# Social Communications Assisted Epidemic Disease Influence Minimization

Bowu Zhang[1], Pei Li[2], Xiuzhen Cheng[1], Rongfang Bie[3], and Dechang Chen[4]

[1] Computer Science, The George Washington University, DC, USA
[2] College of Information Systems and Management,
National University of Defense Technology, ChangSha, China
[3] Information Science and Technology, Beijing Normal University, Beijing, China
[4] Division of Epidemiology and Biostatistics,
Uniformed Services University of the Health Sciences, MD, USA
{bowuzh,cheng}@gwu.edu, pli.nudt@gmail.com,
bierf@163.com, dechang.chen@usuhs.edu

**Abstract.** This work explores the use of social communications for epidemic disease control. Since the most infectious diseases spread through human contacts, we focus on modeling the diffusion of diseases by analyzing the social relationship among individuals. In other words, we try to capture the interaction pattern among human beings using the social contact information, and investigate its impact on the spread of diseases. Particularly, we investigate the problem of minimizing the expected number of infected persons by treating a small fraction of the population with vaccines. We prove that this problem is NP-hard, and propose an approximate algorithm representing a preventive disease control strategy based on the social patterns. Simulation results confirm the superiority of our strategy over existing ones.

**Keywords:** Preventive disease control, social networks, target vaccination.

## 1 Introduction

In this paper, we propose to use communication records to guide the use of vaccines in a population to minimize the impact of epidemic disease. Infectious diseases pose major risks to the human life and social development. In recent years, various infectious diseases such as H1N1 and SARS have caused thousands of deaths and severe economic loss. Most of the infectious diseases can be transmitted from one person to another through personal contacts. With rapidly growing global transportations, infectious diseases that formerly die out in isolated areas may now spread worldwide. In order to stop infectious diseases from spreading, a number of interventions are available with distinct benefits or drawbacks. They either directly impact the transmission of diseases so that the viruses/germs cannot easily spread through the population, or immunize segments of a population. Considering the time that interventions are applied,

strategies for controlling epidemic diseases can be classified into two categories: 1) preventive disease control that takes place before a disease occurs in a population; and 2) reactive disease control that stops a disease from spreading out after the disease outbreak is detected in a population. Reactive strategies focus on individuals who are already infected and their close friends to protect healthy people from being infected. Preventive strategies, on the other hand, identify people at high infection risks to take vaccines, so that the disease can be prevented from spreading out or even happening in a population.

In this paper, we design a preventive disease control strategy that takes action to set up defenses against disease breakout ahead of time. In particular, we attempt to immunize a small number of people who are in danger in advance, so that we can prevent them from developing deadly infections and spreading to others. Such preventive vaccination strategies are widely used to prevent diseases, i.e., people are encouraged to take flu shot before the flu season starts. However, current preventive disease control mainly relies on mass vaccination strategy, which intends to immunize a major fraction of a population, leading to a high economic expense and adverse side effects. Moreover, due to production cycle and population growth, vaccines are often in short supply, making the mass vaccination ineligible for a number of occasions. In light of the problems above, we aim to select individuals to receive vaccines in advance based on information extracted from people's daily life data stream, so that the number of people infected will be minimized if the population is exposed to a disease later.

Prior work that selects target individuals for vaccination mainly fall under two categories: 1) node centrality methods, which rank nodes by measures such as degree, shortest path, or random walk betweenness; and 2) influence cascade methods, which select a subset of nodes that maximize information diffusion using independent cascade models or linear threshold models. The latter has been widely adopted in many recent works due to the rising of the idea of viral marketing, where commercial messages are sent to people who are socially important in order to achieve marketing objectives. In this paper, on the contrary, we try to minimize the diffusion of disease, rather than maximizing the spreading of a piece of information. In particular, we show that our minimization problem is NP-hard, and can be solved by an approximation algorithm following the information cascade model. Our approach is evaluated over real world data set. The results demonstrate that the proposed target immunization outperforms other strategies in providing protection over the population.

Our contributions can be summarized as follows:

- We explore social communications to determine the pattern of disease transmissions among individuals.
- We attempt to minimize the expected number of infected individuals by treating a small fraction of the population. This problem is proved to be NP-hard in this paper and we propose an approximation algorithm that provides a simple preventive disease control strategy.
- A comparison based simulation study over a real-world data set is conducted to evaluate the performance of our preventive disease control strategy and

the results indicate that the proposed strategy is superior over the most popular existing ones.

