Hidden Cluster Detection for Infectious Disease Control and Quarantine Management

Yain-Whar Si, Kan-Ion Leong, Robert P. Biuk-Aghai, and Simon Fong

Faculty of Science and Technology, University of Macau
{fstasp,ma46511,robertb,ccfong}@umac.mo

Abstract. Infectious diseases that are caused by pathogenic microorganisms can spread fast and far, from one person to another, directly or indirectly. Prompt quarantining of the infected from the rest, coupled with contact tracing, has been an effective measure to encounter outbreaks. However, urban life and international travel make containment difficult. Furthermore, the length of incubation periods of some contagious diseases like SARS enable infected passengers to elude health screenings before first symptoms appear and thus to carry the disease further. Detecting and visualizing contact–tracing networks, and immediately identifying the routes of infection, are thus important. We apply information visualization and hidden cluster detection for finding cliques of potentially infected people during incubation. Preemptive control and early quarantine are hence possible by our method. Our prototype Infectious Disease Detection and Quarantine Management System (IDDQMS), which can identify and trace clusters of infection by mining patients' history, is introduced in this paper.

Keywords: Infectious Disease, Cluster Detection, Contact Tracing, SARS, Health Care Information System.

1 Introduction

Infectious disease prevention and control is one of the most important research areas in the 21st century. The SARS (Severe Acute Respiratory Syndrome) outbreak in 2003 and recent world–wide avian flu infections have contributed to the urgent need to search for efficient methods for prevention and control of highly infectious diseases. During the SARS outbreak of 2003, there were 8096 probable cases and 774 deaths reported in 29 countries including China, Taiwan, Hong Kong, Canada, Singapore, and Vietnam [1]. According to a WHO (World Health Organization) report [2], three influenza pandemics were recorded in history: "Spanish influenza" in 1918, "Asian influenza" in 1957, and "Hong Kong influenza" in 1968. The 1918 Spanish influenza pandemic killed an estimated 40 to 50 million people worldwide.
When a patient is diagnosed with a highly infectious disease or is suspected of being infected, it is crucial to locate the source of infection (also known as ground zero or super spreader) and to identify future probable cases within a short time. Such measures aim to limit the secondary spread during the outbreak. Although numerous

M.L. Huang et al. (eds.), *Visual Information Communication*,
DOI 10.1007/978-1-4419-0312-9_10, © Springer Science+Business Media, LLC 2010

researches have been undertaken to find effective vaccines and cure, relatively less attention was devoted to devising software programs which are capable of assisting medical planners with rapid decision making in tracing the source of an outbreak based on information obtained from patients. In addition, currently available health–care information systems from hospitals are not designed to deal with the complex and delicate task of managing quarantine for potentially large numbers of suspected cases.

Given this background, this research aims to develop a decision support system which can be used to locate the source of an outbreak by mining clusters and communities from the patients' past activities (testimonies) using techniques from infectious disease control, information visualization, and database management systems. Our developed system also allows categorizing individuals who are likely to be infected into different risk groups whereby members from the high risk group are subjected to immediate isolation whereas members from the low risk group are monitored periodically. The system also includes visualization capabilities for medical experts to analyze the outbreak as the clusters of infection depicted as inter–connected graphs, and graphs for different risk groups. The system is also designed to provide medical planners with quarantine management capabilities such as administering the data of (1) persons currently being quarantined, (2) persons being monitored for symptoms but not in quarantine, and (3) persons who were recently identified for possible infection but have yet to be contacted.

This paper is structured as follows. We briefly review related work in section 2. An introduction of IDDQMS is given in section 3. The extraction of detailed epidemiological history and data analysis is described in section 4. The identification of index case and tracing the source of outbreak is described in section 5. Identifying clusters of infections is detailed in section 6. The need for a quarantine management system is discussed in section 7. We summarize our ideas in section 8.

2 Related Work

Recent work on infectious disease informatics projects also deploys information visualization methods. In the West Nile Virus and Botulism Portal project [3], spatial temporal data sets are overlaid onto geographic maps including land information and demographic data. The Portal System also allows web–enabled access to data related to West Nile Virus and Botulism [3].

