Vaccination and Quarantine Policies for Controlling Pandemic Disease Outbreak

Shirish Tatikonda, Srinivasan Parthasarathy*, and Sameep Mehta
Department of Computer Science and Engineering,
The Ohio State University, Columbus, OH 43210, USA

ABSTRACT

There have been a number of significant pandemics in human history such as cholera, influenza, and smallpox. These pandemics are widespread, highly infectious, and kill a large number of people. Recently with the outbreak of $H5N1$ influenza virus in avian populations, it is speculated that another serious transmissible pandemic might occur because of the mutations in unstable $H5N1$ strain. Therefore, it is essential to be prepared for such a sudden and fatal transmissible disease outbreak. In this paper, we develop several containment strategies or policies to curb such infectious diseases from spreading. As part of this study, we design and evaluate several vaccination policies which identify individuals and locations which are critical in spreading the disease. We propose a novel quarantining methodology that is based on hierarchical clustering. We demonstrate the effectiveness of the proposed strategies using datasets generated by EpiSims system. We, further, evaluate the robustness of the policies under various practical constraints such as limited number of anti-viral drugs, and delay in implementation of quarantining and vaccination policies.

1. INTRODUCTION

Pandemic diseases such as the *avian influenza* are extremely infectious and lethal. This infectious disease is caused by the type A strains of influenza virus and has now spread to 13 different countries in Asia and Europe. Outbreak of pathogenic $H5N1$ avian influenza or bird flu was first reported in 2003 in South-East Asia. Over a past couple of years, several cases of influenza virus have been observed worldwide such as in China, Indonesia, Egypt etc. Till date, a total of 229 human cases have been reported causing 131 deaths[1]. In the year of 2006, cases of bird flu were also reported in Turkey and Iraq. This was the first time the presence of this virus was recorded in these countries, demonstrating the ability of the virus to spread worldwide and effect individuals across all age groups. Through a process of re-assortment events and adaptive mutations in the influenza virus, outbreak of a fully transmissible large-scale pandemic is not an improbable postulation. Furthermore, increase in global transport, urbanization and overcrowded conditions, can act as catalyst in the spread of the disease. A combination of these factors can, unfortunately, lead to serious outbreak of the disease that can spread more quickly than in the past, overwhelming countries and health systems that are not adequately prepared.

Given such a lethal threat, it is essential to be prepared for the potential pandemic outbreak. A straightforward and foolproof strategy is to vaccinate the whole population. However, the cost associated with this approach is extremely high. Apart from the prohibitive cost, man power and logistics required to administer a population wide vaccination makes this policy an impractical option. Please note that population wide vaccination for diseases like smallpox and polio is implemented in a very controlled and precise fashion. The vaccines are given to infants during their early years. Moreover, since the virus also mutates frequently, the flu vaccine must be concocted anew each year, which makes the option of vaccinating the whole population improbable (if not impossible). Therefore, it is imperative to develop effective and efficient policies to control the outbreak of new diseases with limited resources.

Typically, an infected individual is likely to transmit the disease to a healthy person in close spatial vicinity if there is an interaction. The actual length and nature of the interaction leading to disease transmission depends on the disease specific properties. Therefore, understanding and modeling the interactions among people and their movement in spatially proximate geographic regions is the key for assessing the transmissibility of the virus and using that information to develop containment policies. In this paper, we focus on *People-People Interactions*(PPI) and *People-Location Interactions*(PLI).

PPI, as the name suggests, captures the relationships among individuals. These interactions are used to find people who are either more susceptible to the disease or are capable of infecting several other individuals. In case of limited vaccines, highest priority should be given to vaccinate these individuals.

PLI models the relationships between people and locations. This is used to find locations where the presence of an infected person can be extremely hazardous. For example, an infected individual going to school or mall. Again, in case of limited resources, these locations should be quarantined first. Apart from above mentioned interactions, the inherent properties of spatial regions can also augment the disease spread in that region. The importance of this analysis was exemplified by John Snow[2] who traced the source of Cholera to water sources by studying the disease spread rate in London during 1854 epidemic.

The key contribution of this study are:

- We have developed several vaccination and quarantin-

*Contact email: srini@cse.ohio-state.edu

[1] http://www.who.int/csr/disease/avian_influenza/en/

[2] http://en.wikipedia.org/wiki/John_Snow_(physician)

ing policies based on PPI and PLI. We evaluated the effectiveness of these policies by calculating the prevention rate in each case.

- We explored the possibility of combining these policies to derive hybrid policies which can result in higher prevention rates.
- We evaluated the proposed strategies under several practical constraints like limited number of anti-viral drugs and the delay in implementation of the policies.
- We also employed hierarchical clustering to discover the spatial regions which require immediate attention due to high disease spread rate.

This paper is organized as follows: In Section 2, we describe the simulation and data models. Section 3 presents various vaccination and quarantining policies. Finally, we present the study of effectiveness of proposed policies in Section 4 followed by related work and conclusions.

2. DATA AND MODELS

We use the simulation data generated using Episims[6][3]. EpiSims is a tool for simulating the spread of epidemics at the level of individuals in a large urban region, taking into account the realistic contact patterns and the disease transmission characteristics. It simulates the disease dynamics over an instance of time varying social contact network of individuals. These simulations are intended to model influenza virus, and the model assumes that an infected person can not be re-infected. During the course of simulation, an infected person can infect other individuals with whom that person is involved in certain activity (as defined by the contact network). When a person gets infected, the time and place at which the infection occurred is recorded by the system. The dataset used in this paper represent the synthetic population of the city of Portland, USA and captures the interactions among individuals. The dataset includes the following geographical and demographic information about locations and individuals:

- a set of individuals in the city of Portland with demographic information like gender, age, income, and house address(id).
- a set of daily activities of each individual.
- a set of aggregated activity locations with associated geographic information i.e., x, y coordinates.
- a social contact network of people representing the interactions among them. These interactions are tagged with the duration for which the contact is made.
- a description of disease outbreak and spread that includes the time and place when a person is infected. It also contains the information for other individuals present at that location.

The data consists of 1.6 million people spread over 246,000 different locations. At any given time during the simulation, an individual is involved in one of the following activities: *Home* (id: 0), *Work (1)*, *Shop* (2), *Visit* (3), *Social/Recreation* (4), *Other* (5), *Pick up* (6), *School* (7) and *College* (8). These activities define the person's interactions, if any, with other people. At the start of the simulation, 100 individuals are selected and are marked as diseased. The system is, then, run to simulate 100 days. With no containment policies, about 565,600 people ($\sim$ 35% of whole population) are infected by the end of 100^{th} day.

The given simulation data can be modeled in many different ways. The models should capture the interactions among people and the interplays between people and locations. We use two representations, *People - Locations Activities Graph* (PLA) and *People - People Contacts Network* (PPC) (Figure 1). The models are generic in nature and hence many graph theoretic algorithms can easily be applied to them. Though, these two models seem to capture different interactions, one model can be derived from the other. In our discussion, we chose to differentiate these two models for the ease of exposition.

