Forecasting the Flu: Designing Social Network Sensors for Epidemics

Huijuan Shao[1,2,*], K.S.M. Tozammel Hossain[5,*], Hao Wu[2,4,†], Maleq Khan[3], Anil Vullikanti[3], B. Aditya Prakash[1,2], Madhav Marathe[3], Naren Ramakrishnan[1,2]
[1]Department of Computer Science, Virginia Tech, Arlington, VA, USA
[2]Discovery Analytics Center, Virginia Tech, Arlington, VA, USA
[3]Biocomplexity Institute, Virginia Tech, Blacksburg, VA, USA
[4]Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA
[5]Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA

ABSTRACT

Early detection and modeling of a contagious epidemic can provide important guidance about quelling the contagion, controlling its spread, or the effective design of countermeasures. A topic of recent interest has been the design of social network sensors, i.e., identifying a small set of people who can be monitored to provide insight into the emergence of an epidemic in a larger population. We formally pose the problem of designing social network sensors for flu epidemics and identify two different objectives that could be targeted in such sensor design problems. Using the graph theoretic notion of dominators we develop an efficient and effective heuristic for forecasting epidemics at lead time. Using six city-scale datasets generated by extensive microscopic epidemiological simulations involving millions of individuals, we illustrate the practical applicability of our methods and show significant benefits (up to twenty-two days more lead time) compared to other competitors.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Information systems** → *Data mining*; *Social networks*;

KEYWORDS

Social Network Sensors, Epidemic prediction, Health Informatics

ACM Reference Format:
Huijuan Shao[1,2,*], K.S.M. Tozammel Hossain[5,*], Hao Wu[2,4,†], Maleq Khan[3], and Anil Vullikanti[3], B. Aditya Prakash[1,2], Madhav Marathe[3], Naren Ramakrishnan[1,2]. 2018. Forecasting the Flu: Designing Social Network Sensors for Epidemics. In *Proceedings of ACM International Workshop on Epidemiology meets Data Mining and Konwledge Discovery (KDD epiDAMIK'18).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.475/123_4

[*]These authors contributed equally.
[†]Now working at Google Inc.

KDD epiDAMIK'18, August 2018, London, UK

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Motivated by complicated public health concerns during the initial stages of a pandemic (other than just detecting if there is an epidemic at all) [11], public health officials are usually interested in the questions: Will there be a large disease outbreak? Or, has the epidemic reached its peak? These are important questions from a public health perspective [3]; the answers can help determine if costly interventions are needed (e.g., school closures), the strategies to organize vaccination campaigns and distributions, locations to prioritize efforts to minimize new infections, the time to issue advisories, and in general how to better engineer health care responses.

Given a graph and a contagion spreading on it, can we answer such questions by monitoring some nodes to get *ahead* of the overall epidemic? A social sensor is a set of individuals selected from the population which could indicate the outbreak of the disease under consideration, thus giving an early warning. Many existing methods for such detection problems typically give indicators which lag behind the epidemic. Recent work by Christakis and Fowler [5] has made some advances. They first proposed the notion of social network sensors for monitoring flu based on the friendship paradox: your friends have more friends than you do. They proposed a so-called 'Friend-of-Friend' approach to use the set of friends nominated by the individuals randomly sampled from the population as the social sensor. After implementing it among students at Harvard, Christakis and Fowler found that the peak of the daily incidence curve (the number of new infections per day) in the sensor set occurs 3.2 days earlier than that of a same-sized random set of students.

Figures 1 and 2 depict the results of experiments we did on two large contact networks—Oregon and Miami (see Table 1 for details)—using the SEIR model. We formed the sensor set using the approach given in [5] and measured the *average lead time* of the peaks for 100 runs (hence the results are robust to stochastic fluctuations). For the Oregon dataset, Fig. 1 shows that there is a lead time of 11 days on average for the peak in the sensor set with respect to the random set (see Fig. 1(c)). In contrast, for the Miami dataset, no lead time for the sensor set is observed (see Fig. 2(c)).

There may be several possible reasons for these inconsistencies. First, the 'Friend-of-Friend' approach implicitly assumes that the lead time always increases as we add more sensors into the set. Second, the lead time observation is assumed to be independent of the underlying network topology structures, which is clearly not

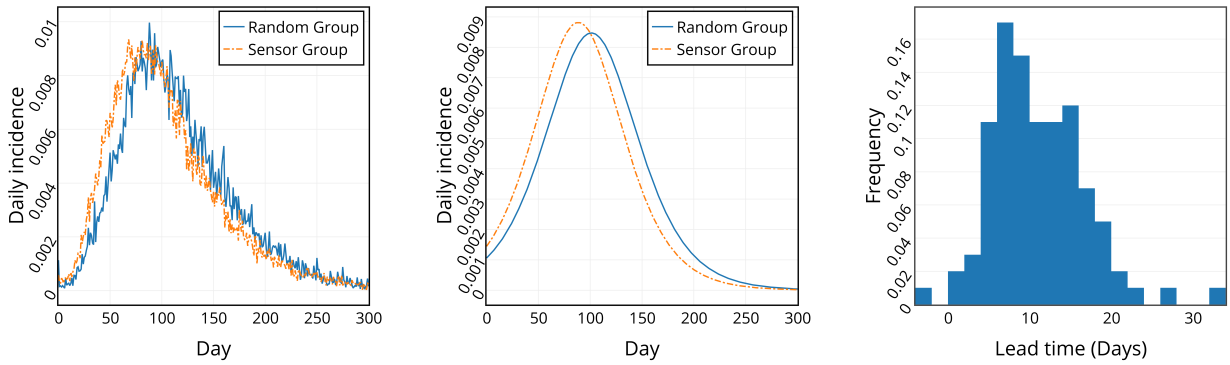

Figure 1: Illustration of the Friend-of-Friend approach [5] on the Oregon dataset. (a) True daily incidence curve (left), (b) fitted daily incidence curve with logistic function (middle), and (c) distribution of lead time over 100 experiments (right). Note that there is a non-zero lead time observed, i.e., the peak of the sensor curve occurs earlier than the peak of the curve for the random group.

