Maximizing the Likelihood of Detecting Outbreaks in Temporal Networks

Sterchi, Martin ; Sarasua, Cristina ; Grütter, Rolf ; Bernstein, Abraham

Abstract: Epidemic spreading occurs among animals, humans, or computers and causes substantial societal, personal, or economic losses if left undetected. Based on known temporal contact networks, we propose an outbreak detection method that identifies a small set of nodes such that the likelihood of detecting recent outbreaks is maximal. The two-step procedure involves i) simulating spreading scenarios from all possible seed configurations and ii) greedily selecting nodes for monitoring in order to maximize the detection likelihood. We find that the detection likelihood is a submodular set function for which it has been proven that greedy optimization attains at least 63% of the optimal (intractable) solution. The results show that the proposed method detects more outbreaks than benchmark methods suggested recently and is robust against badly chosen parameters. In addition, our method can be used for outbreak source detection. A limitation of this method is its heavy use of computational resources. However, for large graphs the method could be easily parallelized.

DOI: https://doi.org/10.1007/978-3-030-36683-4_39

Maximizing the Likelihood of Detecting Outbreaks in Temporal Networks

Martin Sterchi[1,2,3], Cristina Sarasua[1], Rolf Grütter[2], and Abraham Bernstein[1]

[1] University of Zurich, Zurich, Switzerland,
sterchi@ifi.uzh.ch,
[2] Swiss Federal Research Institute WSL, Birmensdorf, Switzerland,
[3] University of Applied Sciences and Arts Northwestern Switzerland FHNW, Olten, Switzerland

Abstract. Epidemic spreading occurs among animals, humans, or computers and causes substantial societal, personal, or economic losses if left undetected. Based on known temporal contact networks, we propose an outbreak detection method that identifies a small set of nodes such that the likelihood of detecting recent outbreaks is maximal. The two-step procedure involves i) simulating spreading scenarios from all possible seed configurations and ii) greedily selecting nodes for monitoring in order to maximize the detection likelihood. We find that the detection likelihood is a submodular set function for which it has been proven that greedy optimization attains at least 63% of the optimal (intractable) solution. The results show that the proposed method detects more outbreaks than benchmark methods suggested recently and is robust against badly chosen parameters. In addition, our method can be used for outbreak source detection. A limitation of this method is its heavy use of computational resources. However, for large graphs the method could be easily parallelized.

Keywords: temporal networks, epidemic spreading, outbreak detection, source detection, submodular set functions, greedy optimization

1 Introduction

Spreading processes on networks can occur in various contexts, such as infectious disease spreading among humans or animals [3, 14], computer virus spreading over the internet [13], or misinformation spreading in online social networks [5]. In those examples, the spreading process can have disastrous consequences and stopping it effectively is of great importance. For example, if the spread of a new infectious disease remains undetected, it can grow into a global pandemic. Likewise, the undetected spread of misinformation, can have significant financial or political consequences. In some cases, however, the spreading phenomenon can be seen as desirable, such as in the case of viral marketing [8]. Whether desired or harmful, the mentioned examples are conceptually similar and share the idea that a network of physical or virtual contacts acts as the substrate for

the spreading process. While much of the past research considers static contact networks, a recent surge in research focusing on temporal contact networks sheds light on the importance of temporal structure for spreading processes (e.g., [14]).

Several attempts have been made to select nodes for optimal outbreak detection. Most notably, in their seminal work, Leskovec et al. [10] propose a near-optimal outbreak detection strategy that is based on selecting a small set of nodes for monitoring using greedy optimization. Leskovec et al. focus on three optimization objectives: detection likelihood, time until detection, and the population that is affected by an outbreak. While their approach works well in a static water distribution network, where edges do not change over time, it presents a critical shortcoming when applied to a temporal blog network where bloggers post and repost stories. In that case, Leskovec et al. identify nodes for sensor placement based on past (observed) data which are then used to detect future outbreaks. However, the generalizability to future data can be unsatisfactory, especially if the topology of the underlying temporal network is changing rapidly [3, 10]. Another study suggests that central nodes tend to be infected sooner and, therefore, monitoring them may be an effective strategy [6]. Finally, Bajardi et al. [3] suggest monitoring so called sentinel nodes, i.e., nodes exhibiting both a large probability of being infected and little uncertainty about the seed node of the outbreak. The last two methods may lead to unsatisfactory results as they do not involve optimizing the set of nodes selected but are instead based on heuristics.