The paper is organized as follows. Section 2 presents the most related work. Sections 3 discusses the preliminary knowledge to be used in this work. An important metric, the transition probability, is defined and analyzed in section 4. The disease control minimization problem and an approximation solution are detailed in section 5. The proposed strategy is evaluated under various scenarios in section 6. We conclude our work in section 7.

## 2   Related Work

Our work is built on considerable prior research on disease propagation and social relationship identification. In this section, we review the most related work along these directions.

Disease propagation has been studied for a long time in human contact networks [1–3], where human beings are interpreted as nodes and their interactions are edges connecting nodes. Various mathematical models have been developed to characterize the disease transmission in a contact network. Most of the existing models [4],[5] describe the spread of disease in a homogeneously mixed network, where an infected individual infects each of his/her neighbors in a uniformly random and independent way. Under the constraint of fixed network size, differential equations can be written down to represent the movement of disease in a network, from which the number of infected people can be estimated. However, in real world, an infected person does not infect his contacts with an equal probability, as the length and the nature of the interactions among people can vary greatly from one to another. In recent years, a few investigations have been made to study the impact of variations in degree, infectiousness, and closeness of interactions. For instance, different network topologies such as sparse networks [6], clustered networks [2], and power-law networks [7], have been examined for their effect on disease propagation process. In this research, we construct a contact network based on social communications, and model the spread of disease as a function of the social relationship between individuals.

Many works have offered practical insights into the impact of social relationship on the development of applications in various domains, ranging from monitoring/tracking applications, to medical, emergency, and military applications [8–10]. Among them, a large body of research has addressed the problem of viral marketing [11, 12], where the social connections are used to spread the product information to achieve business objectives. The work by [13] studies the information propagation in social networks using two information diffusion models: independent cascade (IC) and linear threshold. The paper proves the NP-hardness of the problem of finding a small number of influential nodes as initial information adopters to maximize the expected number of nodes that would adopt the information. In contrast to influence maximization (i.e., recent research on viral marketing mentioned above) that attracts attention from finance and social network communities, influence minimization has relatively

been less investigated. Misinformation such as scams, false twitter or facebook posts, have led to enormous social and economic issues. In order to minimize the impact of misinformation, researchers have considered to block critical nodes to prevent the false message from spreading out. Such an idea has been applied in a number of contexts. In this work, we consider a similar minimization problem which we prove it to be NP-hard.

## 3    Preliminary

### 3.1    Contact Network

We investigate the spread of diseases over a contact network constructed over social communication records. Define the contact network as $G(V, E, W)$, where $V$ denotes the node set, $E$ denotes the edge set, and $W$ represents the edge weight set. A node $v \in V$ represents an individual, while an edge $e_{ij}$ between nodes $i$ and $j$ indicates that there exist interactions between the two nodes. Attached to the edge $e_{ij}$ is the weight $w_{ij}$, which is used to quantify the interaction between $i$ and $j$. If the interaction is made through phone calls, then $w_{ij}$ can be the total time used in the phone conversations. If the interaction is done through messages, then $w_{ij}$ can include information such as the number of the messages and the sizes of the messages.

### 3.2    The SIR Model

We use a SIR model to describe the progress of a disease in a single node. Nodes may be either *Susceptible* to the disease, or *Infected* with the disease, or *Recovered* from the disease. Assume an infective person stays *Infected* for $\alpha$ time intervals. During each time interval, an *Infected* node $i$ infects a *Susceptible* node $j$ that has interactions with $i$ with a probability $p$. After $\alpha$ time intervals, an *Infected* node becomes *Recovered*. A *Recovered* node can not get infected or infect others. Nodes that receive vaccinations become *Recovered* automatically. Notice that we differentiate $p$ from the transition probability defined in section 4, where $p$ stands for the probability that a disease is transmitted between two nodes during one time interaction, while the transition probability states in general the probability for a node to pass a disease to another node. For the sake of simplicity, we assume $p$ is the same for any two nodes. However, the transition probability, that we will discuss later, relies on the variations in closeness of human connections.