The PLA graph, (V_P, V_L, E_P), is a bi-partite graph and models the relationship between people and locations in the city. V_P and V_L represents the set of all people and locations, respectively. An edge $(P_i, L_j) \in E_P$ implies that the person P_i is *connected* to the location L_j. Each edge is weighted with the type, start time, and duration of the corresponding activity. Therefore, an edge $(P_i, L_j) \in E_p$, weighted with a, s, and d implies that person P_i was at location L_j for d hours starting at s to perform activity a.

The PPC network, (V_C, E_C) captures the social interactions among individuals, where the individual are represented by nodes and an edge represent an interaction between two individuals. The edges in this network are weighted with type, start time and duration of the contact. Contact type is the purpose of interaction, i.e., it defines the activity type for which the contact is made. It is one of the 9 different activities (mentioned above). Contact hours $(0-23)$ is the duration for which the contact is made. An edge $(P_i, P_j) \in E_C$ weighted with c, s, h implies that the person P_i is in contact with P_j for the purpose of c and for the duration of h hours, starting at s.

We observe that PLA graph and PPC networks are scale-free networks. Their degree distributions follow the power law, i.e., number of nodes with degree k falls as $k^{-\alpha}$ for some constant α. For example, in PLA graph, we observe the value of α to be between $1.7 - 2.1$ for locations and around 1.8 for people. Interested readers are referred to Barabasi[1] and Newman[12] for an excellent discussion on scale-free networks and power law.

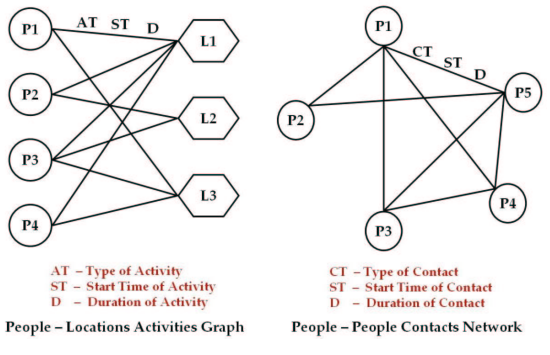


Figure 1: Data Models (a) People - Locations Activities graph and (b) People - People Contacts network

We now present the effect of demographic attributes like

[3] http://ndssl.vbi.vt.edu/opendata/

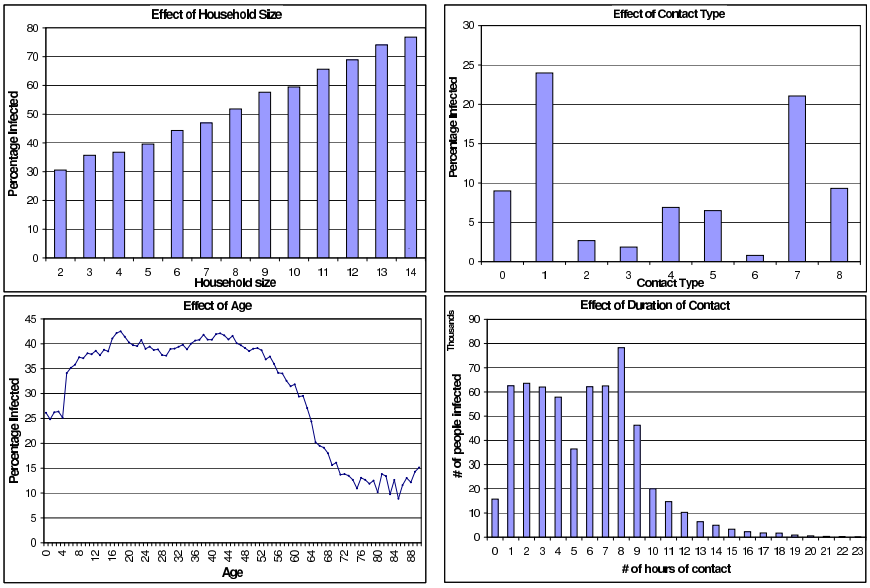


Figure 2: Effect of attributes on disease outbreak (a) Household size (b) Contact Type (c) Age (d) Contact Hours

age, household size etc. on the disease spread. Such an analysis can help in developing probability functions which can aid in gaining better insights into the data. We concentrate on four attributes viz., *household size, contact type, age,* and *hours of contact*. Figure 2 illustrates the effect of these attributes on disease spread. X-axis in each graph shows the type of attribute and Y-axis gives the number of people who got infected. Not surprisingly, an infected person in a larger household is more infectious than an infected person in a smaller household, because the person in larger household can spread the infection to more number of people. For example, an infected person in an household of size 14 can infect 80% of the other household members. Figure 2 (b) illustrates how the contact type effects the probability of a person to get infected. All types of contacts are not equally infectious. Contacts at schools and work places are far more infectious than other contacts. Similarly, a person's age can also effect the susceptibility of the person to the disease. We can infer that the people who are between the age of 5 and 60 are more susceptible to the infection than the people from other age groups. This can be explained by the fact that people from age group 5 - 60 are probable candidates to visit places with high probability in spreading the infection like schools and work places. The downward trend at higher ages should not be mistaken for less number of people in that age group. The y axis in this case depicts the percentage of infected people and not the actual counts. Along with the activity type, duration of the contact also effects the person's susceptibility to the disease. Chances of getting the infection is directly proportional to the time spent with the infected person.

Based on this analysis, we develop probability functions which maps the value of attribute to susceptibility to the disease. Let us consider the example of contact types. From figure 2 (b), we can infer that 24% of the people going to the work and 21% of the people attending the school are infected. By normalizing the percentages over all the activities, we can model the effect of contact type as a discrete probability distribution function. In similar spirit, we can construct probability mass functions for other attributes also. Based on a person's attributes (age, household size etc.) and the above developed attribute specific functions, we can derive the probability with which the person gets infected by simply multiplying the corresponding probabilities. These probabilities can then be used to assign weights to edges.

3. STRATEGIES

As mentioned earlier, a naive strategy to contain the disease is by vaccinating the whole population. However, difficulty in implementing and high cost renders this scheme practically infeasible. Another seemingly correct approach is to vaccinate all the household members of the 100 initially infected people. This strategy is easy to implement and cheap but not effective. It could only prevent $66,419$ cases out of a total of $565,685$ cases. Therefore, there is a need for more sophisticated methods which take into account interactions among people, effect of locations, activity type and number of contact hours. In this section, we propose several such policies. Each of the proposed strategy chooses a set of people to vaccinate from the whole population. Intuitively, we would like to choose and vaccinate the people who are more probable to get infected. Sections 3.1 to 3.5 presents the vaccinating policies, which are based on people's interaction network. Vaccinating policy presented in section 3.6 takes both spatial and temporal dimensions of people's movement into consideration. Quarantining policies are described in section 3.7.