In this paper, we investigate the problem of outbreak detection for epidemic spreading in *temporal* contact networks. The central premise of this work is that an outbreak can start at any time and from any node within a certain time period. On a given day, we aim to identify a (small) set of k nodes for monitoring such that the probability of detecting an outbreak that started anywhere within the past b days is maximal. Since monitoring resources are typically scarce, the optimal solution of monitoring all nodes in the network is not feasible. Hence, k is set such that it matches the available monitoring resources. Our approach can be divided into two steps: i) extensive Monte-Carlo simulations of outbreak scenarios for every possible seed configuration in the window over the past b days using a propagation model and ii) greedy optimization of the detection likelihood. An optional third step extends our method to the problem of outbreak source detection (e.g. [2]). Here, we use the stochastic version of the well-known *susceptible-infected-susceptible* (SIS) model [4] as the propagation model but our approach can be used with any propagation model. Our contributions can then be summarized as follows:

- We introduce a novel method for outbreak detection in temporal networks that combines extensive outbreak simulations and greedy optimization in order to maximize the likelihood of detecting recent outbreaks (Sect. 4). We show that the method extends to the problem of source detection (Sect. 4.3).
- We show that the detection likelihood of a set of nodes is a submodular set function for which greedy optimization is guaranteed to achieve at least 63% of the optimal solution (Sect. 4.2).

– Finally, we evaluate our method on two temporal networks: an undirected network describing sexual contacts in Brazil [14] and a directed network describing pig movements in Switzerland. We show that our method outperforms previously suggested heuristics [3, 6]. Moreover, we provide evidence that it is robust against badly chosen parameters (Sect. 5).

2 Related Work

One of the early works suggesting greedy optimization of submodular set functions for optimal node selection is Kempe et al. [8], who consider the problem of maximizing the spread of influence in social networks. This work has been extended to dynamic networks by Aggarwal et al. [1] and more recently, the use of real diffusion cascades instead of simulated ones has been suggested [12]. The problem of influence maximization is conceptually similar to outbreak detection. In the case of influence maximization, the goal is to select nodes such that the size of the spreading cascades originating from those nodes is maximal, whereas for outbreak detection, the goal is to select nodes such that we catch as many spreading cascades as possible. Much of the research in the field of outbreak detection goes back to the seminal work by Leskovec et al. [10], who not only optimize *detection likelihood*, but also *detection time* and the *share of the population that is affected by an outbreak*. Our approach differs from Leskovec et al.'s approach [10] in that we optimize the set of nodes that we can monitor on a given day, whereas Leskovec et al. optimize the set of nodes without reference to a given day. As a consequence, our approach does not apply to other optimization goals, such as the detection time. The problem of outbreak detection has also been addressed in more heuristic ways. In [6], the authors make use of the observation that central individuals are topologically closer to the average individual in a network and are thus more likely to be infected early. This provides the rationale for monitoring high (in-)degree individuals. Bajardi et al. [3] propose sentinel nodes, i.e., nodes that are reached by outbreaks from many different initial conditions. However, in contrast to our work, this approach assumes that we know when an outbreak started and thus fails to address the uncertainty about the temporal origin of an outbreak.

3 Problem Definition

Imagine that we organize information about contacts between individuals as a directed or undirected time-stamped network $G = (V, E)$ where V denotes the set of nodes (individuals) and E the set of edges. An edge triple $(v_i, v_j, t) \in E$ consists of the two nodes $v_i, v_j \in V$ and a time-stamp t indicating when the contact happened. Note that in a directed network, the edge is directed from v_i to v_j. We suspect that one single node introduced a disease that now spreads along edges in the network. Furthermore, we assume that we know the underlying propagation model. Here, we use the stochastic version of the SIS model [4] with a probability p that a susceptible node gets infected upon contact with an infected

node and the time until recovery μ^{-1} that indicates the number of time steps until an infected individual recovers and becomes susceptible again. Our goal is to identify an optimal set of nodes $\mathcal{S}_t \subseteq V$ at the time of monitoring t such that the likelihood of detecting an outbreak that originated within a window over the past b days is maximal. Since monitoring resources are typically restricted, we introduce a maximal number of nodes k that can be monitored at a time. Therefore, the optimization problem becomes:

$$\max_{\mathcal{S}_t \subseteq V} DL(\mathcal{S}_t) \quad \text{subject to } |\mathcal{S}_t| \leq k \,, \tag{1}$$

where $DL(\mathcal{S}_t)$ denotes the detection likelihood associated with the set of nodes $\mathcal{S}_t$, i.e., $DL\colon \mathcal{S}_t \to [0,1]$. If we define $A(s)$ to be the event that the node $s \in \mathcal{S}_t$ detects the outbreak, we can write the detection likelihood as $DL(\mathcal{S}_t) = P(\cup_{s\in\mathcal{S}_t} A(s))$. In other words, the detection likelihood corresponds to the probability that at least one of the nodes $s \in \mathcal{S}_t$ detects an outbreak (cf. Fig. 1 for an example). We use the simplifying assumption that we detect an outbreak if at least one of the nodes $s \in \mathcal{S}_t$ is infected at the date of monitoring.