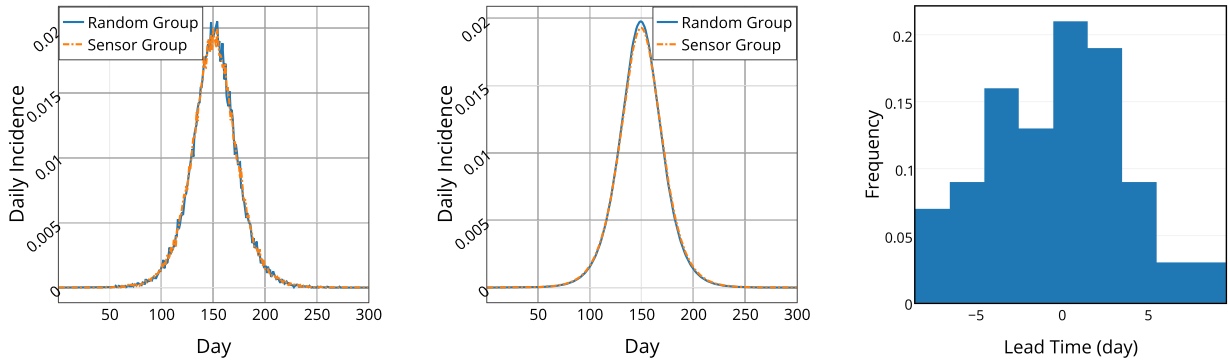

Figure 2: Illustration of the Friend-of-Friend approach on the Miami dataset. (a) True daily incidence curve (left), (b) fitted daily incidence curve with logistic function (middle), and (c) distribution of lead time over 100 experiments (right). Note that this experiment does not reveal any lead time.

the case. Finally, and most importantly, the work in [5] does not formally define the problem it is trying to solve, i.e., what objective does the sensor set optimize?

In this paper, we systematically formalize the problem of picking appropriate individuals to monitor and forecast the disease spreading over a social contact network. Our contributions are:

(1) We formally pose and study three variants of the sensor set selection problem.
(2) We provide an efficient heuristic based on the notion of graph dominators which solves one variant of the social sensor selection problem.
(3) We conduct extensive experiments on city-scale datasets based on detailed microscopic simulations, demonstrating improved lead time over competitors (including the Friend-of-Friend approach of [5]).
(4) We design surrogate/proxy social sensors using demographic information so that it is easy to deploy our approach in practice without knowledge of the full contact network.

2 EPIDEMIOLOGY FUNDAMENTALS

The most fundamental computational disease model is the so-called 'Susceptible-Infected' (SI) model where each individual (e.g. node in the disease propagation network) is considered to be in one of two states: Susceptible (healthy) or Infected. Any infected individual may infect each of its neighbors *independently* with probability β. Also, the SI model assumes every infected individual stays infected forever. For a clique of N nodes, the SI model can be characterized as:

$$\frac{dI}{dt} = \beta \times (N - I) \times I$$

where I is the number of infected nodes at time t. It is easy to prove that the solution for I is the logistic or sigmoid function, and its derivative (or the number of *new* infections per unit time) is symmetric around the peak.

The disease model that we use in this paper is the so-called SEIR model where a node in the disease propagation network is in one of *four* states: Susceptible, Exposed, Infected, and Recovered.The

dynamics of the SEIR model can be described as:

$$\frac{dS}{dt} = -\beta SI \quad \frac{dI}{dt} = \alpha E - \gamma I \quad \frac{dE}{dt} = \beta SI - \alpha E \quad \frac{dR}{dt} = \gamma I,$$

where S, E, I, and R denote the number of individuals in the corresponding states at time t, and $S + E + I + R = N$. Here β, α and γ represent the transition rates between the different states. Notice that since we are considering disease epidemics during a short period of time in this paper, we ignore the birth and death rates in the standard SEIR model here.

3 PROBLEM FORMULATION

Using the SEIR process, let $G = (V, E)$ be a social contact network where V and E represent the vertex set and edge set respectively. We use $f(S)$ to denote the probability that at least one vertex in the sensor set S gets infected, starting the disease spread from a random initial vertex. The most basic problem in such a setting is the *early detection* problem, in which the goal is to select the smallest sensor set S so that some vertices in S get infected within the first d days of the disease outbreak in the network G with probability at least ϵ (here, d and ϵ are given parameters)—this can be used to detect if there is an epidemic at all. This problem can be viewed as a special case of the detection problem in [10], and can be solved within a constant factor by a greedy submodular function maximization algorithm. As we show later, our optimization goal is *non-linear* and *not submodular*, and hence the approach in [10] can not be directly applied. Importantly, the early detection problem does not capture the more important issues about the disease characteristics of relevance to public health officials, and therefore we do not explore this further. For example, just detecting an infection in the population is generally not sufficient justification for doing an expensive intervention by public health officials (as the disease might not spread and may disappear soon). But knowing that the infection will still grow further and peak gives justification for robust infection control measures.