SAHANA [4] is an open source disaster management system which is designed to assist in tracking of missing victims and managing of relief supplies. SAHANA is a web–enabled system powered by a set of modules including volunteer coordination, camp management, and supply chain management. The main difference between our work and SAHANA is that our system focuses on discovering hidden clusters of infection using data mining methods whereas the main objective of SAHANA is to provide web–enabled coordination functions to different organizations during a disaster.

Similar methods have been applied in the context of crime data mining. Chen et al. [5, 6] have proposed an automated Social Network Analysis approach to analyze

structural properties of criminal networks and to investigate patterns of interaction. Chen et al. use the concept space approach to create networks automatically and deploy the complete–link algorithm for partitioning networks. In [5], MDS algorithm [7] is used to visualize criminal networks.

3 Overview

IDDQMS (see Figure 1) consists of four modules; information extraction, data analysis, hidden cluster detection, and quarantine management. The current system was developed in the Java programming language, JUNG [8], and MySQL database management system [9]. JUNG is a JAVA based open–source software library for the modeling, analysis, and visualization of data.

Once an outbreak is reported, one of the crucial steps is to collect information from the patients and transform them into an appropriate format for data mining tasks. During the course of an isolated outbreak, the load of extracted epidemiological and virological information is generally low. However, the amount of data and its growth may become unmanageable in the event of a major outbreak. In the following section, we describe the information extraction tasks and propose an event–based framework for recording patients' activities.

4 Information Extraction

Recording detailed epidemiological history of relevant travel or contact history was identified as one of the key factors in identifying potential SARS cases [10]. When a patient is diagnosed with a highly infectious disease such as SARS, emergency medical planners may need to decide whom they should quarantine after analyzing the patient's history spanning the entire incubation period (e.g. approximately 10 days for the case of SARS). The information extraction phase can be carried out by a medical practitioner to record patient's name, identification number (or Passport number), address, family members, colleagues or classmates, friends, workplace address, and activities happened during the last 10 days. The extracted information is then structured into a set of events which include four primary attributes: (1) duration, (2) persons contacted during that period, (3) location visited (e.g. enclosed spaces such as offices, air planes, etc.), and (4) food taken. A snapshot of information extraction menus designed for the SARS infectious disease for entering visiting activities is depicted in Figure 2. A separate menu is also designed to view all the visiting activities of the persons in the infectious disease as depicted in Figure 3. As a note, the data in Figures 2 and 3 are simulated data which are used for demonstration only. In a similar way, information on persons contacted and food intake are recorded in the other information extraction menus in our system. Since the proposed system is intended for emergency medical planners, we assume that the confidentiality of the extracted information is guaranteed in a way similar to other health care information systems.