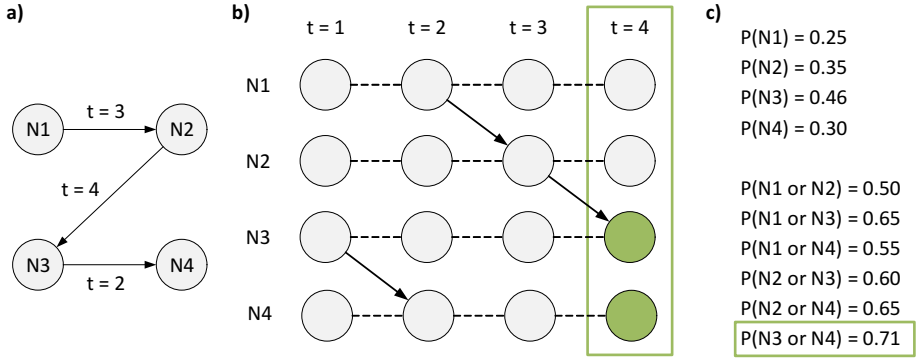

Fig. 1. a) Simple directed network with 4 nodes and 3 time-stamped edges. **b)** The same network but now time-unrolled. The goal is to identify the optimal 2 nodes for monitoring at $t = 4$. We assume a disease spreads as a *susceptible–infectious* (SI) process (individuals do not recover) with $p = 0.6$. The outbreak can happen anywhere within the time period $t = 1, 2, 3$ ($b = 3$). Assuming that all seed configurations are equally likely to be the source, we get the detection probabilities in **c)**. The node with the highest probability to "see" an infection at $t = 4$ is N3. The optimal 2 nodes for outbreak detection are N3 and N4.

4 Proposed Solution

Our solution first simulates outbreaks from every possible seed configuration (cf. Sect. 4.1) and then uses the simulation results to optimally allocate monitoring

resources to nodes (cf. Sect. 4.2). Optionally, a last step identifies the source of the outbreak (cf. Sect. 4.3). Here, we discuss these steps in turn.

4.1 Step 1: Spreading Simulations

The first step of the solution consists of simulating outbreaks from every possible seed node v_0 at every possible starting time t_0 within the window of length b. For the simulation, we use the known parameters of the underlying propagation process. As mentioned before, here, we focus on the SIS model with infection probability p and recovery time μ^{-1}. As a result, we get a large number n of outbreak outcomes for every possible pair $\{v_0, t_0\}$, which can be used to approximate the probability that a certain seed configuration $\{v_0, t_0\}$ infects at least one of the nodes in $\mathcal{S}_t$, i.e., $P(\cup_{s\in\mathcal{S}_t} A(s) \mid \{v_0, t_0\})$. In particular, we simply count the number of simulation runs that infect at least one $s \in \mathcal{S}_t$ and divide it by n. Note that for the sake of notation, we do not condition on the given parameters of the SIS process. By marginalizing out the seed configurations, we can compute the unconditional probability of detecting an outbreak, i.e.,

$$\begin{aligned} DL(\mathcal{S}_t) &= P(\cup_{s\in\mathcal{S}_t} A(s)) \\ &= \sum_{\{v_0,t_0\}} P(\cup_{s\in\mathcal{S}_t} A(s) \mid \{v_0, t_0\})\; P(\{v_0, t_0\})\;. \end{aligned} \tag{2}$$

The running time for the simulation step is $\mathcal{O}(n\, b\, |V|)$ assuming the individual simulation has cost 1. With $|V|$ given by the network and the length of the window b determined by epidemic characteristics, we need to choose the number of simulations n such that the procedure is computationally feasible and results in satisfactory approximations of the conditional probabilities. In order to ease the greedy optimization step (Sect. 4.2), we transform the simulation results into an inverted index, a data structure that facilitates fast lookups [10]. Also note that if certain outbreak configurations $\{v_0, t_0\}$ are more likely than others, we can adjust the prior $P(\{v_0, t_0\})$ accordingly. By default, we use a uniform prior.

4.2 Step 2: Optimal Selection of Nodes for Monitoring

As stated in Sect. 3, we aim to identify a set of k nodes such that the detection likelihood is maximal. Solving such an optimization is NP-hard since the number of possible sets $\mathcal{S}_t$ is prohibitively large [9]. For example, finding the optimal 10 nodes in a small network of 50 nodes, would require evaluating over 10 billion different sets of nodes. However, we can make use of a famous result by Nemhauser and Wolsey [9, 11], which states that for non-negative monotone and submodular set functions, a set selected by greedy optimization is within 63% of the optimal but intractable solution.