### 3.3    IC Model

We employ an independent cascade (IC) model [13] to study the spread of disease among people. In this model, the diffusion proceeds in discrete time intervals. Nodes in this model are either active or inactive. The diffusion starts with an initial active node set. Assume node $i$ becomes active at step $t$. Then $i$ will attempt to activate each of its inactive neighbors, $j$, with a transition probability

$t_{ij}$, which indicates the tendency of $j$ to be activated by $i$. If $i$ succeeds, $j$ becomes active at step $t + 1$. If $j$ has multiple active neighbors at step $t$, their activation contacts with $j$ would be sequenced in a random order. The IC process terminates if no more activations are possible. Denote by $\sigma(i)$ the *influence degree* of a node $i$, which is defined to be the expected number of nodes influenced by $i$ at the end of the IC process. In this work, we aim to minimize the expected number of infected people by vaccinating a small fraction of the population.

## 4    Transition Probability

Since infectious diseases spread through human contacts, the disease transition probability strongly depends on the pattern of contacts between infected persons and others. Many works assume that this probability is the same between any two individuals, while in fact, it is obvious an infected person will not infect all others with the same probability. For example, the probability of a disease transmitted between two close friends is higher than the probability of disease transmission between two strangers. A number of factors influence the transition probability, including the vaccination history, the general health of the normal person, the nature of the disease, and the nature of the interaction (e.g., time, interaction type) between the individual and the infected person(s). In this work, we try to capture these factors based on social communication records so that we can predict how diseases are transmitted in a population.

Let $T$ be a $N \times N$ matrix. Denote by $t_{ij}$ the element of $T$ at row $i$ and column $j$, which states the transition probability describing how likely an epidemic disease will be passed from node $i$ to node $j$. Considering the IC process which determines the diffusion of a disease, during each step, a disease only moves from the current node to its immediate neighbors. Therefore, for a pair of nodes that are not connected by edges, the transition probability between them is zero. For a pair of connected nodes, it is widely accepted that the transition probability between them depends on their interaction patterns [3, 14], since epidemic diseases spread through human contacts. In this work, we model the transition probability among individuals by analyzing the social relationship information in $G$ over their contact records. Following our preliminary work [15, 16], we define the transition probability $t_{ij}$ as follows:

$$t_{ij} = f(d_{ij}, N_i, N_j, w_{ij}, \sum_{k \in V, k \neq i} w_{ik}, \sum_{k \in V, k \neq j} w_{jk}) \tag{1}$$

where $f$ is a multivariate function ranging from 0 to 1, $N_i$ represents the set of nodes which communicate with $v_i$ directly, and $d_{ij}$ denotes the physical distance between $i$ and $j$, which can be roughly estimated through cell phone's built-in GPS, or through cell phone tower log. Note that two individuals who have a large number of phone interactions may not have physical interactions through which epidemic disease can be transmitted. Therefore, we include $d_{ij}$ to ensure that $t_{ij}$ in the above case is not high. In addition, we let $t_{ii} = 0$ to avoid self-loop. The computation of $t_{ij}$ can be completed locally at $i$ and $j$, without requiring

the global network information, which leads to low communication overhead. Further discussions on $t_{ij}$ can be found in our preliminary work [15, 16].

## 5   Disease Influence Minimization

Different from target selection in viral marketing, which attempts to maximize the number of nodes that can be affected by the target set, we try to minimize the impact of disease over $G$ by immunizing a small number of nodes. Denote by $K$ a pre-defined constant satisfying $|K| < N$. Given a network $G(V, E, W)$ and the transition matrix $T$, we aim to immune a set $A$ of $K$ nodes, $A \subset V$, so that the number of nodes infected will be minimized if disease occurs. Assume that disease appears at each node with a probability $q$, where $q$ is a positive constant less than 1. Thus after vaccinating nodes in $A$, the expected number of nodes infected by $v, v \in A^-$, is $q \cdot \sigma(v)_{\text{over } G^{A^-}}$, where $A^-$ is the complementary set of $A$, and $G^{A^-}$ is the subgraph of $G$ induced by all the nodes in $A^-$. Therefore, the sum of the expected number of nodes infected by each node in $A^-$, denoted by $AVG(A^-)$, is $\sum_{v \in A^-} q \cdot \sigma(v)_{\text{over } G^{A^-}}$. Then our goal is to find $A$ of $K$ nodes for vaccination, so that $AVG(A^-)$ is minimized. Notice here $AVG(A^-)$ measures the sum of the expected number of nodes infected by each node in $A^-$, not the sum of the union of the expected number of nodes infected, in case that the nodes infected/influenced by $i \in A^-$ and the nodes infected/influenced by $j \in A^-$ overlap. The definition of $AVG(A^-)$ is based on the consideration that we attempt to minimize the impact of disease no matter which node the disease starts with. Then the problem can be mathematically described as:

$$\min_{A \subseteq V} AVG(A^-) \text{ s.t } |A| = K \tag{2}$$

To solve (2), intuitively, we can check every possible set of $K$ nodes in $V$, which takes $\binom{N}{K} = O(n^K)$ time. We will prove in the following section that this minimization problem is NP-hard, which can not be solved efficiently. An approximation algorithm is then proposed to provide an approximate solution.