3.1 Random Vaccination

This strategy randomly chooses the set of people to vaccinate. A simple random sampling without replacement (SRSWR) is employed for this purpose. Random numbers are generated using a Uniform distribution.

3.2 Contacts Driven Vaccination

Another vaccinating policy is to select the people who are directly in contact with an infected person. One can

continue to higher levels by choosing the people who are in direct contact with the contacts of infected person, and so on. In graph theoretic terms, this algorithm is same as Breadth-First Search (BFS). We start the BFS from each of the 100 people who are initially infected (*sources* of BFS). At level i, we select nodes which are connected to the *source* by $i-1$ nodes or i edges.

3.3 Sociability Driven Vaccination

This vaccinating policy is based on the sociability of individuals. Intuitively, a more sociable person has higher chances of getting infected than a person who is less sociable. We use the degree of a person in *PPC* graph (i.e., the number of contacts) as a measure of person's sociability. As mentioned in section 2, *PPC* network is a scale-free network. The distinguishing feature of a scale free networks is the presence of centrally located and highly connected nodes known as *hubs*. In the PPC network, vertex degrees (i.e., number of contacts) varies from 0 to 277 with the $\mu = 39$ and the $\sigma^2 = 33.28$. There are 956 people (0.06% of population) with number of contacts $\geq$ to 200. These hubs correspond to people who are much more societal compared to others. A straight forward strategy will be to vaccinate these hubs to control the outbreak. But, vaccinating these 956 people could only prevent 4655 cases, which is less than 1% of total number of infected people. Vaccinating people with number of contacts ≥ 150 could prevent only 12% of the cases. Therefore, vaccinating just the hubs is not very effective in containing the spread of the disease. Instead of concentrating on just the hubs, we propose to vaccinate all the nodes with degree greater than a fixed *cutoff degree*. The cutoff degree should be chosen carefully. Ineffective policy of vaccinating hubs corresponds to the choice of higher d. Too small of a cutoff degree will amount to vaccinating a very large percentage of the population, which will be cost ineffective. However, it can lead to better prevention rate. We discuss the detailed results highlighting the effect of chosen degree on the prevention rate in section 4.

3.4 Profile Based Vaccination

People with similar contact networks have similar probabilities of getting infected. In this strategy, we vaccinate the set of people whose contact network is similar to the contact network of an infected person. We make use of random walks on the PPC network to select the set of such people. Random walk is a process consisting of sequence of discrete steps of fixed length. In our context, each discrete step corresponds to traversing an edge of the graph. Random walk starts at a node (i.e., person) referred to as *source* of the walk. At each step, next node in the walk is chosen randomly from all the nodes, which are adjacent (connected directly) to the current node. Multiple random walks from a given node can be performed using restarts, i.e., at each step during the walk we can restart from the source with certain probability known as *restarting probability*. Two people are considered to have similar profiles if their contact networks are similar. Sun et. al [13] have demonstrated, in context of bi-partite graphs, that random walk with restarts is very effective in determining nodes which are most relevant (i.e., more similar) to the source node. If the source is an infected person, random walk visits a set of people who have similar profile as of the source. Therefore, people visited during such random walks are more susceptible to the infection and should be vaccinated first.

We treat each of the 100 initially infected people as sources of the random walks. In traditional random walk, probability of an edge to be taken is same for all the edges incident on a given node. From Figure 2(b), we know that the contact type and hours of contact effects a person's susceptibility to infection. Each edge in the contact network is thus weighted based on the probability distributions constructed for attributes, contact type and the duration of contact (as described in section 2). Assume that v is an intermediate node in the random walk and $E_v = \{e_1, e_2, ..., e_n\}$ is the set of edges incident on v. Let edge e_i corresponds to one of the v's contact with type X and duration H. Assume that p_x and p_h are the probabilities of getting infected for contact type X and the duration of contact H, respectively. Each edge e_i is then weighted with the product $p_x \times p_h$. We normalize the weights of all the edges incident on v such that the sum of weights is 1. With such weighing method, higher edge weights implies higher probability of getting infected. These edge weights are, then, used in choosing the nodes to be visited and, subsequently, vaccinated. In effect, the node selection process is biased towards the individuals with higher susceptibility.

3.5 Hybrid Policies

The above mentioned policies exploits the properties of the data models, PPC network and PLA graph, in devising effective containment policies. Each of these policies enjoy certain advantages and also suffers from some limitations. For example, sociability driven vaccination presented in section 3.3 is simple as it *only* depends on the sociability (degree) alone. In the context of more generic models such as time-varying social networks, such simplicity makes the policy implementation a difficult task. Hybrid strategies can be developed by combining two or more policies. These hybrid policies can leverage the positive features of each policy while reducing the detrimental effects of the individual policies. Above described four fundamental policies can be combined in many different ways. We present only the most effective combinations viz. sociability driven vaccination combined with profile and contacts based vaccination schemes. We also evaluated other combinations such as "Random + Sociability based scheme" and "Random + Profile based scheme" but we observed these combinations to be not as effective.

3.5.1 Profile + Sociability Based Vaccination

As described in sections 3.4 and 3.3, the profile based scheme locates the people with similar contact networks and the sociability driven policy selects the people solely based on their degree. When these two policies are combined, *cutoff degree d* can be considered as a constraint on random walks i.e., a random walk started from an initially infected person visits a node v in PPC network only if the $degree(v) \geq d$. Alternatively, sociability driven policy can also be considered as a post-processing step on the resulting set of nodes from the profile based policy. At any point during the random walk, the next node to be visited is solely based on the vertices adjacent to the current node. When the walk encounters a node v where most of the adjacent nodes of v have degree $< d$, the algorithm spends a lot of time in finding the nodes with degree $\geq d$. Such difficulty is not encountered when the cutoff degree is applied as a

filtering criterion in the post-processing step. Effectiveness of the second approach is thus slightly better than the first approach even though both of these alternatives seem identical.

3.5.2 Contacts + Sociability Based Vaccination

Approaches similar to the ones mentioned above (section 3.5.1) can also be applied to combine contacts based and sociability driven vaccination policies. Contacts based vaccination policy can first be used to get a candidate set of people to vaccinate. Sociability driven policy can then be used to choose people from the candidate set based on the cutoff degree and vaccine. The sociability constraint can also be incorporated in contacts based policy by making the BFS to visit a node only if that node's degree is $\geq d$.