Monotonicity of a set function F is defined as follows. For any two sets of nodes $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A} \subseteq \mathcal{B} \subseteq V$, it holds that $F(\mathcal{A}) \leq F(\mathcal{B})$. Moreover, F is submodular if it satisfies

$$F(\mathcal{A} \cup \{v\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{v\}) - F(\mathcal{B})\;. \tag{3}$$

for $v \in V \setminus \mathcal{B}$. Intuitively, this means that adding a new node v to a smaller set $\mathcal{A}$ results in a marginal gain at least as large as the one resulting from adding v to $\mathcal{B}$.

Theorem 1. *The set function $DL(\mathcal{S}_t)$ is monotone and submodular.*

Proof. As a probability, $DL(\mathcal{S}_t)$ is also a measure and thus monotone [15]. For submodularity, we first note that a non-negative linear combination of submodular set functions is submodular [9]. From (2), we can see that $DL(\mathcal{S}_t)$ corresponds to a linear combination of probabilities $P(\cup_{s\in\mathcal{S}_t} A(s) \mid \{v_0, t_0\})$ with weights $P(\{v_0, t_0\}) \geq 0$. It thus remains to show that the probability $P(\cup_{s\in\mathcal{S}_t} A(s) \mid \{v_0, t_0\})$ is submodular. As mentioned in Sect. 4.1, this probability is estimated by counting how many of the n simulation runs (or cascades) infect at least one $s \in \mathcal{S}_t$. For the sake of notation, we will denote $P(\cup_{s\in\mathcal{S}_t} A(s) \mid \{v_0, t_0\})$ as F. Consider two sets of nodes $\mathcal{A} \subseteq \mathcal{B} \subseteq V$ and a node $v \in V \setminus \mathcal{B}$. We can distinguish three cases: i) the new node v either detects no new cascades or only detects cascades that are already detected by $\mathcal{A}$ and $\mathcal{B}$. In that case, $F(\mathcal{A}\cup\{v\}) - F(\mathcal{A}) = 0 = F(\mathcal{B}\cup\{v\}) - F(\mathcal{B})$. ii) the new node v detects cascades that $\mathcal{B}$ detects, but not $\mathcal{A}$. Then, $F(\mathcal{A}\cup\{v\}) - F(\mathcal{A}) = F(\{v\}) \geq 0 = F(\mathcal{B}\cup\{v\}) - F(\mathcal{B})$. iii) the new node v detects cascades that neither $\mathcal{B}$ nor $\mathcal{A}$ detects. In that case, $F(\mathcal{A}\cup\{v\}) - F(\mathcal{A}) = F(\{v\}) = F(\mathcal{B}\cup\{v\}) - F(\mathcal{B})$. Hence, in all possible cases the inequality in (3) is satisfied. This concludes the proof. □

Based on Theorem 1, we can use the greedy optimization algorithm that simply consists of iteratively adding the node that maximizes the marginal increase in detection likelihood to the initially empty set $\mathcal{S}_t$. The optimal set is found when $|\mathcal{S}_t| = k$. The complexity of the greedy optimization procedure is $\mathcal{O}(k|V|)$ if the evaluation of the set function is considered a constant operation. Note that we can improve the scalability of the algorithm with lazy forward evaluations that have been suggested in [10].

4.3 Step 3 (optional): Outbreak Source Detection

Based on the simulation results from Sect. 4.1 and a set of observed infected nodes, we can compute the source probability for every seed configuration. If we denote the set of observed infected nodes at time t as $\mathcal{I}_t$ and we want to compute the probability that $\{v_0^*, t_0^*\}$ was the source of this outbreak, we can use Bayes' rule as follows,

$$P(\{v_0^*, t_0^*\} \mid \mathcal{I}_t) = \frac{P(\mathcal{I}_t \mid \{v_0^*, t_0^*\})\, P(\{v_0^*, t_0^*\})}{\sum_{\{v_0,t_0\}} P(\mathcal{I}_t \mid \{v_0, t_0\})\, P(\{v_0, t_0\})} \,. \tag{4}$$

$P(\mathcal{I}_t \mid \{v_0, t_0\})$ can be easily approximated by the relative number of simulation runs that infect all nodes in $\mathcal{I}_t$ for a given seed configuration $\{v_0, t_0\}$. Typically, we would assume a uniform prior $P(\{v_0, t_0\})$, i.e., all seed configurations are equally likely a priori. However, it is possible to have different priors for different

seed configurations. For example, we may know that more recent outbreaks are more likely and, therefore, we set higher prior probabilities for more recent seed configurations. Or, in case of an animal transport network, we surmise that bigger farms are more likely to start an outbreak than smaller ones.