In our formulation, we use the term *epicurve* $I(t)$ to refer to the time series of number of infections by day. The *peak* of an epicurve is its maximum value, i.e., $\max_t I(t)$. Note that it is possible for an epicurve to have multiple peaks, but for most epidemic models in practice, the corresponding epicurves usually have a single peak. The derivative of the $I(t)$ with respect to t is called the *daily incidence* curve (number of new infections per day). The "time of peak" of the epicurve corresponding to the entire population is the time when the epicurve first reaches its peak, and is denoted by $t_{pk} = \text{argmax}_t\, I(t)$. Similarly, we use $t_{pk}(S)$ to denote the time-of-peak of the epicurve restricted only to a set S. The lead time of the epicurve peak for sensor set S compared to the entire population is then simply $t_{pk} - t_{pk}(S)$. The problem we study in this paper is:

(ϵ, k)-**Peak Lead Time Maximization (PLTM)**
Given: Parameters ϵ and k, network G, and the epidemic model
Find: A set of nodes S from G such that

$$S_{max} = \underset{S}{\text{argmax}}\, E[t_{pk} - t_{pk}(S)]$$
$$\text{s.t. } f(S) \geq \epsilon,\ |S| = k$$

Here, k is the budget, i.e. the required size of sensor set. Notice that we need the $f(S)$ constraint so that we only choose sets which have a minimum probability of capturing the epidemic—intuitively, there may be some nodes which only get infected infrequently, but the time they get infected during the disease propagation might be quite early. Such nodes are clearly not good 'sensors.'

4 PROPOSED APPROACH

Unfortunately, the peak of an epicurve is a high variance measure, making it challenging to address directly. Further, the expected lead time, $E[t_{pk} - t_{pk}(S)]$ is not non-decreasing (w.r.t. $|S|$) and non-submodular, in general. Hence we consider a different but related problem as an intermediate step. Let $t_{inf}(v)$ denote the expected infection time for node v, given that the epidemic starts at a random initial node. Then:

(ϵ, k)-**Minimum Average Infection Time (MAIT)**
Given: Parameters ϵ and k, network G, and the epidemic model
Find: A set S of nodes such that

$$S_{min} = \underset{S}{\text{argmin}} \sum_{v \in S} t_{inf}(v)/|S|$$
$$\text{s.t. } f(S) \geq \epsilon,\ |S| = k$$

Justification: In contrast to the peak, note that the *integral* of the epicurve restricted to S, normalized by $|S|$, corresponds to the *average infection time* of nodes in S, which is another useful metric for characterizing the epidemic. Further, if the epicurve has a sharp peak, which happens in most real networks and for most disease parameters, the average infection time is likely to be close to t_{pk}.

Approximating MAIT: The MAIT problem involves $f(S)$, which can be seen to be submodular, following the same arguments as in [7], and can be maximized using a greedy approach. However, the objective function — average infection time $\sum_{v \in S} t_{inf}(v)/|S|$ is non-linear as we keep adding nodes to S, which makes this problem challenging, and the standard greedy approaches for maximizing submodular functions and their extensions [8] do not work directly. In particular, we note that selecting a sensor set S which minimizes $\sum_{v \in S} t_{inf}(v)$ (with $f(S) \geq \epsilon$) might not be a good solution, since it might have a high average infection time $\sum_{v \in S} t_{inf}(v)/|S|$. We discuss below an approximation algorithm for this problem. For graph $G = (V, E)$, let $m = |E|$, $n = |V|$.

Lemma 1. *It is possible to obtain a bi-criteria approximation $S \subseteq V$ for any instance of the (ϵ, k)-MAIT problem on a graph $G = (V, E)$, given the $t_{inf}(\cdot)$ values for all nodes as input, such that $\sum_{v \in S} t_{inf}(v)$ is within a factor of two of the optimum, and $f(S) \geq c \cdot \epsilon$, for a constant c. The algorithm involves $O(n^2 \log n)$ evaluations of the function $f(\cdot)$.*

Proof. (Sketch) Let $t_{inf}(v)$ denote the expected infection time of $v \in V$, assuming the disease starts at a random initial node. Let B_{opt} be the average infection time value for the optimum; we can "guess" an estimate B' for this quantity within a factor of $1 + \delta$, by trying out powers of $(1 + \delta)^i$, for $i \leq \log n$, for any $\delta > 0$, since $B_{opt} \leq n$. We run $O(\log n)$ "phases" for each choice of B'.

Within each phase, we now consider the submodular function maximization problem to maximize $f(S)$, with two linear constraints: the first is $\sum t_{inf}(v)x(v) \leq B'k$ and $\sum_v x(v) \leq k$, where

$x(\cdot)$ denotes the characteristic vector of S. Using the result of Azar et al. [1], we get a set S such that $f(S) \geq c\mu(B')$, for a constant c, and $\sum_{v \in S} t_{inf}(v) \leq B'k$ and $|S| \leq k$, where $\mu(B')$ denotes the optimum solution corresponding to the choice of B' for this problem. If we have $|S| < k$, we add to it $k - |S|$ nodes with the minimum $t_{inf}(\cdot)$ values, which are not already in S, so that its size becomes k. Note that for the new set S, we have $\sum_{v \in S} t_{inf}(v) \leq 2B'k$, since the sum of the infection times of the nodes added to S is at most $B'k$.