3.6 Location Based Vaccination

Till now, we have presented vaccination policies which are based on the contact network of people. In this section, we propose a policy based on locations. In order to study the people-locations relationship, we make use of *PLA* graph. We first identify the critical locations based on the number of cases reported at each location. Assume that a location, L is connected to P_L number of people (i.e., the degree of L in PLA graph) out of which I_L are infected. We calculate the measure *Infection Ratio* at location L as on day D, $IR_L^D = \frac{I_L}{P_L}$ to quantify each locations infectiousness. A location is declared as *critical* if IR_L^D exceeds a certain threshold value, $Threshold_{IR}$. D is referred to as *Policy Effective Date* (PED). Even though IR_L^D can be calculated every time I_L changes, we calculate IR_L^D at the end of each day for computational efficiency. It is important to note that the cases reported till policy effective day *can not* be prevented. Once the critical locations are identified, people visiting the locations can be vaccinated in various ways. One possible way is to vaccinate all the people who are connected to critical locations through some activity. This method can easily be extended to employ more sophisticated policies. For example, the concept of IR_L^D can be extended to take the type, start time and duration of the activity into consideration.

3.7 Quarantining Locations

In context of many diseases, the specific properties of a location may be conducive to spread the disease (e.g. Cholera and water wells). In such cases, it might be cost-effective to close the place or take any other remedial action at that place instead of vaccinating people. We refer to such an action as *quarantining a location*. Specifically, quarantining a location L refers to a process by which people are kept away from L (either by closing or by other means) so that no individual can get infected at L. Once the locations are identified, remedial measures can be employed to "clean" the location.

Our approach of quarantining locations can be thought of as an extension to the location based vaccinating policy. Locations to be quarantined are identified by analyzing the infection ratios calculated in section 3.6. For a given PED, each location L is tagged with $R_L = (x, y, r)$ where x, y are geographic coordinates of L and r is its infection ratio, IR_L. The resulting dataset, R is then examined to determine the set of locations to be quarantined. We apply hierarchical spatial clustering algorithms on R to form clusters of locations based on geographical distance and also on similarity in infection ratios. The order in which clustering is performed on (x, y) and r attributes define a particular hierarchy. In $(x, y) - r$ clustering, locations are first clustered by the x, y coordinates. The set of locations in each of the resulting cluster are, further, clustered based on r. Similarly we can define $r - (x, y)$ clustering, where the first level of clustering is performed on r. Each of the resulting cluster is, then, spatially segmented to generate the regions. These two types of clustering schemes help in identifying the spatial regions which have similar geographic coordinates and similar infection ratio. Quarantining policy maker can focus on the resulting regions of interest instead of analyzing the locations distributed all over the space. Please note that any vaccinating or containment policy can also be applied on the regions obtained from hierarchical spatial clustering.

4. RESULTS

In this section, we first compare the effectiveness of proposed policies. We then evaluate our strategies under different constraints like limited number of anti-viral drugs and delay in response time.

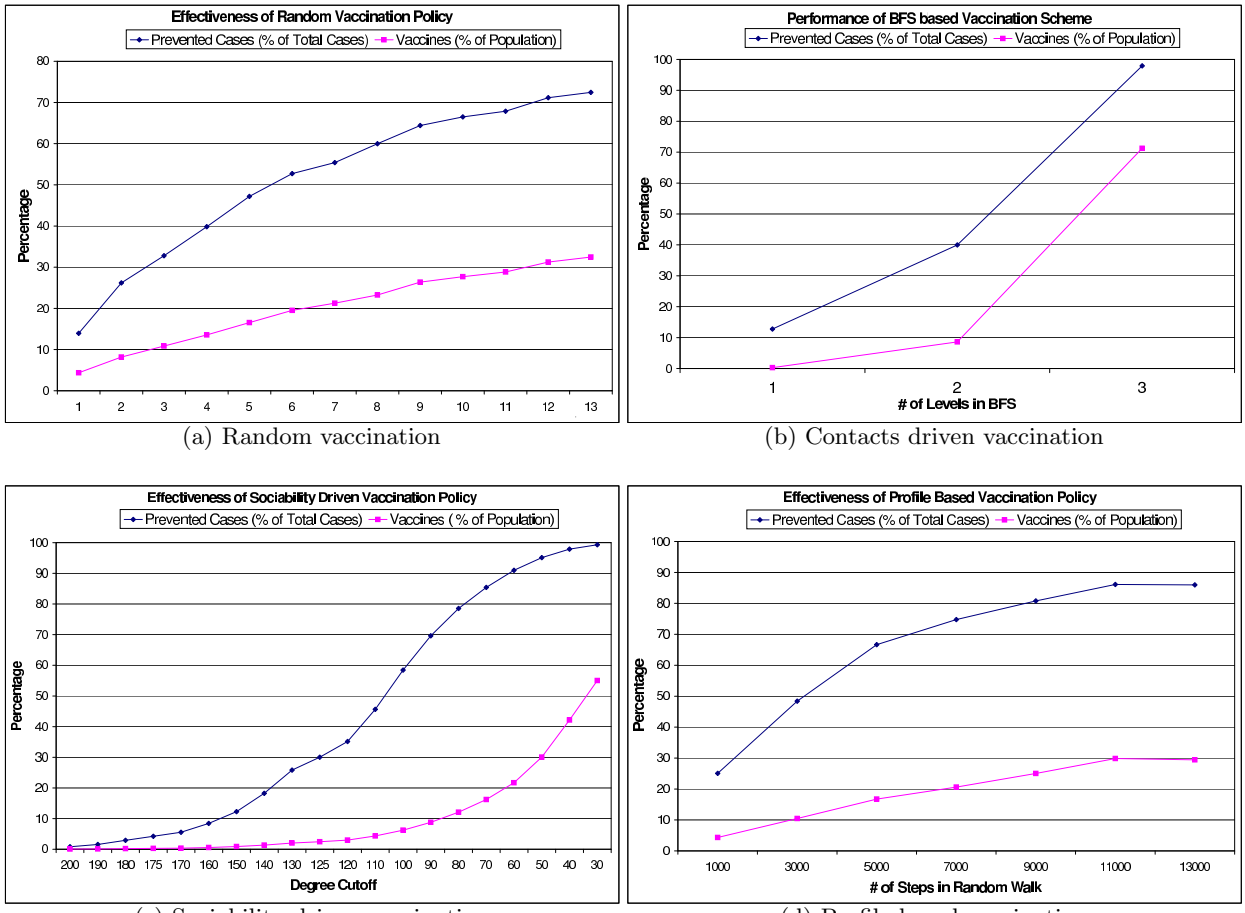
4.1 Effectiveness Measures

We compare the effectiveness of different methods using the measure, *Percentage Prevention*. It is the percentage of people who are prevented from the disease. To calculate this measure, we make use of disease evolution data from EpiSims simulation system [6]. This data provides insight into when, where and from whom a person got infected. Simulation is started at $t = 0$ with 100 initially infected people. When any person is infected during the simulation, system records the time and location at which the person got infected along with the list, L of *already* infected people who are currently in the same location. Assume that a person, P is infected at time T_P during the simulation. With no containment policies, P can infect other people in the contact network. Assume that P infects a person Q at T_Q ($> T_P$). Now, assume that using one of the disease containment strategies, we vaccinate P at time T_V ($T_V < T_P$). Since P is vaccinated, P is prevented from the disease *directly*. But vaccinating P in turn also prevents Q from getting infected because P is no longer infectious. i.e., Q is prevented from the disease *indirectly*. Therefore, vaccinating a person not only prevents that person but can also prevent other people indirectly. *Percentage Prevention* includes total number of people who are prevented from infection both directly and indirectly. Note that for a given person P, if the set L contains more than one individual then P is considered to be prevented (indirectly) only if *all* the people in L are prevented, either directly or indirectly. Another closely related measure is the *cost incurred* by the containment strategy. It is inversely proportional to number of cases prevented (both directly and indirectly) per vaccine.