5 Evaluation

The primary goal of our evaluation is to show that our method is more effective than previously suggested methods and that it can be applied in practical settings where, for example, wrong parameters are chosen for the propagation model. We operationalize these claims with the following hypotheses:

- **H1.** Our method detects a higher fraction of outbreaks than state-of-the-art methods. We specifically compare with methods that are based on central nodes [6] or sentinel nodes [3], or that use past contact data to detect future outbreaks [10].
- **H2.** Our method is robust against badly chosen parameters for the spreading model.

In the remainder of this section, we introduce the datasets used in the evaluation and then discuss the empirical results for each hypothesis.

5.1 Datasets

We evaluate our method with two different datasets. The first dataset —widely used in work on temporal contact networks— describes time-stamped sexual contacts between male sex-buyers and female sex-sellers [14] and thus corresponds to an undirected bipartite network. We assume that this network acts as the substrate for a spread of sexually transmitted diseases (STDs). The full dataset covers a period of 6 years between September 2002 and October 2008. However, in what follows, we only show the results for the last 30 days of the dataset (October 2008) such that $b = 30$. The results for other 30-day periods are similar.[1] The 30-day network we analyze consists of 1,573 nodes and 1,463 edges. Ignoring the temporal dimension, we find that the network consists of 257 connected components and is thus highly fragmented. The density is 0.1131%. We refer to this network as the *escort network*.

The second dataset describes pig movements in Switzerland [16].[2] Movements are directed and time-stamped with the day of transport. Movements to slaughterhouses are removed and we only consider the farm-to-farm network. Moreover, as in [3], we ignore within-farm dynamics and simply consider a movement as

[1] We provide the dataset, the code, and additional results and figures under the following link: `https://github.com/martinSter/Outbreak-Detection`.

[2] The pig movement data contain private information and cannot be shared publicly. For research purposes, a data request can be sent to Identitas AG, Stauffacherstrasse 130A, 3014 Bern, Switzerland.

a directed contact between two farms. As with the first dataset, we consider a window of $b = 30$ days (October 2017) since this typically corresponds to the silent spread phase, i.e., the time between introduction of a disease and its first detection [7]. This 30-day network consists of 3,055 nodes 4,048 edges. Compared to the first dataset, the density is 0.0360% and thus lower which is partly due to the directionality of the second network. We refer to this network as the *pig network*.

5.2 Hypothesis H1 (Comparison with Benchmarks)

Benchmark methods. In order to test *H1*, we use two benchmark methods that have been suggested in the related work [3, 6] and we additionally use a random set of nodes as a baseline. The first benchmark method corresponds to selecting nodes based on their *(in-)degree*. As stated in [6], we expect more central nodes, i.e., nodes with a high (in-)degree, to be good indicators for detecting outbreaks. The second benchmark method [3] identifies so called *sentinels*, i.e., nodes that are reached by many different infection paths and that exhibit a low degree of uncertainty with respect to the seed cluster. The method consists of two key steps. First, a deterministic SIS process with infection probability $p = 1$ and recovery time μ^{-1} yields all possible infection paths starting from any node at some time t_0. In a second step, all seed nodes with similar infection paths, measured by the Jaccard index (threshold of 0.8 as in [3]), are grouped into seed clusters. We adjust their approach to our setting by using a SIS instead of a SIR model. We expect the method of Bajardi et al. [3] to perform worse than our method for three reasons: i) their method only considers outbreaks starting at the beginning of the considered period and thus fails to address the fact that an outbreak can start at any time, ii) their method trades some of the outbreak detection power for lower uncertainty about the seed of the outbreak, and iii) their method is heuristic and does not maximize an objective function. Since their method does not uniquely focus on outbreak detection, we adjust it slightly and define sentinels as the nodes that are reached most often by infection paths, hereby neglecting the uncertainty criteria.

Methodology. In order to test our method, we stochastically simulate 10,000 outbreak scenarios with a random seed node and starting time within the 30 days considered. We assume that the disease propagates according to a SIS model with known infection probability $p = 0.6$ and recovery time $\mu^{-1} = 15$ days. The same parameter values are used in Step 1 of our method to simulate $n = 1{,}000$ spreading simulations from every seed configuration $\{v_0, t_0\}$. Since both networks have a temporal resolution of one day, the order of contacts of an individual on a given day is not known and, therefore, we assume that there is no same-day-spread from individuals with multiple contacts on a given day. With the greedy optimization procedure outlined in Step 2 of our method, we find k optimal nodes for monitoring. To compare the different methods, we count the number of detected outbreaks for varying k. Our method is implemented in Python 3.6 and is executed on a Intel Core i7 machine (1.80GHz) with 16GB

RAM. The execution time is about 22 minutes for both the escort and pig network. Note however, that this may vary depending on the parametrization of the propagation model.

