is indirectly tied to $P(X_A)$ in the prevalence mutual information expression in Equation 1.

Evaluation metrics. For evaluating performance, we use two metrics:
1. Prevalence mutual information: Given a set of nodes A, $\mathcal{M}(A) = I(Z; X_A)$. If this is high, the chosen node subset has high mutual information with the prevalence.
2. Expected standard deviation of the conditional prevalence: If A is the subset of nodes selected by a method, this is computed as $E[\sigma(Z|X_A)] = \sum_{\mathbf{x}\in\mathcal{X}_A} P(X_A = \mathbf{x})\sigma(Z|X_A = \mathbf{x})$, where $\sigma^2(Z|X_A = \mathbf{x})$ is the variance of Z conditioned on $X_A = \mathbf{x}$. This measures the expected spread around the mean of the prevalence upon querying a node set.

6.1 Results

Performance of GREEDYMI versus baselines. Figure 2 shows the performance of GREEDYMI against baselines under the `Known-source` seeding. We report the average scores over 10 replicates. We observe that GREEDYMI consistently outperforms baselines, with the performance gap widening with increasing budget. We can observe a reduction in the expected standard deviation in prevalence ranging from 5% in `ER` to 80% in `HospICU`. We also observe that this expected standard deviation rapidly decreases with the first few node selections, followed by a more gradual decrease. This "diminishing returns" effect is especially pronounced in `PowLaw` and `HospICU` networks while it is nearly absent in `ER` networks. This is likely due to the shape of the degree distribution in each network. On `PowLaw`, DEGREE is nearly as good as GREEDYMI, which is linked to the structure of the network where there is a core of high-degree nodes. On `HospICU`, we observe that VULNERABLE is superior to DEGREE, indicating that dynamics-based selection can outperform structure-based selection. In Figure 3, we show the performance of the methods under `Random-source` seeding. For `PowLaw` and `ER`, we average the scores over 10 network replicates. We observe that GREEDYMI still performs better than the baselines, although the gap is smaller than before. In the `Known-source` scenario, the source node may have low-degree neighbors–nodes that are highly vulnerable but contribute little to the overall cascade size. This structural limitation can reduce the informativeness of such nodes. Such situations rarely arise in the `Random-source` scenario, where sources are selected randomly.

Analysis of solutions. We analyze the solutions obtained by GREEDYMI and the baselines to derive further insights into what characteristics make for a good surveillance node set. Specifically, the properties we consider for each node are the degree, vulnerability, and node-wise influence (expected size of a cascade initiated from the node). Figure 4 shows

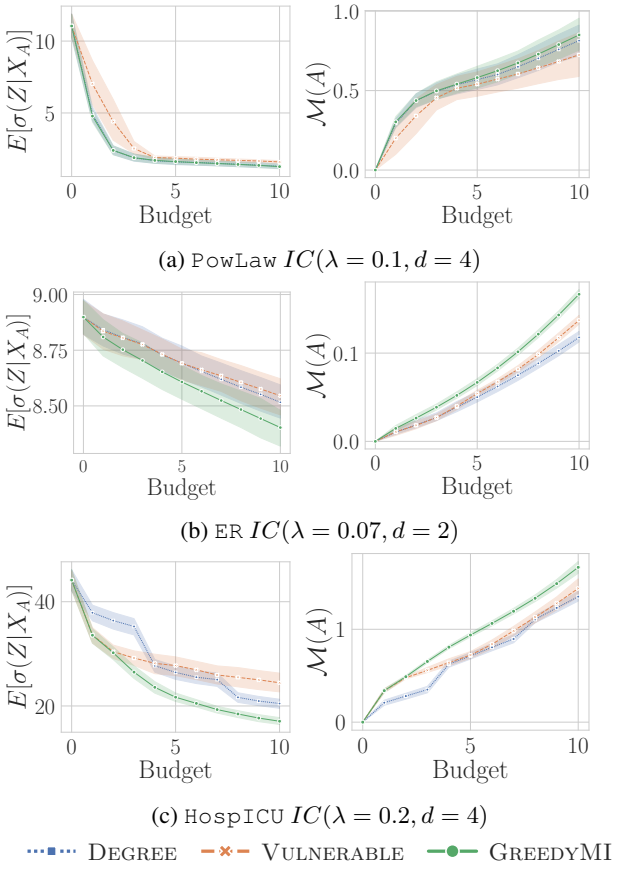

Figure 2: Performance of GREEDYMI vs baselines under `Known-source` seeding (averaged over 10 runs).