### 5.1   NP-hardness

Consider the following sum-of-squares partition problem [17]. Let $G(V, E)$ be an undirected graph, with a node set $V$ and an edge set $E$. Given a constant $K$, the sum-of-squares problem attempts to partition $G$ into disconnected components $C_1, \ldots, C_i, \ldots$ by removing a set $A$ of at most $K$ nodes such that $\sum_i |C_i|^2$ is minimized. This problem is known to be NP-hard. We present in the following, that we can reduce the sum-of-squares partition problem to the proposed minimization problem in polynomial time when the transition probability is 1.

**Theorem 1.** *When the transition probability is 1, finding a node set $A \subseteq G, |A| = K$ so that $AVG(A^-) \leq AVG(\Theta^-)$ for $\forall \Theta \subseteq V, |\Theta| = K$ is a NP-hard problem.*

*Proof.* As the transition probability between any two nodes in $G$ is 1, once $v \in G$ is infected, all the nodes that can be reached from $v$ will be infected. Then for any node $v$ in $C_i$, if $v$ is infected, the expected number of influenced nodes is $|C_i|$. Thereby the sum of the expected number of infected nodes is $AVG(A^-) = \sum_i q \cdot |C_i|^2$. As a result, minimizing $\sum_i |C_i|^2$ is equivalent to minimizing $AVG(A^-)$ in $G$. Through the above steps, we successfully transform the sum-of-squares partition problem to the problem of minimizing $AVG(A^-)$ when the transition probability is 1 in polynomial time. Therefore, the problem of minimizing $AVG(A^-)$ is NP-hard when the transition probability is 1.

**Theorem 2.** *The influence minimization problem defined by (2) is NP-hard.*

*Proof.* Since the problem considered in Theorem 1 is a special case of minimizing $AVG(A^-)$ when the transition probability varies at edges, the general $AVG(A^-)$ minimization problem (2) is also NP-hard.

### 5.2 Approximation Algorithm

We design a simple approximation algorithm to find a set $A$ of $K$ nodes to solve (2). Initially $A = \emptyset$. We add $K$ nodes into $A$ to maximize $\sum_{v \in A} q \cdot \sigma(v)$ over $G$. The details of the algorithm are presented in Algorithm 1.

---

**Algorithm 1.** Approximation Algorithm($G$)

**Input:**

– $G$: the contact network

**Output:**

– $A$: a set of $K$ nodes for vaccination

1: **function** ANPROXIMATION ALGORITHM($G$)
2:     Let $A = \emptyset$.
3:     Sort nodes in $V$ in a decreasing order of $q \cdot \sigma(v)$ over $G$.
4:     Add the first $K$ nodes in the sorted list into $A$.
5:     Output $A$;
6: **end function**

---

## 6    Simulation Study

### 6.1    Simulation Set-Up

We validate the proposed disease control strategy over a real-world data set from Facebook (http://snap.stanford.edu/data/egonets-Facebook.html), where facebook friend information of 3959 individuals are collected. In this simulation, we let

$$t_{ij} = \frac{|N_i \bigcap N_j| + 1}{|N_i \bigcup N_j| + 1}.$$

Here, $N_i$ represents the set of nodes that are directly attached to $i$. The number '1' appearing on the numerator and denominator is used to prevent $p_{ij}$ from becoming 0. This action is based on the idea that a node always has a potential influence on any other node.

To verify the strength of the proposed strategy, we implement two other approaches for performance comparison. One approach employs a random alert strategy in the sense that a number of individuals are randomly chosen to be alerted for vaccination according to the number of available vaccines. This strategy has been widely used in the literature and is denoted as RD in this paper. Another implemented approach uses degree centrality, where the nodes with the largest node degrees are chosen for vaccination according to the number of available vaccines. It is argued by the previous work [18] that degree centrality yields promising results in predicting the risk of infection, compared to other centrality metrics.