Note that the resulting set S corresponds to one 'guess' of B'. We take the smallest value of B', which ensures $f(S) \geq c\epsilon$. It follows that for this solution S, we have $\sum_{v \in S} t_{inf}(v)/|S| \leq 2B_{opt}$ and $|S| = k$. The algorithm of Azar et al. [1] involves a greedy choice of a node each time; each such choice involves the evaluation of $f(S')$ for some set S', leading to $O(n^2)$ evaluations of the function $f(\cdot)$; since there are $O(\log n)$ phases, the lemma follows. □

Heuristics. Though Lemma 1 runs in polynomial time, it is quite impractical for the kinds of large graphs we study in this paper because of the need for a super-quadratic number of evaluations of $f(\cdot)$. Therefore, we consider faster heuristics for selecting sensor sets. The analysis of Lemma 1 suggests the following significantly faster greedy approach: pick nodes in non-decreasing $t_{inf}(\cdot)$ order till the resulting set S has $f(S) \geq \epsilon$. In general, this approach might not give good approximation guarantees. However, when the network has "hubs", it seems quite likely that the greedy approach will work well. However, even this approach requires repeated evaluation of $f(S)$, and can be quite slow. The class of social networks we study has the following property: nodes v which have low $t_{inf}(v)$ are usually hubs and have relatively high probability of becoming infected. This motivates the following simpler and much faster heuristic, referred to as the **Transmission tree (TT) based sensors** heuristic:

(1) generate a set $\mathcal{T} = \{T_1, \ldots, T_N\}$ of dendrograms; a dendrogram $T_i = (V_i, E_i)$ is a subgraph of $G = (V, E)$, where V_i is the set of infected nodes and an edge $(u, v) \in E$ is in E_i iff the disease is transmitted via (u, v);
(2) for each node v, compute d_v^i, which is its depth in T_i, for all i, if v gets infected in T_i;
(3) compute $t_{inf}(v)$ as the average of the d_v^i, over all the dendograms T_i, in which it gets infected;
(4) discard nodes v with $t_{inf}(v) < \epsilon_0$, where ϵ_0 is a parameter for the algorithm;
(5) order the remaining nodes $v_1, \ldots, v_{n'}$ in non-decreasing $t_{inf}(\cdot)$ order (i.e., $t_{inf}(v_1) \leq t_{inf}(v_2) \leq \ldots \leq t_{inf}(v_{n'})$)
(6) Let $S = \{v_1, \ldots, v_k\}$

We also use a faster approach based on dominator trees, which is motivated by the same greedy idea. We referred to it as the **Dominator tree (DT) based sensors** heuristic:

(1) generate dominator trees corresponding to each dendrogram;
(2) compute the average depth of each node v in the dominator trees (as in the transmission tree heuristic);
(3) discard nodes whose average depth is smaller than ϵ_0;
(4) order nodes based on their average depth in the dominator tree, and pick S to be the set of the first k nodes.

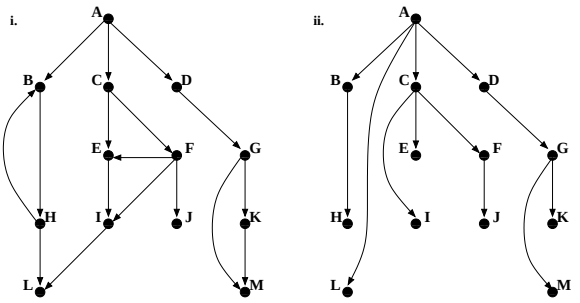

Figure 3: (i) An example graph and (ii) its dominator tree. In practice, the dominator will have a significantly reduced number of edges than the original graph.

Formally, the dominator relationship is defined as follows. A node x dominates a node y in a directed graph iff all paths from a designated start node to node y must pass through node x. In our case, the start node indicates the source of the infection or disease. Consider Fig. 3 (left), a schematic of a social contact network. All paths from node A (the designated start node) to node H must pass through node B, therefore B dominates H. Note that a person can be dominated by many other people. For instance, both C and F dominate J, and C dominates F. A node x is said to be the unique immediate dominator of y iff x dominates y and there does not exist a node z such that x dominates z and z dominates y. Note that a node can have at most one immediate dominator, but may be the immediate dominator of any number of nodes. The dominator tree $D = (V^D, E^D)$ is a tree induced from the original directed graph $G = (V^G, E^G)$, where $V^D = V^G$, but an edge $(u \rightarrow v) \in E^D$ iff u is the immediate dominator of v in G. Fig. 3 (right) shows an example dominator tree.

The computation of dominators is a well studied topic and we adopt the Lengauer-Tarjan algorithm [9] from the Boost graph library implementation. This algorithm runs in $O((|V|+|E|)\log(|V|+|E|))$ time, where $|V|$ is the number of vertices and $|E|$ is the number of edges.