the degrees of the selected nodes for different approaches along with those of the remaining nodes. Figure 5 shows the node-wise influence and vulnerability of the subset of nodes selected by each of the methods, with a random sample of network nodes forming the backdrop. We observe that although DEGREE nodes tend to have high influence, they are not usually picked by GREEDYMI, which has more overlap with VULNERABLE node sets. Any node selection algorithm must balance *relevance* (i.e., information about prevalence) with *redundancy* (i.e., low marginal information gain). Relevance depends on how vulnerable a node is and if infected, to what extent it can influence the cascade size. Very high or very low vulnerability, or low influence, all correspond to less information. Similarly, when states of two nodes are highly correlated, it makes sense to monitor one of the nodes, thus reducing redundancy. Given the corresponding MI performance gaps, this suggests that GREEDYMI balances this tradeoff better than the top-k methods.

Sample size vs. performance. To assess the effect of sample size on the performance of GREEDYMI, we adopt the following progressive sampling-based approach. We choose 1000 samples in each round. In iteration i, we have a total of $1000 \cdot i$ samples. We find the GREEDYMI solution A_{greedy} corresponding to these samples and note its

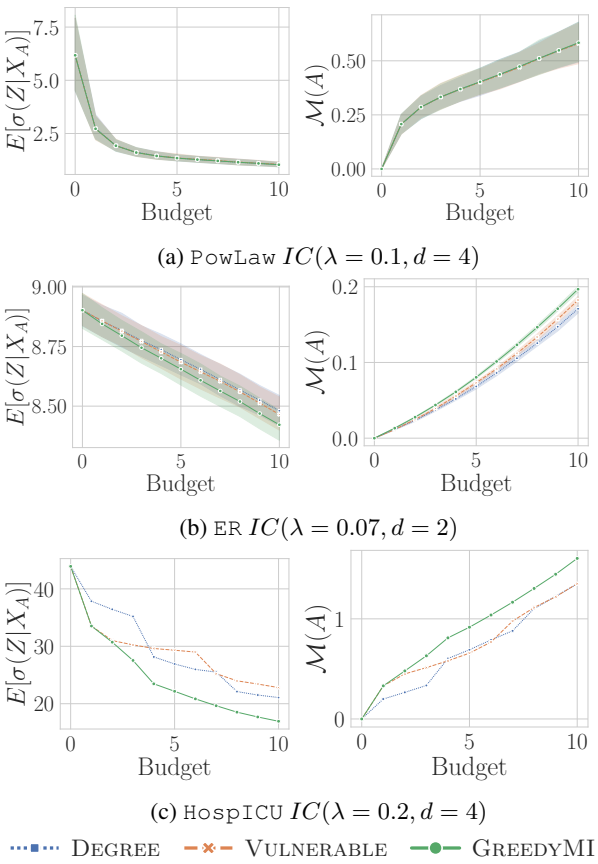

Figure 3: Performance of GreedyMI vs baselines under Random-source seeding.

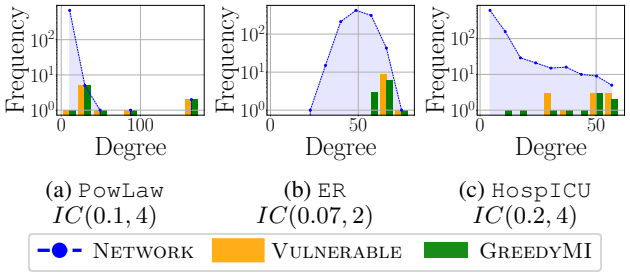

Figure 4: Comparison of the degree distributions of GreedyMI and Vulnerable in Random-source seeding (Y-axis in log-scale).