Results. Figure 2 presents the fraction of detected outbreaks for both networks. It is apparent that our method (*greedy max.*) performs better compared to the different benchmark methods in both data sets (of different sizes) and coincides with the expected detection likelihood (*dashed line*), which can be computed by plugging the k optimal nodes into (2). If k is small, high (in-) degree nodes achieve a similar detection quality because our approach initially selects mostly high-degree nodes. For larger k, the curves diverge as the greedy selection approach allocates monitoring resources to more remote parts of the network in order to maximize detection likelihood. As expected, sentinel nodes detect substantially fewer outbreaks but still perform better than the random node set. However, note that even with $k = 100$ our approach results in a detection likeli-

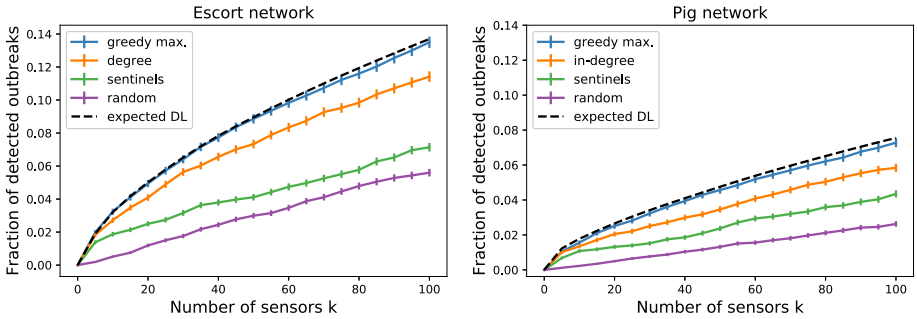

Fig. 2. Fraction of detected outbreaks for the escort network (**left**) and the pig network (**right**). The error bars represent ± one standard deviation. The disease propagates according to a SIS model with $p = 0.6$ and $\mu^{-1} = 15$ days.

hood of only roughly 14% and 7% for the escort and the pig network, respectively. Two factors contribute to these results. First, as seen in Sect. 5.1, both networks are highly fragmented, making outbreak detection difficult. Second, depending on the concrete parametrization, there is a considerable fraction of outbreak scenarios that die out before they could even be detected at the monitoring date. If we skew our analysis to outbreaks that are still active at the monitoring date, the fraction of detected outbreaks increases to 20% in the escort network and to 13% in the pig network, both for $k = 100$. We can go even further and only consider outbreaks with a certain minimal size. To this end, we again simulate outbreak scenarios for the escort network based on a SIS model with $p = 0.6$ and $\mu^{-1} = 15$. For every minimal outbreak size, we simulate 3,000 scenarios and count the number of outbreaks that the different methods (with $k = 10$ nodes) detect. Figure 4 (left) shows that all methods detect larger outbreaks more effec-

tively in the escort network. Our method detects more than half of all outbreaks that involve at least 5 nodes at the time of monitoring. This makes sense for two reasons: First, the more nodes that are affected, the more likely it is that one of the nodes is in our set of monitored nodes. Second, larger outbreaks tend to happen in more connected parts of the network where our method is better at detecting outbreaks.

Next, we compare the optimal node sets from Fig. 2 with node sets that were selected based on 5 previous 30-day periods. If the contact patterns in the networks were roughly constant, we would expect a similar performance for node sets chosen based on previous data. However, Fig. 3 shows that the performance is generally worse if we use nodes selected based on previous 30-day periods. Interestingly, for the pig network the node set that is closest to the reference period (October 2017) is the one selected based on June 2017. Additionally considering the node set based on February 2017 reveals that the pig network may exhibit some 4-month seasonalities that are associated with the production cycle in the pig industry. Overall, however, those findings imply that contact patterns vary significantly and node sets chosen based on past data do not generalize well for the two networks considered here.