We define the final infection ratio as the ratio of the total number of infected persons during the time of evaluation to the size of the entire population. The final infection ratio will be used as the primary performance metric for the evaluation of disease control strategies in our simulations.

We examine these strategies by varying different parameters such as the number of available vaccines, $q$, as well as the infection probability $p$ mentioned in section 3.2. The initial infection ratio is defined as the total number of infected persons on the first day divided by the size of the population. The initial infected persons are chosen randomly in our simulations. We report our experimental results by an average of 50 runs.

## 6.2   Simulation Results

Fig. 1 reports the final infection ratio *vs.* the number of available vaccines, where we fix $q$ to be 0.003, and the infection probability $p$ to be 0.05. From the results, we observe that the final infection ratio declines as the number of available vaccines increases. When the number of people who receive vaccines is growing, more people are protected, and therefore the number of infected nodes is decreasing. Overall, the target vaccination strategies (degree and the proposed strategy) achieve a lower number of infected nodes than the random strategy, which is consistent with the previous work [15], [16], [3] whose strategies have resulted in less number of infections than the random strategy with the same cost of vaccines. The proposed strategy achieves the lowest final infection ratio when the number of available vaccines is less than 1000, and has almost the same final infection ratio as the degree strategy when the number of vaccines is larger than 1000. The results demonstrate that compared to the other two strategies, the proposed method protects more people with the same number of vaccines when there is a limited supply of vaccines.

In Fig. 2, we plot the final infection ratio *vs.* $q$, where the number of available vaccines is 200, and $p$ is 0.05. Notice that for all three strategies, the final infection ratio increases as $q$ increases. This is because when the number of vaccines is fixed, an increasing number of initial infected nodes will lead to a larger chance
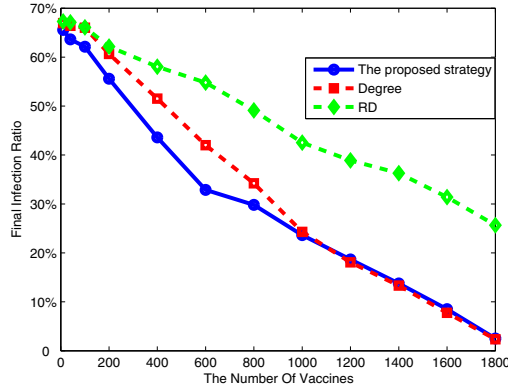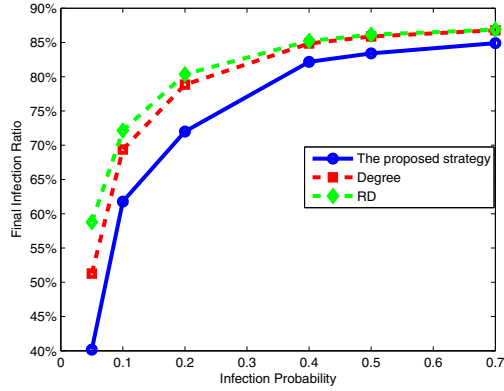
**Fig. 1.** Final Infection Ratio *vs.* Number Of Available Vaccines ($q = 0.003$, infection probability $= 0.05$).
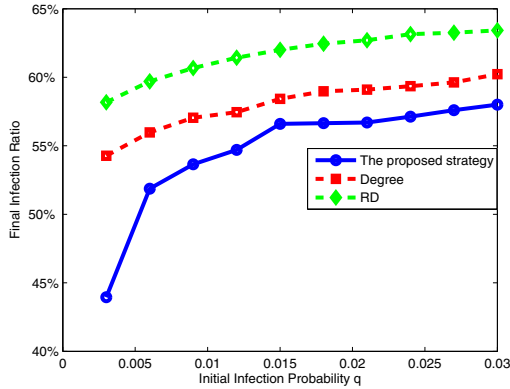


**Fig. 2.** Final Infection Ratio *vs.* $q$ (Number of available vaccines $= 200$, infection probability $= 0.05$)

of spreading diseases to more nodes. In general, target vaccination strategies perform better than the random strategy, which is consistent with the results in Fig. 1. We also observe that the proposed strategy stands out from all others under different $q$. This indicates that our approach has a more effective capability to prevent disease from spreading no matter how seriously the disease starts initially in a population.