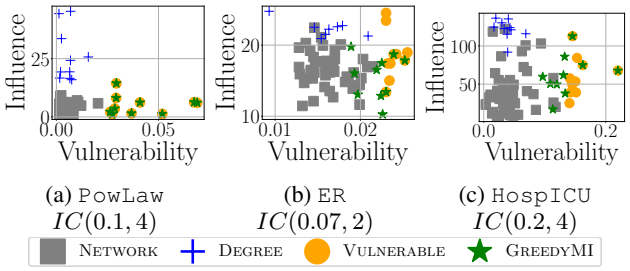

Figure 5: Vulnerability vs Influence of GreedyMI and Vulnerable in Random-source seeding. The "Network" corresponds to a sample of nodes that do not feature in any of the solution sets.

conditional entropy score, $H(Z|X_{A_{greedy}})$. Figure 6 plots the empirical entropy of the prevalence conditioned on the GreedyMI solution of budget $k = 10$, $H(Z|X_{A_{greedy}})$, computed using the combined set of samples at the end of each round of sampling. Firstly, we observe that the entropy estimate increases with the number of samples as more of the unseen joint distribution $P(Z; X_A)$ is sampled. We observe that the regimes in which large cascades are possible take more rounds of sampling to converge. In each round of sampling, even a few large cascades can push up the estimate due to the presence of the logarithm term in the entropy, triggering a new round of sampling. For a network of size $n = 1000$ and budget $k = 10$ the maximum joint alphabet size $|Z| \times |X_A| = n2^k$ can be very large. Yet, the number of cascade samples (30,000) required for the conditional entropy to converge is about 34 times fewer in size. This shows that in many realistic problem setups, sampling-based methods are not prohibitively expensive.

7 Conclusions

Our work presents an information-theoretic framework for active disease surveillance, offering – for the first time in the literature – novel, distribution-level insights into outbreak size. As our results demonstrate, solutions to TestPrev yield results with substantially lower variance than strategies such as node degree, which have been frequently proposed in the literature (Browne et al. 2024; Bai et al. 2017). At the same time, our findings underscore the challenges of adopting this approach. First, developing algorithms with provable performance guarantees remains an open problem. Unlike problems such as influence maximization, desirable properties such as submodularity may not hold, making algorithmic analysis more difficult. Second, closed-form solutions for graph families beyond paths may lead to further insights. From a practitioner's perspective, there is also a need to speed-up the greedy method by, for example, reconstruction-aware methods (Mishra et al. 2023) for estimating $P(Z|X_A)$. Finally, extending this mutual information framework to outbreak characteristics such as time to peak, spatial spread etc. offers a promising future direction.

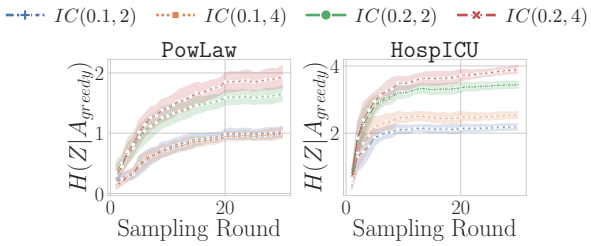

Figure 6: Convergence of the conditional entropy estimate $H(Z|A_{greedy})$.

Acknowledgments

This research is partially supported by NSF grants CCF-1918656 and CNS-2317193, and DTRA award HDTRA1-24-R-0028, Cooperative Agreement number 6NU50CK000555-03-01 from the Centers for Disease Control and Prevention (CDC) and DCLS, Network Models of Food Systems and their Application to Invasive Species Spread, grant no. 2019-67021-29933 from the USDA National Institute of Food and Agriculture.

References

Adhikari, B.; Lewis, B.; Vullikanti, A.; Jiménez, J. M.; and Prakash, B. A. 2019. Fast and near-optimal monitoring for healthcare acquired infection outbreaks. *PLoS computational biology*, 15(9): e1007284.