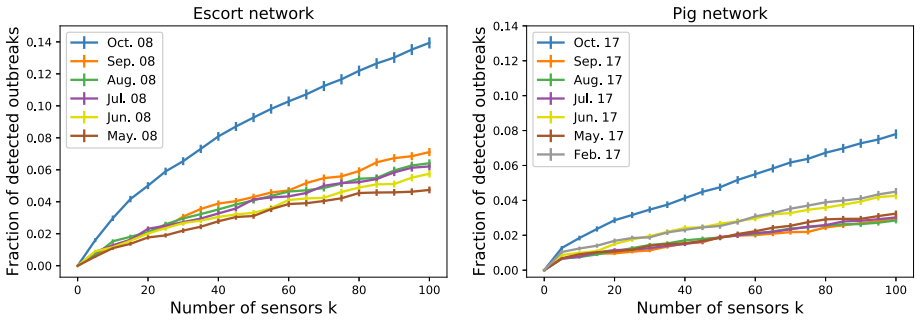

Fig. 3. Fraction of detected outbreaks with corresponding error bars for the escort network (**left**) and the pig network (**right**). The performance of the reference period (October) is compared to optimal node sets based on previous 30-day periods.

Finally, our method not only serves as an outbreak detection mechanism, it also detects the source of an outbreak given a sample of infected nodes at the time of monitoring. In order to test the source detection extension, we simulate 1,000 large outbreak scenarios (at least 8 infected nodes at the time of monitoring) according to the SIS model parametrized as above. We then randomly pick 1, 2, and 3 nodes from the simulated infection cascades as observed infected nodes and compute the posterior distribution according to (4), using a uniform prior. Note that for every possible source node, we sum the posterior probabilities over the different starting times in order to get an aggregated source probability per node. Figure 4 (right) shows the results. If we only observe one infected node,

the accuracy of the prediction is unsatisfactory because in more than half of all simulated outbreaks the true source is not ranked among the five top nodes. However, for two or three observed infected nodes, the accuracy of the source detection improves substantially. For example, we detect the source perfectly in more than 30% of all cases if we observe 3 infected nodes.

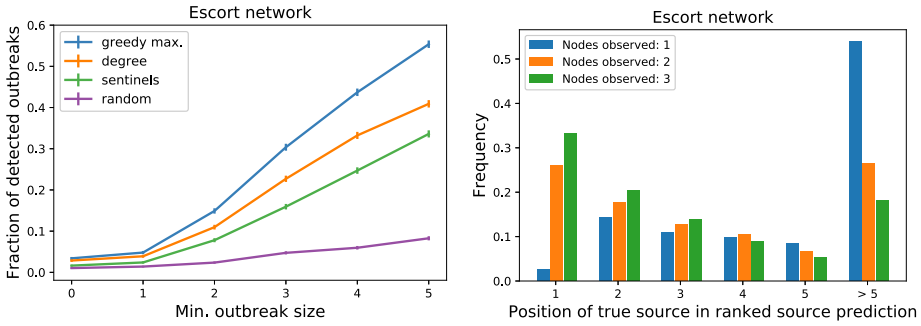

Fig. 4. Fraction of detected outbreaks with corresponding error bars for 10 nodes and varying minimal outbreak size (**left**). Source detection quality with varying numbers of nodes observed (**right**).

5.3 Hypothesis H2 (Robustness)

Methodology. In order to test *H2*, we use our method (Step 1 and 2) to identify nodes with the true underlying infection probability $p = 0.6$ as well as with misspecified probabilities $p = \{0.1, 0.3, 0.9\}$, keeping $\mu^{-1} = 15$ fixed. Likewise, we identify monitoring nodes with the true underlying recovery time $\mu^{-1} = 15$ as well as with misspecified $\mu^{-1} = \{5, 30\}$, keeping $p = 0.6$ fixed. We simulate 10,000 outbreak scenarios based on the SIS model with the true underlying values of $p = 0.6$ and $\mu^{-1} = 15$ and compare the fraction of detected outbreaks.

Results. Our method seems to be robust against misspecified infection probabilities p. The fraction of detected outbreaks in the escort network is only slightly smaller if we use nodes found based on misspecified values of p (Fig. 5, left). However, the detection performance can deteriorate considerably if μ^{-1} is misspecified (Fig. 5, right). Note that the decrease in detection likelihood is more significant if μ^{-1} is smaller than the true value. This makes sense intuitively as assuming a too small μ^{-1} will neglect many possible infection paths.

6 Conclusions and Future Work

The goal of this work is to identify a small set of nodes for monitoring such that the likelihood of detecting recent outbreaks in temporal networks is maximized.