5 EXPERIMENTAL RESULTS

Our experimental investigations focus on addressing the following questions:

(1) How do the proposed approaches perform when forecasting the epidemic in terms of the lead time?
(2) How large should our sensor set size be?
(3) How many days are necessary to observe a stable lead time?
(4) What is the predictive power of the sensor set in estimating the epidemic curve over the full population?
(5) Is it possible to employ surrogates for sensors?

Table 1 shows some basic network statistics of the datasets we used in our experiments. The Oregon AS (Autonomous System) router graph is an AS-level connectivity network inferred from Oregon route-views [4]. Although this dataset does not relate to epidemiological modeling, we use it primarily as a testbed to understand how (and if) graph topology affects our results due to the relatively small size and neat graph structure. The rest of the datasets are synthetic but realistic social contact networks (see [2, 6]) for six large cities

Table 1: Characteristics of datasets used in the experiments.

Dataset	Nodes	Avg. deg	Max deg
Oregon	10,670	4.12	2,312
Miami	2,092,147	50.38	425
Boston	4,149,279	108.32	437
Dallas	5,098,598	113.10	477
Chicago	9,047,574	118.83	507
Los Angeles	16,244,426	113.08	463
New York	20,618,488	93.14	464

in the United States. These six US city datasets are generated with specific aim at modeling epidemics in human populations.

In our experimental study, we evaluated our two proposed approaches: the transmission tree based heuristic and the dominator tree based heuristic. For comparison, we also implemented two strategies as baseline methods: (i) a **Top-K high degree sensors** heuristic used in [5] where a set $P \subseteq V$ is first sampled and for each $v \in P$ its K neighbors with largest degree are selected, and (ii) a **Weighted degree (WD) sensors** heuristic, which is similar to the previous heuristic except that the K neighbors are chosen based on largest weighted degree. The weight we use here is the durations of the activities indicated by edges of the graphs in the datasets mentioned in Table 1. However, since we don't have these weights for the Oregon dataset, we will omit the results of the WD sensor heuristic on the Oregon dataset.

Our primary figure of merit is the lead time, calculated as follows. For each run of the disease model in a social contact network, we fit a logistic function curve to the cumulative incidence of the chosen sensor set and a random sampled set from V. Here, we use the random sampled set to represent the entire population for the large city-level datasets we used in our experiments. (It is usually impossible to track the entire population in practice.) We then derive daily incidence curves for both the sensor set and the random set (we will refer to this set as random set in the rest of this paper). Let t_s and t_r represent the peak times of the daily incidence curves for the sensor and random sets respectively, and the lead time is defined as $\Delta t = t_r - t_s$.

For all the experiments in this section, the parameters for the epidemic simulations are set as follows unless specified. We set $\epsilon = 0.8$ (see the definitions of the PLTM and MAIT problems) and flu transmission rate to be 4.2×10^{-5} for the SEIR disease model. The size for the sensor set and random set (k) is 5% of the entire population, and the epidemic simulations start with five randomly infected vertices in the networks. All the results were obtained by averaging across $1,000$ independent runs.

5.1 Performance of predicted epidemic lead time

In this experimental study, we set the flu transmission rate to 0.05 for the SEIR model in the Oregon dataset due to its relatively small size compared to the Miami dataset. Fig. 4 depicts the daily incidence curves of the four sensor selection heuristics and the random set on Oregon and Miami datasets, and Fig. 5 describes the corresponding peak time of the daily incidence curves shown in Fig. 4. As we can see from these figures, on the Oregon dataset, the performance of the proposed heuristics and baseline heuristics is comparable where they both predict the peak of the epicurves about five days earlier when compared to the ground truth. However, on the Miami dataset, the proposed TT and DT heuristic approaches give a much larger lead time, around 10 days, compared to the about two-day and almost zero day lead time in the WD and Top-K baseline heuristics. This is because, as described earlier, our approaches are precisely designed to try to pick vertices with early expected infection time from the disease propagation network as social sensors. We also study whether the number of the initial infected vertices will affect the predicted lead time. Table 2 shows the predicted lead time of the two proposed and the two baseline heuristics for 1, 5, and 10 initial infected vertices in the epidemic simulations. As the results in this table show, the number of initial infected vertices would not have too much impact on the predicted lead time.

5.2 How many sensors to choose?

Since we have already demonstrated the influences of the network topology on social sensor selection strategies, we will put the Oregon dataset aside, and focus on the social contact network datasets for US cities in the rest of the experiments. An interesting conundrum is the number of sensors to select in a design. Fig. 6 depicts the mean lead time and the inverse of variance-to-mean ratio of the lead time v.s. the sensor size for the Miami datasets. The results show that the variance of the lead time estimate is high for small size of sensor sets and decreases as the sensor set size increases. This suggests a natural strategy of scaling the lead time against the variance, thus helps establish a sweet spot in the trade-off. This variance-to-mean ratio is also known as the *Fano factor*, which is widely used as an index of dispersion. In the result for the Miami dataset, there is a clear peak in the figure of the inverse of variance-to-mean ratio, which suggests a suitable size of sensors to pick.