We also evaluate the vaccination strategies under different infection probabilities in Fig. 3, where the number of available vaccines is fixed to be 400, and $q$ is fixed to be 0.003. It is observed that the final infection ratio increases when the infection probabilities grow larger. This is reasonable because for a single

**Fig. 3.** Final Infection Ratio *vs*. Infection Probability (Number of available vaccines = 400, $q = 0.003$)

infected node, the number of nodes that might be infected by this node is a monotonically increasing function of the infection probability. Since $AVG(A^-)$ is the sum of the expected number of nodes infected by each node in $A^-$, it is also increasing as the infection probability ones. Our proposed strategy outperforms the other two under different $p$, again confirming the superiority of the proposed strategy to existing strategies. However, as the infection probability rises up, the superiority of the proposed strategy becomes less apparent. It can be observed that the performance differences among all the three strategies decline when the infection probability goes up. This reflects the real life scenario that when the disease is highly transmissible, there is less need to target people at risks, because everybody who has interactions with the infected person are all likely to be infected.

## 7   Conclusion

In this paper, we design a preventive disease control strategy to set up defenses against disease breakout ahead of time. The social communications are explored to extract social information such that we can determine the pattern of disease transmissions among individuals. We attempt to minimize the expected number of infected individuals by treating a small fraction of the population; this minimization problem is proved to be NP-hard in this paper and thus we propose an approximation algorithm. Simulations and comparisons are conducted to evaluate the performance of the proposed disease control strategy over a real-life data set. The results indicate that the proposed strategy is superior over existing ones.

# References

1. Newman, M.E.J.: Spread of epidemic disease on networks. Phys. Rev. E 66, 016128 (2002)
2. Miller, J.C.: Spread of infectious diseases through clustered populations. Journal of the Royal Society Interface 6(41), 1121–1134 (2009)
3. Ren, Y., Yang, J., Chuah, M.C., Chen, Y.: Mobile phone enabled social community extraction for controlling of disease propagation in healthcare. In: 2011 IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), pp. 646–651. IEEE (2011)
4. Dimitrov, N.B., Meyers, L.A.: Mathematical approaches to infectious disease prediction and control. In: Hasenbein, J.J. (ed.) Informs Tutorials in Operations Research, vol. 7, pp. 1–25 (2010)
5. Hethcote, H.: The mathematics of infectious diseases. SIAM Review 42(4), 599–653 (2000)
6. Anderson, R., May, R.: Infectious diseases of humans: dynamics and control. Oxford University Press (1991)
7. Zhou, T., Yan, G., Wang, B.H.: Maximal planar networks with large clustering coefficient and power-law degree distribution. Phys. Rev. E 71, 046141 (2005)
8. Meyers, L.A.: Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bull. Amer. Math. Soc. 44, 63–86 (2007)
9. Perisic, A., Bauch, C.T.: Social contact networks and disease eradicability under voluntary vaccination. PLoS Comput. Biol. 5(2), e1000280 (2009)
10. Huang, S.: Probabilistic model checking of disease spread and prevention. In: Scholarly Paper for the Degree of Masters in University of Maryland (2009)
11. Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th International Conference on World Wide Web, pp. 721–730. ACM (2009)
12. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 241–250. ACM (2010)
13. Kempe, D., Kleinberg, J., Tardos, É.: 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. ACM (2003)
14. Liang, X., Lu, R., Chen, L., Lin, X., Shen, X.: Pec: A privacy-preserving emergency call scheme for mobile healthcare social networks. Journal of Communications and Networks 13(2), 102–112 (2011)
15. Zhang, B., Cheng, X., Bie, R., Chen, D.: A community based vaccination strategy over mobile phone records. In: Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare. mHealthSys 2012, pp. 2:1–2:6. ACM (2012)
16. Zhang, B., Gilani, S.M., Wu, D., Cheng, X., Bie, R.: Mobile phone based social relationship identification for target vaccination in mobile healthcare. In: Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones. PhoneSense 2012, pp. 5:1–5:5. ACM (2012)
17. Aspnes, J., Chang, K., Yampolskiy, A.: Inoculation strategies for victims of viruses and the sum-of-squares partition problem. Journal of Computer and System Sciences 72(6), 1077–1093 (2006)
18. Han, B., Hui, P., Kumar, V.A., Marathe, M.V., Shao, J., Srinivasan, A.: Mobile data offloading through opportunistic communications and social participation. IEEE Transactions on Mobile Computing 11(5), 821–834 (2012)