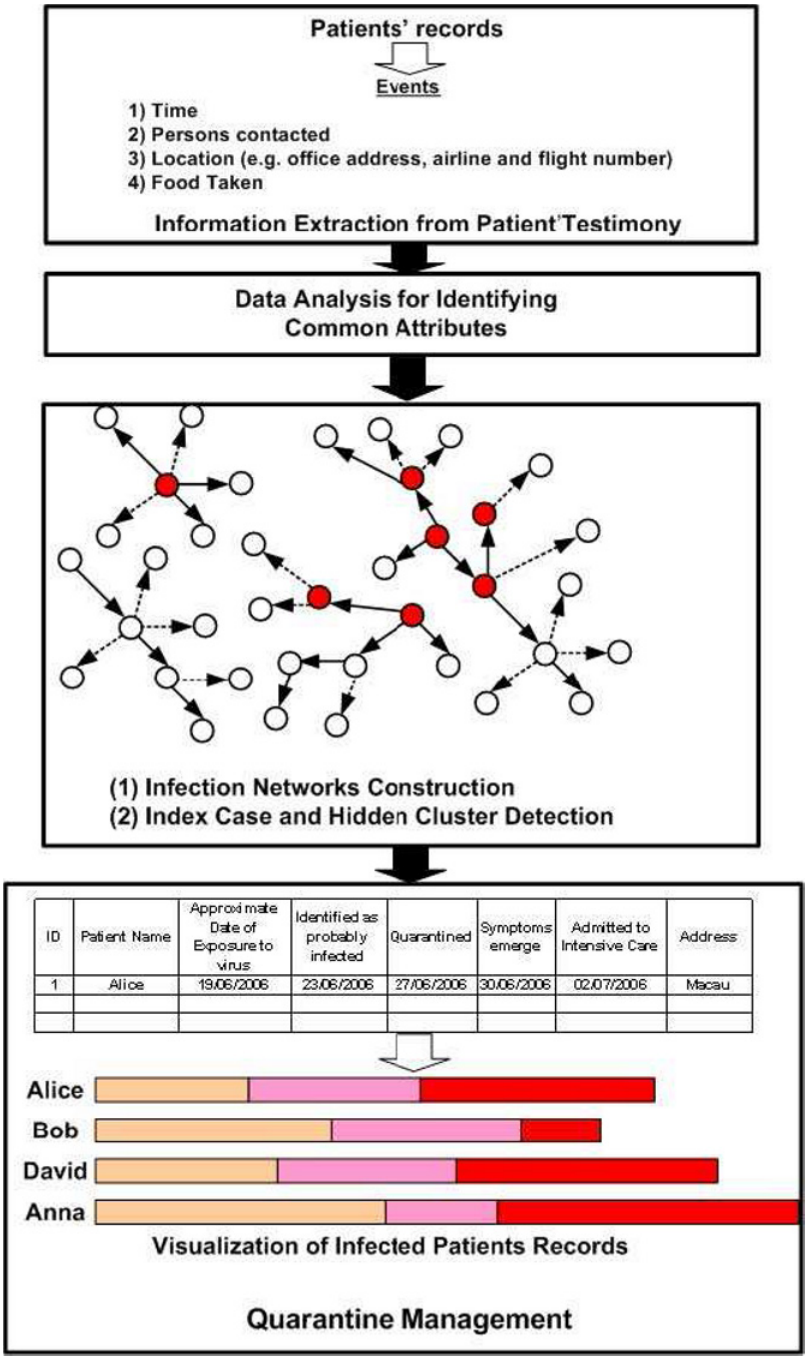


Fig. 1. Overview of the IDDQMS system.

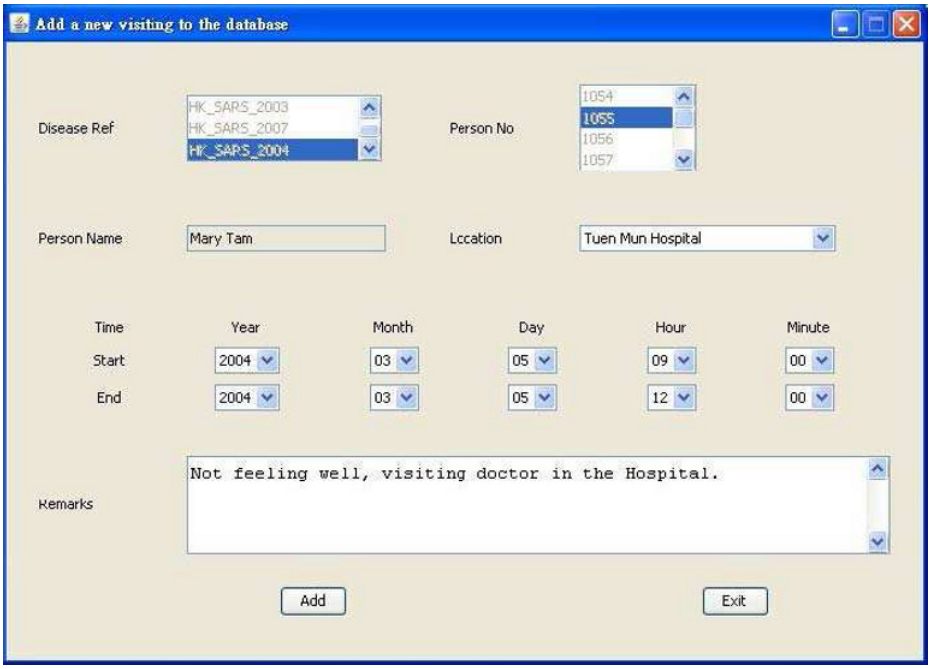
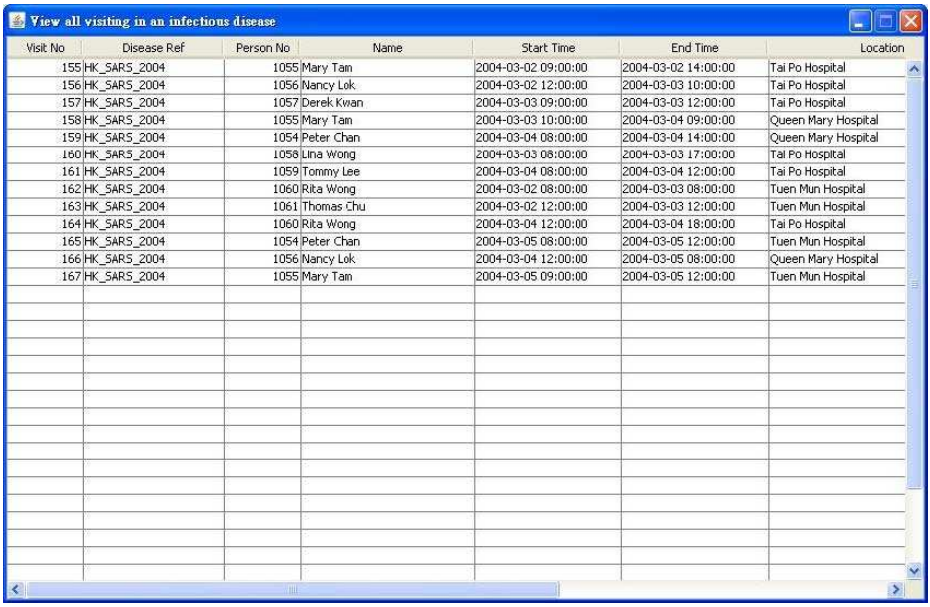