4.2 Effectiveness of Policies

In every strategy, percentage prevention increases as more people are vaccinated. Random vaccination policy gives the theoretical upper bound on number of vaccines needed to achieve the given prevention rate (Figure 3 (a)). In practice, this strategy will not be effective as it can not take any extra knowledge about the data or the disease into account.

Figure 3 (b) shows how the number of vaccines and percentage prevention changes as we change the number of lev-



(a) Random vaccination

(b) Contacts driven vaccination

(c) Sociability driven vaccination

(d) Profile based vaccination

Figure 3: Effectiveness of various vaccination policies

els of BFS. Exponential increase in number of nodes visited (i.e., number of vaccines) by BFS illustrates shorter path lengths highlighting the small-world properties of the PPC network. Although we prevent 98% of cases at level 3, the cost incurred is very high because we vaccinate 71% of population. From level 2 to level 3, number of cases prevented per vaccine dropped from 162 to 48. This method is very costly and, therefore, may not be very practical. We later show that even with a fixed number of vaccines, it performs poorly when compared to the profile based and sociability driven strategies. Note that, BFS with 1 level will vaccinate the direct contacts of an infected person. Since a person can pass on the infection only to direct contacts, BFS with 1 level should achieve 100% prevention rate. But, this is not true for the given data. This is due to the presence of people who are neither infected from others nor in the set of initially infected people. We attribute them as people who got infected naturally but not from contacts made with an infected person.

Figure 3 (c) illustrates the variation in percentage prevention and number of vaccines given as we vary d. Trend representing the number of vaccines expounds the power-law degree distribution in PPC network. Presence of highly connected nodes (*hubs*) can be seen from the slow increase in the number of vaccines given, initially. Since there are very few people with high degree of contacts, number of vaccines spent increases very slowly in the beginning. A quick increase after the cutoff degree of 100 is due to large number of nodes with smaller degrees of contact. Hence, a smaller cutoff degree leads to higher percentage prevention. Sociability driven prevention strategy offers the lower bound on number of vaccines to be spent for achieving a given percentage of prevention. Prevention of 99.21% is achieved at cutoff degree of 30. This can be achieved by vaccinating *at least* 55% of population.

Effectiveness of the profile based vaccinating policy with varying number of steps is shown in figure 3 (d). We have set the restarting probability to be 0.15 for all our experiments. As we increase the number of steps taken during the walk, number of nodes visited and hence the prevention rate goes up. As more number of nodes are visited, more number of vaccines are used. There is a trade-off between the cost incurred and the percentage prevention achieved. Thus, constraints such as availability of vaccines and other resources should be taken into account when determining the number of steps. It is important to note that the increase in percentage prevention is slow when compared to increase in number of vaccinations. In other words, number of cases prevented per vaccine reduces as the number of steps increases. In practice, profile based strategy might work better than any sociability driven strategy. Because, sociability driven methods require the exact knowledge of contacts of a person and also assume that the contact network is static. In real life, it is very difficult to keep track of exact informa-

tion of contacts as the contact network changes over time. In such cases, we can expect the profile based policies to be more practical than others.

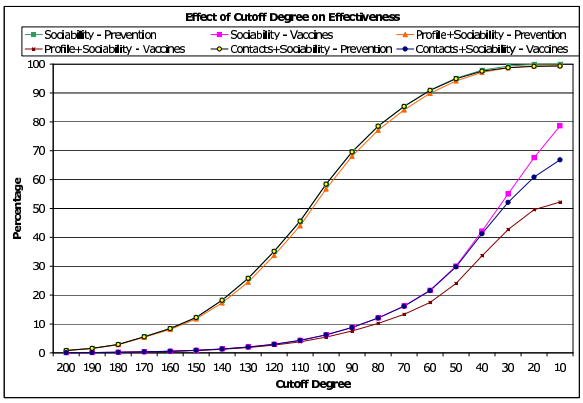


Figure 4: Effect of Degree Cutoff on Effectiveness of Vaccination Policies

The effectiveness (in terms of percentage prevention) of the hybrid strategies described in section 3.5 is higher compared to the performance of the fundamental policies. Figure 4 shows the differences in performance of hybrid strategies and the sociability driven policy (the best among fundamental policies) as the cutoff degree d is varied. Results are obtained using BFS with 3 levels and random walks with $90,000$ steps in case of contacts and profile based policies, respectively. The sociability driven policy is applied as a post-processing step in filtering the candidate set of people obtained from either contacts or profile based policy. When d is very high, effectiveness of vaccinating hubs is very low, as presented in section 3.3. For all levels of d, percentage prevention obtained by all three policies is similar. In fact, percentage prevention curves of sociability driven and contacts+sociability based policies are *mostly* overlapping. This suggests that the set of people with high degree are located within 3 hops from 100 initially infected people. Difference in number of vaccines spent by these policies is very small at high values of d. This difference increases with the decrease in d depicting the cost effectiveness by the hybrid strategies. For example, at degree cutoff of 20, sociability driven policy achieves 99.9% prevention rate by vaccinating 67.6% of population whereas a hybrid strategy with profile based approach gives 99.4% prevention rate by vaccinating just 49.5% of population.

Furthermore, hybrid strategies need to look at a smaller section of population in determining the set of people to vaccine. This is because the contacts or profile based policy is first applied to obtain the candidate set of people and then the sociability driven policy is applied. For example, when $d = 10$, the sociability based policy need to examine every individual to see if their degree is greater than 10. The Contacts+Sociability based policy need to concentrate only on the set of people returned by 3-level BFS, which is just 71% of population (figure 3 (b)). i.e., to achieve 99.9% of prevention rate the sociability based policy examines 100% of population and vaccinates 78.6% of them. The Contacts+Sociability based policy vaccinates examines only 71% of population and vaccinates just 66.8% of population. Therefore, hybrid strategies *examines smaller section* of population and spends *smaller number of vaccines* in achieving comparable prevention rates.