5.3 Empirical study on stability of lead time

In this experiment, we study the stability of the estimated lead time as we observe more data on the sensor group when the number of monitoring days increases. As is well known, the cumulative incidence curve of flu epidemics can be modeled by a logistic function where the dependent and independent variables are the flu cumulative incidence and the time of the epidemic (days in our context). Here, we vary our flu epidemic simulation time from 2 days to 300 days on the Miami dataset, estimate cumulative incidence curves (with logistic function) for both the sensor and the random set based on the simulated cumulative flu incidence data, and then compute the lead time. Fig. 7 shows the lead time vs. the flu epidemic simulation time. As we can see from this figure, the estimated lead time fluctuates a lot when the simulation time is short and stabilizes at around 12 days when the epidemic simulation time is more than around 80 days. Such results provide some insights for public health officials on how much epidemic data they should collect in order to make an accurate estimation of the flu outbreak from the time domain perspective.

5.4 Predicting population epidemic curve from sensor group epidemic curve

In this experiment, we study the relationship between the flu cumulative incidence curve of sensor and that of random group. As

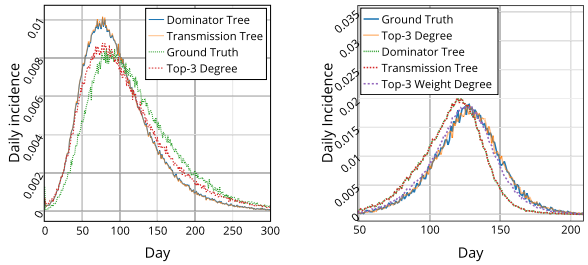

Figure 4: Daily incidence of sensor sets selected by the heuristic approaches compared to the true daily incidence in the simulated epidemic on (a) Oregon dataset (left), (b) Miami dataset (right).

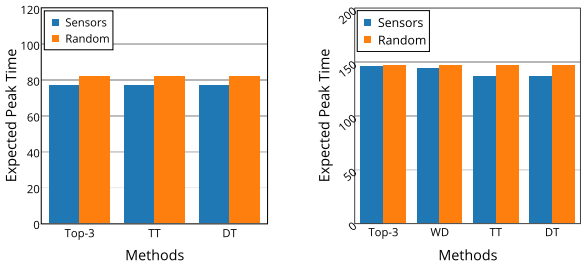

Figure 5: The expected peak time of the daily incidence curve on (a) Oregon dataset (left), (b) Miami dataset (right). Here Top-3, WD, TT, and DT denote Top-3 high degree, Top-3 weighted degree, Transmission tree based, and Dominator tree based heuristic respectively.

Table 2: Comparison of the lead time across four different social sensor selection heuristics when the number of initial infected vertices vary.

Dataset	Seed	Lead time			
		Top-K degree	Weight degree	Transmission tree	Dominator tree
Oregon	1	13.13	n/a	10.10	9.91
	5	8.85	n/a	7.93	7.75
	10	11.00	n/a	8.63	8.55
Miami	1	0.29	3.38	10.46	10.08
	5	0.39	3.41	10.15	10.19
	10	0.62	3.41	10.13	10.13

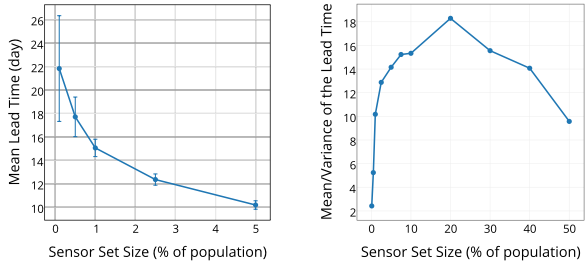

Figure 6: Mean lead time (left) and inverse of variance-to-mean ratio (right) v.s. the sensor size for the Miami dataset. When sensor set size is less than 1.0% of the entire population we observe higher (good) lead time, but also with high variances. Scaling the mean lead time by the variance, i.e., the reciprocal of the Fano factor, shows a clear peak with the sensor set size at approximately 20% of the population, the position where we can obtain substantial gains in lead time with correspondingly low variances.

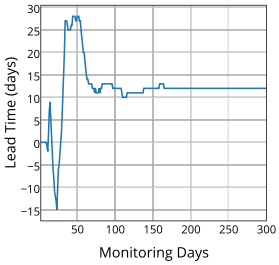

Figure 7: Stability of the lead time estimation. The estimated the lead time fluctuates initially. As the number of monitoring days increases, it stabilizes quickly.

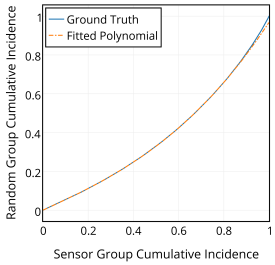

Figure 8: Predicting cumulative incidence of random group with sensor group for the Miami dataset.

we mentioned before, we use random set to represent the entire population since it is usually quite difficult to characterize the entire population in practice when the dataset is quite large. We try to estimate a polynomial regression model with degree of three where the observed cumulative incidence of the sensor group serves as predictor and that of the random group serves as responses. Here, the sensor group is selected by the dominator tree heuristic from the Miami dataset. Over the 300 simulated days, we use the data of the first 150 days to estimate our polynomial regression model, and make predictions of the cumulative incidence of random group for the rest of the 150 days. Fig. 8 shows the fitted polynomial regression model compared to the true relation curve of the flu cumulative incidences between sensor group and random group. As we can see from this figure, the polynomial regression model with degree of three could capture the relationship between the cumulative incidences of random group and sensor group quite well, which can help us predict the epidemic curve of entire population with epidemic data collected from the sensor group.