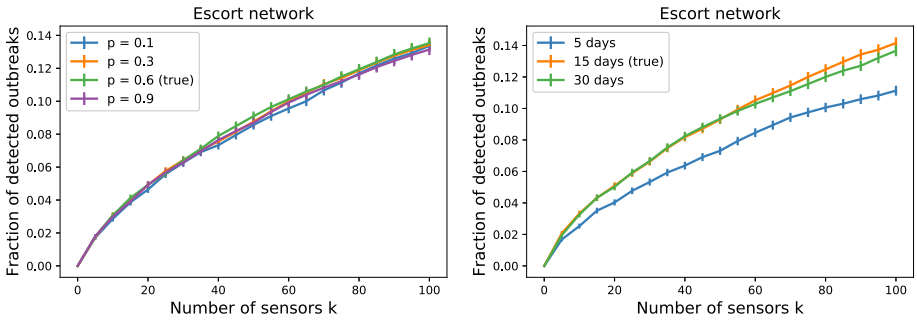

Fig. 5. Fraction of detected outbreaks with corresponding error bars for the escort network when the infection probability p (**left**) or the recovery period μ^{-1} (**right**) is misspecified.

Based on an evaluation with two different datasets, we find that our method outperforms other methods and it is especially effective for detecting large outbreaks. Moreover, it is robust with respect to the choice of parameters in the SIS model, as long as a large enough recovery time is chosen. Additionally, our method naturally extends to the problem of infection source detection. Although we restrict our examples to the SIS model, our method can be used with any propagation model that can be simulated. The findings reported here are crucial for the development of new outbreak detection strategies because our approach performs well in different contexts and applies to the realistic scenario where, on a given day, we need to find the optimal individuals to monitor. Moreover, our method is derived from basic concepts in probability theory rather than based on heuristics. The major limitation of our method compared to heuristics suggested previously is its computational intensity that can be prohibitive for large networks. However, the expensive simulation procedure could be easily parallelized. Another limitation of our method is that it only applies to the maximization of the detection likelihood. Other maximization objectives, such as the detection time, cannot be transferred easily to our problem. Further experiments, using different datasets and more complex propagation models, are an essential next step in confirming the generalizability of this method. Overall, we are convinced that our method can be an invaluable tool in a practical disease surveillance context.

Acknowledgement

This work was supported by the Swiss National Science Foundation (SNSF) NRP75, Project number 407540_167303. M. Sterchi was partially supported by the Hasler foundation. We would like to thank Identitas AG for providing the pig movement data and Emily E. Raubach, Heiko Nathues, Beat Hulliger, and the anonymous reviewers for helpful comments.

References

1. Aggarwal, C.C., Lin, S., Yu, P.S.: On influential node discovery in dynamic social networks. In: Proceedings of the 2012 SIAM International Conference on Data Mining. pp. 636–647 (2012)
2. Antulov-Fantulin, N., Lančić, A., Šmuc, T., Štefančić, H., Šikić, M.: Identification of patient zero in static and temporal networks: Robustness and limitations. Phys. Rev. Lett. 114, 248701 (2015)
3. Bajardi, P., Barrat, A., Savini, L., Colizza, V.: Optimizing surveillance for livestock disease spreading through animal movements. J. R. Soc. Interface 9(76), 2814–2825 (2012)
4. Barrat, A., Barthlemy, M., Vespignani, A.: Dynamical Processes on Complex Networks. Cambridge University Press, New York, NY, USA, 1st edn. (2008)
5. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web. pp. 665–674. ACM, New York, NY, USA (2011)
6. Christakis, N.A., Fowler, J.H.: Social network sensors for early detection of contagious outbreaks. PLOS ONE 5(9), 1–8 (2010)
7. Dubé, C., Ribble, C., Kelton, D., McNab, B.: Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in Ontario, Canada. Transboundary and Emerging Diseases 55(9-10), 382–392 (2008)
8. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 137–146. KDD '03, ACM, New York, NY, USA (2003)
9. Krause, A., Golovin, D.: Submodular function maximization. In: Tractability: Practical Approaches to Hard Problems. Cambridge University Press (2014)
10. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 420–429. ACM, New York, NY, USA (2007)
11. Nemhauser, G.L., Wolsey, L.A.: Best algorithms for approximating the maximum of a submodular set function. Mathematics of Operations Research 3(3), 177–188 (1978)
12. Panagopoulos, G., Malliaros, F.D., Vazirgiannis, M.: DiffuGreedy: An influence maximization algorithm based on diffusion cascades. In: Aiello, L.M., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L.M. (eds.) Complex Networks and Their Applications VII. pp. 392–404. Springer International Publishing, Cham (2019)
13. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200–3203 (2001)
14. Rocha, L.E.C., Liljeros, F., Holme, P.: Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLOS Computational Biology 7(3), 1–9 (2011)
15. Schilling, R.L.: Measures, Integrals and Martingales. Cambridge University Press (2005)
16. Sterchi, M., Faverjon, C., Sarasua, C., Vargas, M.E., Berezowski, J., Bernstein, A., Grütter, R., Nathues, H.: The pig transport network in Switzerland: Structure, patterns, and implications for the transmission of infectious diseases between animal holdings. PLOS ONE 14(5), 1–20 (2019)