4.3 Location Based Vaccination

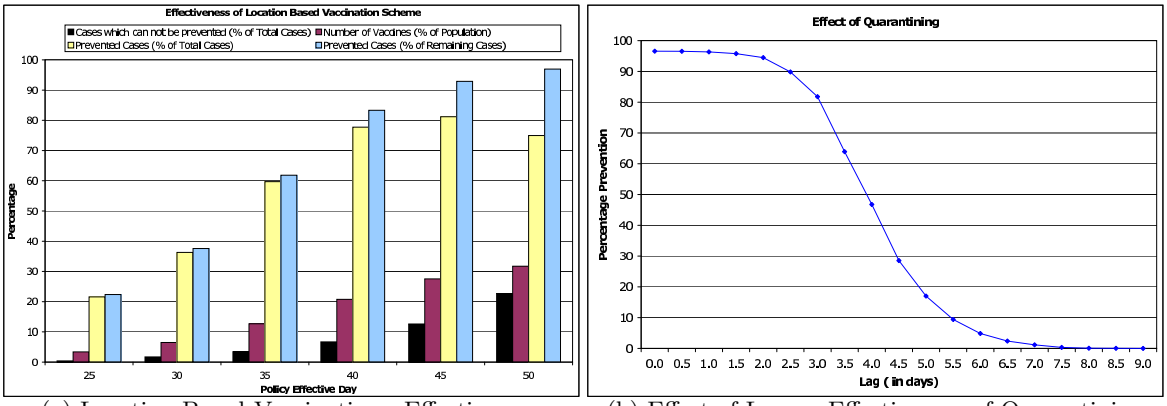
Figure 5 (a) shows the effectiveness of location based vaccination scheme for various values of policy effective day (PED). We have set $Threshold_{IR}$ to 0.01 i.e., if 1% of people connected a location are infected then that location is declared as critical. Let T be the set of people who got infected during the 100 day simulation. And, let S is the set of people who got infected before PED. Hence, people present in S can not be prevented from disease. We define a set R as $T - S$ that represents the set of people who can be prevented. Percentage prevention can be calculated based on both T and R. We refer to prevention rate as a percent of T and R as PP_T and PP_R, respectively. As mentioned earlier, we vaccinate all the people connected to critical locations.

As we increase the PED, number of cases which can not be prevented increases quickly. For example, if we delay the policy for 50 days then almost 23% of the cases can not be prevented. As the time progresses, disease spreads among the people and so the number of locations with infected people increases. Since we vaccinate all the people connected to infected locations, number of vaccines given and, hence, the PP_R increases. Note that the amount of increase in PP_R reduces as the PED increases. But the PP_R does not give the overall effectiveness of the vaccinating policy. To analyze the exact behavior or to compare against other policies, one should use PP_T. As we change PED from 45 to 50, difference between PP_T and PP_R becomes evident. Though the PP_R increases from 92.9% to 96%, PP_T actually decreases from 81.2% to 75%. This is due to the quick increase in number of cases which can not be prevented, from 12.6% to 22.7%.

Figure 6 (a) shows the distribution of infectiousness across various locations. For example, locations in red color have more infected people. We use the $K\text{-}Means$ algorithm for hierarchically clustering the locations and to determine critical regions over space. Other spatial clustering algorithms such as $DBSCAN$ can also be used for clustering the locations.

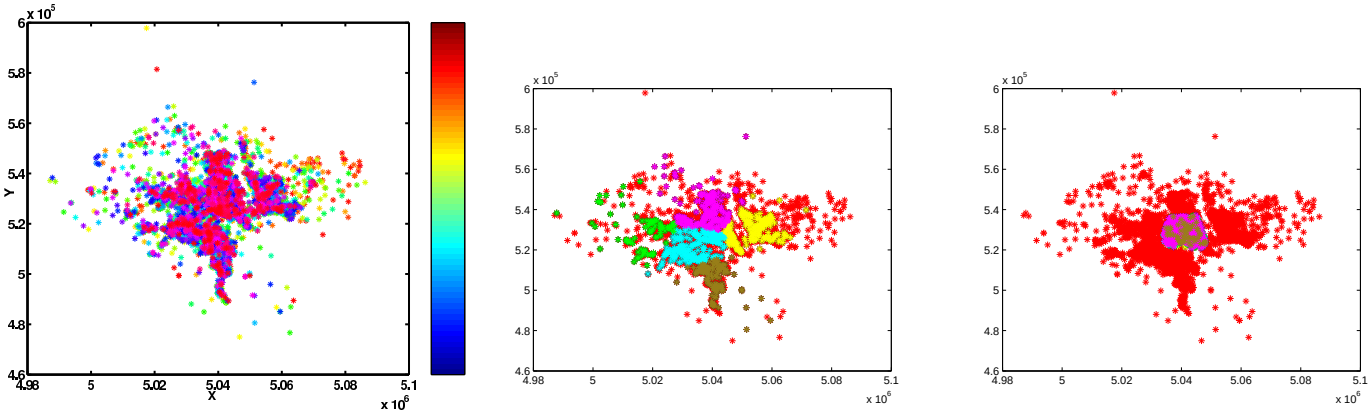
Once the infection ratios are calculated for each location, spatial clustering algorithms can be applied to obtain regions of interest. Figure 6 shows the representative clusters obtained using both $r-(x,y)$ and $(x,y)-r$ clustering methods. In figure 6 (b), locations are first clustered on infection ratio and hence the sub-clusters obtained after (x,y) clustering are distributed all over the space. Whereas in figure 6 (c), (x,y) clustering is done first and hence sub-clusters are concentrated in a partition of space. Resulting clusters are then considered for quarantining based on the cluster centers obtained from clustering based on r. Effectiveness of our clustering based quarantining policy is shown in Figure 7. These results are obtained using a PED of 35 days where the lowest infection ratio obtained is 0.01. Intuitively, as the cutoff ratio is decreased more and more locations gets quarantined and hence the prevention rate goes up. At a cutoff ratio of 0.015, quarantining by $r-(x,y)$ clustering yields 44% of prevention percentage where as quarantining by $(x,y)-r$ clustering gives 59% percentage. Since the lowest r is 0.01, when cutoff ratio is set to 0.01 the effectiveness of quarantining by both clustering methods is identical.

It is worth noting that the prevention rates obtained by quarantining policies are less compared to vaccination poli-



(a) Location Based Vaccination - Effectiveness (b) Effect of Lag on Effectiveness of Quarantining

Figure 5: Effectiveness of Vaccinating and Quarantining policies



(a) Distribution of infectious locations at PED = 35 (b) $r-(x,y)$ Clustering (c) $(x,y)-r$ Clustering

Figure 6: Quarantining by Clustering - Representative Clusters

cies. Due to the delay enforced by PED people who got infected before PED, can not be prevented which partly reduces the prevention percentage. Furthermore, these quarantining strategies are more localized than vaccination policies, which are global in nature. The quarantining policy needs to concentrate only on the set of people adjacent to a quarantined location in the PLA graph. It need not take the entire structure of the PLA graph into consideration. On the contrary, vaccination policies determines the set of people to be vaccinated only by considering the entire structure of the graph. Though the prevention rates of vaccination policies are high, in general, they incur higher costs. Key components of such costs includes the cost incurred in preparation and distribution of vaccines. Whereas implementing a quarantining policies such as closing down a location are simple in nature, easy to implement, and incurs very less cost. Quarantining policies are thus *more easier to implement and cheaper* than vaccination policies which makes them more cost-effective.