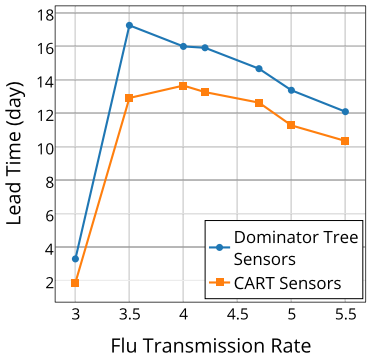

Figure 9: Mean lead times estimated with surrogate sensor set S'' and dominator tree based social sensors for various flu transmission rates.

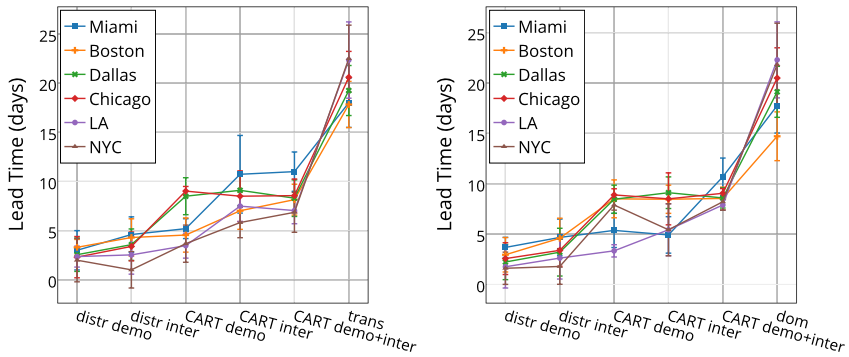

Figure 10: The lead time of transmission tree based (left) and dominator tree based (right) sensor selection strategies using different combinations of individual demographic and interaction information on Miami, Boston, Dallas, Chicago, Los Angeles and New York City datasets.

5.5 Surrogates for social sensors

In reality, the structures of large scale social contact networks are usually unknown or difficult to obtain, which makes it difficult to directly apply our proposed methods as we have done thus far. In order to make the proposed approaches deployable and solve realistic public health problems, we now relax this key assumption, and develop a *surrogate* approach to select social sensors. In this case, the policy makers can implement their strategies without detailed (and intrusive) knowledge of people and their activities. Surrogates are thus an approach to implement privacy-preserving social network sensors.

The key idea of our surrogate approach is to utilize the demographic information. Here, we use the Miami dataset as an example to explain our surrogate approach. We extracted the following 16 demographic features from the Miami dataset: age, gender, and income; number of meetings with neighbor nodes; total meeting duration with neighbor nodes; number of meetings whose durations are longer than 20000 seconds; number of meetings of types 1–5; and percent of meetings of types 1–5. The meeting types of 1–5 refer to home, work, shop, visit, and school, respectively. To select surrogate sensors using demographic information, we use classification and regression trees (CART); any other supervised classification algorithm can also be substituted here. The 16 attributes mentioned above are used as independent variables in our CART model, and the response variable is binary to indicate whether a person should be selected as a sensor or not. In order to learn the CART model, we create the training data as follows. We choose 0.1% of the entire population ($\approx$ 2000) from the US city dataset with our proposed heuristics as the training data with positive responses (social sensors), and choose another 0.1% randomly as the training data with negative responses (not social sensors). Then, separate CART models were learned to select the surrogate sensor set S' for each transmission rate ranging from 3.0×10^{-5} to 5.5×10^{-5} with a step size of 5×10^{-6}. Such transmission rates are the typical values used in various flu epidemic studies. Among all the surrogate sensors chosen by each of these CART models, we choose the common individuals across all the CART models as the final surrogate sensor set S''.

Fig. 9 compares the estimated lead time between the surrogate sensor set S'' and the sensor set selected by the dominator tree heuristic for various flu transmission rates. As we can see from this figure, although the surrogate sensor set S'' does not perform as well as the proposed dominator tree based sensor set, it still provides a significant lead time, which is good enough to give early warning to public health officials for the potential incoming flu outbreak. Most important, since the CART based surrogate sensor approach does not require the information of the social contact network structures, it is easy to implement and deploy in reality compared to the transmission tree and dominator tree based heuristic approaches. This makes it a promising candidate for predicting flu outbreaks for public health officials.

5.6 What information should be used to select surrogate sensors?

Notice that in the last section, when we select the surrogate sensors, both demographic (e.g. age of individuals) and interaction (e.g. total meeting duration and meeting types with neighboring individuals) information is taken into account. However, which kind of information is more important in terms of estimating the lead time of flu epidemics? In this experiment, we focus on all our social contact network datasets for large US cities, i.e., Miami, Boston, Dallas, Chicago, Los Angeles, and New York. For each city, we selected the surrogate sensor set and the random set with the fixed size of 10, 000. The sensor set was selected with the following six strategies: 1) using empirical distributions of demographic information (distr demo); 2) using empirical distributions of interaction information (distr inter); 3) using CART with demographic information (CART demo); 4) using CART with interaction information (CART inter); 5) using CART with both demographic and interaction information (CART demo+inter); 6) using transmission tree or dominator tree based heuristic (trans or dom). We computed the lead time for each of the six surrogate sensor selection strategies mentioned above, and the results were averaged across 100 independent runs. Fig. 10 shows the lead time of the different approaches over the six US city datasets. As we can see from the figure, our proposed approaches (CART based approaches and transmission/dominator

tree based approaches) outperform the two baseline methods (distr demo/inter), and in general, as more information is taken into account, the larger estimated lead time could be achieved (since the transmission/dominator tree based heuristics assume known social contact network structures, they could be thought of possessing the most information about epidemics). Furthermore, the individual interaction information seems to be more important than the demographic information from the perspective of obtaining larger lead time.