Bai, Y.; Yang, B.; Lin, L.; Herrera, J. L.; Du, Z.; and Holme, P. 2017. Optimizing sentinel surveillance in temporal network epidemiology. *Scientific reports*, 7(1): 4804.

Biscarri, W.; Zhao, S. D.; and Brunner, R. J. 2018. A simple and fast method for computing the Poisson binomial distribution function. *Computational Statistics & Data Analysis*, 122: 92–100.

Brown, G.; Pocock, A.; Zhao, M.-J.; and Luján, M. 2012. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, 13(1): 27–66.

Browne, A.; Buttsa, D.; Jaramillo-Rodriguez, E.; Parikh, N.; Fairchild, G.; Needell, Z.; Poliziani, C.; Wenzel, T.; Germann, T. C.; and Valle, S. D. 2024. Evaluating disease surveillance strategies for early outbreak detection in contact networks with varying community structure. *Social Networks*, 79: 122–132.

Burkholz, R.; and Quackenbush, J. 2021. Cascade size distributions: Why they matter and how to compute them efficiently. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6840–6849.

Caselton, W.; and Zidek, J. 1984. Optimal Monitoring Network Designs. *Statistics & Probability Letters*, 2(4): 223–227.

Christakis, N. A.; and Fowler, J. H. 2010. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9): e12948.

Chung, F.; and Lu, L. 2002. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25): 15879–15882.

Cover, T. M.; and Thomas, J. A. 1991. *Elements of Information Theory*. New York, NY: John Wiley and Sons, Inc.

Heavey, J.; Cui, J.; Chen, C.; Prakash, B. A.; and Vullikanti, A. 2022. Provable sensor sets for epidemic detection over networks with minimum delay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10202–10209.

Jang, H.; Pai, S.; Adhikari, B.; and Pemmaraju, S. V. 2021. Risk-aware Temporal Cascade Reconstruction to Detect Asymptomatic Cases : For the CDC MInD Healthcare Network. In *2021 IEEE International Conference on Data Mining (ICDM)*, 240–249.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '03*, 137. Washington, D.C.: ACM Press. ISBN 978-1-58113-737-8.

Krause, A.; and Guestrin, C. 2007. Near-optimal observation selection using submodular functions. In *AAAI*, volume 7, 1650–1654.

Krause, A.; and Guestrin, C. E. 2012. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*.

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *J. Mach. Learn. Res.*, 9: 235–284.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 420–429.

MacKay, D. J. 2003. *Information Theory, Inference, and Learning Algorithms*. New York, NY: Cambridge University Press.

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

Mishra, R.; Heavey, J.; Kaur, G.; Adiga, A.; and Vullikanti, A. 2023. Reconstructing an Epidemic Outbreak Using Steiner Connectivity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10): 11613–11620.

Paninski, L. 2003. Estimation of entropy and mutual information. *Neural computation*, 15(6): 1191–1253.

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

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8): 1226–1238.

Shao, H.; Hossain, K.; Wu, H.; Khan, M.; Vullikanti, A.; Prakash, B. A.; Marathe, M.; and Ramakrishnan, N. 2018. Forecasting the Flu: Designing Social Network Sensors for Epidemics. In *Proceedings of SIGKDD workshop on epidemiology meets data mining and knowledge discovery*.

Tsui, J. L.-H.; Zhang, M.; Sambaturu, P.; Busch-Moreno, S.; Suchard, M. A.; Pybus, O. G.; Flaxman, S.; Semenova, E.; and Kraemer, M. U. 2024. Toward optimal disease surveillance with graph-based active learning. *Proceedings of the National Academy of Sciences*, 121(52): e2412424121.

Valiant, G.; and Valiant, P. 2011. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, 685–694. New York, NY, USA: Association for Computing Machinery. ISBN 9781450306911.

Wang, C.; and Ding, B. 2019. Fast Approximation of Empirical Entropy via Subsampling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 658–667. Anchorage AK USA: ACM. ISBN 978-1-4503-6201-6.

Wu, Y.; and Yang, P. 2016. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6): 3702–3720.