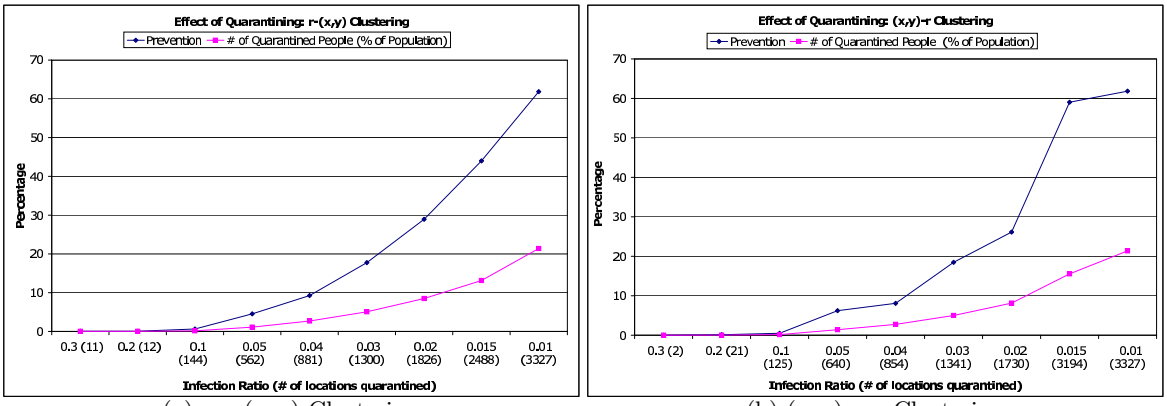
4.4 Effect of Resource Constraints

Anti-viral drugs are often limited in number because of various reasons like mass production cost, inventory cost etc. In this section, we fix the number of vaccines available and evaluate how the percentage prevention changes across different strategies. Such a comparison enables us to analyze the effectiveness and feasibility of various strategies given resource constraints. Figure 8 (a) shows the differences in effectiveness when we fix the number of anti-viral drugs to be used. For all the strategies, total number of prevented cases increases and number of cases prevented per vaccine decreases as the number of vaccines used increases. As we vaccinate larger section of population, difference between their effectiveness decreases. For any given number of vaccines, sociability driven vaccination policy is the clear winner among fundamental policies retaining high levels of effectiveness. Performance of hybrid strategies from the figure further corroborates the argument presented in section 4.2. i.e., hybrid strategies achieves better prevention rates for a fixed number of vaccines. As shown in section 3.2, the reach of BFS is very high but for a fixed number of anti-viral drugs, its effectiveness falls behind the sociability driven and profile based vaccination policies.

4.5 Effect of Delay in Response

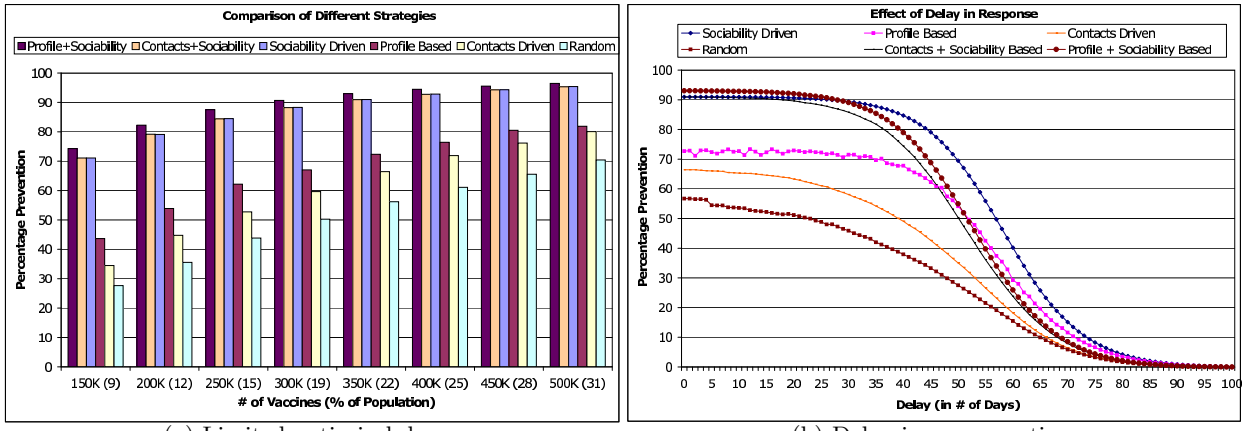
It is not practical to assume that the disease containment policies can be implemented as soon as the first case of the infection is reported. The delay can be due to various constraints like distance between anti-viral drug inventory and the location at which the infected person resides or might be due to late diagnosis. Therefore, it is important to evaluate the tolerance levels of our strategies to the delay in response after the first case has been reported.

Figure 8 (b) shows the effectiveness of different methods



(a) $r - (x, y)$ Clustering (b) $(x, y) - r$ Clustering

Figure 7: Quarantining by Clustering - Effectiveness



(a) Limited anti-viral drugs (b) Delay in response time

Figure 8: Effectiveness under constraints

as a function of reaction time (in terms of days). Number of anti-viral drugs has been fixed at 350, 000 courses for this experiment. Clearly, number of cases prevented goes down as we delay the implementation of containment policies. From the graph, it can be inferred that, in general, a *delay up to* 35 *to* 40 *days* is acceptable for sociability driven and profile based vaccination policies without foregoing significant prevention percentage. For other two policies, number of people prevented from infection continuously decreases as the response time increases. After 40 days, prevention rate decreases at a faster rate especially in sociability driven scheme. Difference in effectiveness between degree and profile based policies decreases as the delay increases. Whereas hybrid strategies can withstand a *delay up to* 30 *days* without loosing too much effectiveness. Though the prevention rate is high initially, it drops at much faster pace compared to sociability driven policy as the delay is increased. This can be explained by examining the reason behind higher prevention rates achieved by hybrid strategies. As noted before, hybrid strategies vaccinates a smaller section of population in achieving a comparable prevention rate i.e., they consider only the *critical* individuals. Therefore, even if a very few number of these critical people gets infected the overall prevention rate gets affected by a large amount. Since the delay in policy implementation affects the number of infected people, hybrid strategies are very sensitive to the delay. Therefore, hybrid strategies are *less robust* compared to sociability driven policy but they are more *cost-effective*.