6 DISCUSSION

The most closely related work to ours is Christakis and Fowler [5], where a simple heuristic that monitors the friends of randomly chosen individuals from a social network as sensors was adopted to achieve early detection of epidemics. However, they only demonstrated their proposed approach on a relatively small social network, e.g. a student network from Harvard College. As we have shown earlier, their friend heuristic fails on large social contact networks of US cities. We have also demonstrated that although the Christakis and Fowler's approach works well over small networks like the Oregon dataset, it provides almost no lead time over large scale social contact networks like the Miami dataset. To explain why the proposed social sensor selection heuristics work better, we start from analyzing the structures of the disease propagation networks. Comparing the graph statistics of the Oregon dataset with the Miami dataset shown in Table 1, we can observe that the graph in the Oregon dataset has a quite different topology structure from the graphs in the Miami datasets. The graph in the Oregon dataset has relatively small average degree but very large maximum degree, which indicates this graph has a star-like topology where few of the central vertices have very large degrees. On the other hand, many vertices in the graphs of the Miami datasets have large degrees, and they spread all over the entire graph. Thus, for the top-K degree based sensor selection approach, it is relatively easy to include the central vertices with high degrees into the sensor set in the Oregon dataset, but for the transmission tree and dominator tree based approaches, whether the high degree vertices are included into the sensor set will heavily depend on the choices of initial seeds of the epidemics in the Oregon network. Such central vertices with high degree are usually very important for the epidemics in such star-like networks, which explains why the top-K degree approach works better than the transmission tree and dominator tree approaches. On the contrary, in the Miami dataset, the total number of vertices is large, and it is quite difficult for the top-K degree approach to select sensors that could represent the entire graph only based on local friend-friend information. However, the transmission tree and dominator tree based sensor selection strategies take the global epidemic spread information into account, which chooses the sensor set that could represent the entire graph. That's why they perform better in terms of the lead time than the top-K degree based approach on the large simulated US city networks. The interesting insight revealed by such results is that the network topology must be considered when designing social sensor selection strategies. The results also demonstrate that the proposed TT and DT based sensor selection heuristics are more robust to the underlying network topologies, and thus more suitable to be deployed in practice, such as monitoring and forecasting epidemics in large cities.

7 CONCLUSION

In this paper, we studied the problem of predicting flu outbreaks with social network sensors. Compared to previous works, we are the first to systematically formalize and study this problem. By leveraging the graph theoretic notion of dominators, we developed an efficient heuristic to select good social sensors to forecast the flu epidemics when the structure of flu propagation network is known. Redescription of the dominator property in terms of demographic information enables us to develop a truly implementable and deployable strategy to select surrogate social sensors to monitor and forecast flu epidemics, which will benefit public health officials and government policy makers.

ACKNOWLEDGMENTS

This work is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, by the National Science Foundation via grants DGE-1545362, IIS-1633363, and by the Army Research Laboratory under grant W911NF-17-1-0021. The US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, NSF, Army Research Laboratory, or the U.S. Government.

This paper is also based on work partially supported by the NEH (HG-229283-15), NSF CAREER (IIS-1750407), ORNL (Task Order 4000143330), and a Facebook faculty gift

REFERENCES

[1] Y. Azar and I. Gamzu. 2012. Efficient Submodular Function Maximization under Linear Packing Constraints. In *ICALP*.
[2] C. Barrett, D. Beckman, M. Khan, V.S. Anil Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. 2009. Generation and analysis of large synthetic social contact networks. In *Winter Simulation Conference*.
[3] CDC. 2012. Flu Activity During the 2012-2013 Season. http://www.cdc.gov/flu/about/season/flu-season-2012-2013.htm
[4] Q. Chen, H. Chang, R. Govindan, and S. Jamin. 2002. The Origin of Power Laws in Internet Topologies Revisited. In *INFOCOM '02*. IEEE, 608–617.
[5] N.A. Christakis and J.H. Fowler. 2010. Social Network Sensors for Early Detection of Contagious Outbreaks. *PloS one* 5, 9 (2010), e12948.
[6] S. Eubank, H. Guclu, V. S. Anil Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 180-184 (2004).
[7] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Conference of the ACM Special Interest Group on Knowledge Discovery and Data Mining*. ACM Press, New York, NY.
[8] Andreas Krause and Carlos Guestrin. 2008. Beyond Convexity - Submodularity in Machine Learning. In *ICML*. http://submodularity.org/
[9] T. Lengauer and R. Tarjan. 1979. A fast algorithm for finding dominators in a flowgraph. *ACM Trans. Program. Lang. Syst.* 1, 1 (1979), 121–141.
[10] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie S. Glance. 2007. Cost-effective outbreak detection in networks. In *KDD*. 420–429.
[11] P. Nsubuga, M. White, and S. Thacker. 2006. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. In *Disease Control Priorities in Developing Countries, World Bank*. http://www.ncbi.nlm.nih.gov/books/NBK11770/.