4.6 Effect of Lag in Quarantining

In order to implement quarantining policies, one has to identify the infected people. Such an identification task is not straight forward as it might take a while for the person to exhibit the symptoms of disease after acquiring it. We define the time interval between the time at which a person is infected and the time at which that person is quarantined as *Lag* in identifying the infection. Lag can also be due to physical constraints such as moving the patients to secluded places etc. When the lag is high, an infected person will have higher chance of propagating the disease before quarantining effecting the performance of quarantining policy. The relation between the lag and the percentage prevention is depicted in figure 5 (b). With the increase in lag, more and more people gets infected by others thereby reducing the percentage prevention.

A person can develop the infection on their own (by some external effect). Therefore the prevention rate is not 100% even if the person is quarantined as soon as that person gets infected (i.e., lag=0 days). In the simulated data used, 3.4% of the reported cases developed the infection by some exter-

nal effects. Please note that a person is quarantined after acquiring the infection where as a person is vaccinated before getting infected. Those 3.4% of cases *can be* prevented from infection by means of vaccination policies. From the figure, it can be inferred that a *lag up to* 2.5 *days* can be tolerated by quarantining policy before effecting the prevention rate.

It is worth noting the difference between *Delay* as defined here and the *Lag* as defined in section 3.7. Delay defines the time at which the policy is implemented and the lag represents the difference in times at which a person is infected and the time at which a person is quarantined. When *delay* = 3 days all people infected before 3^{rd} day *can not* be prevented. If *lag* = 3 days and the a person A gets infected on 10^{th} day of simulation, A would be quarantined on 13^{th} day. Hence, all the people who are infected from A between 10^{th} and 13^{th} *can not* be prevented. But, people who are infected after 13^{th} day from A would be prevented.

5. RELATED WORK

Mining for knowledge in both spatial and temporal dimensions has gained interest in many other application domains like bioinformatics [8], computational fluid dynamics [11] and traffic modeling. Researchers have focused on developing mining algorithms for modeling and uncovering the patterns in spatio-temporal data [15, 16]. Several researchers have also applied the spatio-temporal mining techniques to model and analyze the disease outbreaks. Eubank et. al. [2], [3] have developed disease outbreak models for generating large-scale synthetic data. They also proposed fast algorithms for computing basic structural properties such as clustering coefficients and shortest paths distribution. Ferguson et. al. [4] have proposed strategies to contain the emerging influenza pandemic. Longini et. al. [9] have used the stochastic epidemic simulations to investigate the effectiveness of targeted antiviral prophylaxis to contain influenza. Hartke [7] studied and proposed various mathematical models for disease spread which are motivated from classical models such as *firefighter model* [5, 10]. We presented some of the initial ideas on this problem in Tatikonda et. al. [14].

6. CONCLUSIONS

In this paper, we proposed and evaluated several effective containment policies to curb the disease from spreading. We demonstrated how the fundamental policies can be combined to devise even more effective hybrid strategies. Among all the proposed vaccination policies, the hybrid of sociability driven and profile based approach is the most effective policy. We have also developed and examined the quarantining policies by leveraging hierarchical clustering algorithms. We evaluated the proposed containment policies under various practical constraints such as delay in implementation of policies, limited number of anti-viral drugs, and the lag in quarantining. We showed that the implementation of vaccination policies can be delayed up to 35 days and a lag of up to 2.5 days can be tolerated by quarantining policies without foregoing the effectiveness.

In practice, the social contact network is highly dynamic in nature and hence would be changing over time. Furthermore, exact model of the contact network of a given person is extremely difficult to build. In light of such complex scenarios, application of strategies like sociability based policy can prove to be difficult. In future, We would like to explore and develop containment (both vaccination and quarantining) policies in context of such complex models.

7. ACKNOWLEDGMENTS

This work is supported by NSF grants CAREER-IIS-0347662, RI-CNS-0403342, and NGS-CNS-0406386. The authors are thankful to Professor Madhav Marathe, Virgina Tech for providing the datasets. The authors would also like to thank Professor Naren Ramakrishnan of Virginia Tech, Professor Chris Bailey-Kellogg of Dartmouth College, Sitaram Asur, Duygu Ucar and Greg Buehrer of The Ohio State University for their useful comments and suggestions.

8. REFERENCES

[1] R. Albert and A. Barabasi. Statistical mechanics of complex networks. In *Review Modern Physics*, 2002.

[2] S. Eubank, H. Guclu, V. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling disease outbreaks in realistic urban social networks. In *Nature*, volume 429, pages 180–184, 2004.

[3] S. Eubank, V. Anil Kumar, M. Marathe, A. Srinivasan, and N. Wang. Structural and algorithmic aspects of massive social networks. In *Symposium on Discrete Algorithms*, 2004.

[4] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in southeast asia. In *Nature*, volume 437, pages 209–214, 2005.

[5] S. Finbow, A. King, G. MacGillivray, and R. Rizzi. The Firefighter Problem For Graphs of Maximum Degree Three, 2004.

[6] Synthetic Data Products for Societal Infrastructures and NDSSL-TR-06-006 Proto-Populations: Data Set 1.0. Network dynamics and simulation science laboratory, virginia polytechnic institute and state university.

[7] Stephen G. Hartke. *Graph-theoretic Models of Spread and Competition*. PhD thesis, Rutgers University, 2004.

[8] J. Hu, X. Shen, Y. Shao, C. Bystroff, and M.J. Zaki. Mining Protein Contact Maps. *2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2002)*, 2002.

[9] I. M. Longini, M. E. Halloran, A. Nizam, and Y. Yang. Containing pandemic influenza with antiviral agents. In *Americal journal of epidemiology*, 2004.

[10] G. MacGillivray and P. Wang. On the Firefighter Problem. *J. Combin. Math. Combin. Comput*, 47:83–96, 2003.

[11] S. Mehta, S. Parthasarathy, and R. Machiraju. Visual Exploration of Spatio-temporal Relationships for Scientific Data. *IEEE Symposium on Visual Analytics Science and Technology*, 2006.

[12] M. Newman. The structure and function of complex networks. In *SIAM Review*, 2003.

[13] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Fifth IEEE International Conference on Data Mining*, 2005.

[14] S. Tatikonda, S. Mehta, and S. Parthasarathy. Containment Policies for Transmissible Diseases. *Spatial Data Mining Workshop held with SIAM Conference on Data Mining*, 2006.

[15] H. Yang, S. Parthasarathy, and S. Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. *Conference on Knowledge Discovery in Data*, pages 716–721, 2005.

[16] H. Yang, S. Parthasarathy, and S. Mehta. Mining Spatial Object Patterns in Scientific Data. *International Joint Conference of Artificial Intelligence (IJCAI)*, 2005.