Fig. 2. Entering patient's visiting activities.



Visit No	Disease Ref	Person No	Name	Start Time	End Time	Location
155	HK_SARS_2004	1055	Mary Tam	2004-03-02 09:00:00	2004-03-02 14:00:00	Tai Po Hospital
156	HK_SARS_2004	1056	Nancy Lok	2004-03-02 12:00:00	2004-03-03 10:00:00	Tai Po Hospital
157	HK_SARS_2004	1057	Derek Kwan	2004-03-03 09:00:00	2004-03-03 12:00:00	Tai Po Hospital
158	HK_SARS_2004	1055	Mary Tam	2004-03-03 10:00:00	2004-03-04 09:00:00	Queen Mary Hospital
159	HK_SARS_2004	1054	Peter Chan	2004-03-04 08:00:00	2004-03-04 14:00:00	Queen Mary Hospital
160	HK_SARS_2004	1058	Lina Wong	2004-03-03 08:00:00	2004-03-03 17:00:00	Tai Po Hospital
161	HK_SARS_2004	1059	Tommy Lee	2004-03-04 08:00:00	2004-03-04 12:00:00	Tai Po Hospital
162	HK_SARS_2004	1060	Rita Wong	2004-03-02 08:00:00	2004-03-03 08:00:00	Tuen Mun Hospital
163	HK_SARS_2004	1061	Thomas Chu	2004-03-02 12:00:00	2004-03-03 12:00:00	Tuen Mun Hospital
164	HK_SARS_2004	1060	Rita Wong	2004-03-04 12:00:00	2004-03-04 18:00:00	Tai Po Hospital
165	HK_SARS_2004	1054	Peter Chan	2004-03-05 08:00:00	2004-03-05 12:00:00	Tuen Mun Hospital
166	HK_SARS_2004	1056	Nancy Lok	2004-03-04 12:00:00	2004-03-05 08:00:00	Queen Mary Hospital
167	HK_SARS_2004	1055	Mary Tam	2004-03-05 09:00:00	2004-03-05 12:00:00	Tuen Mun Hospital

Fig. 3. Viewing all the visiting activities of the persons of the infectious disease; the entry in Figure 2 corresponds to the last record in current figure.

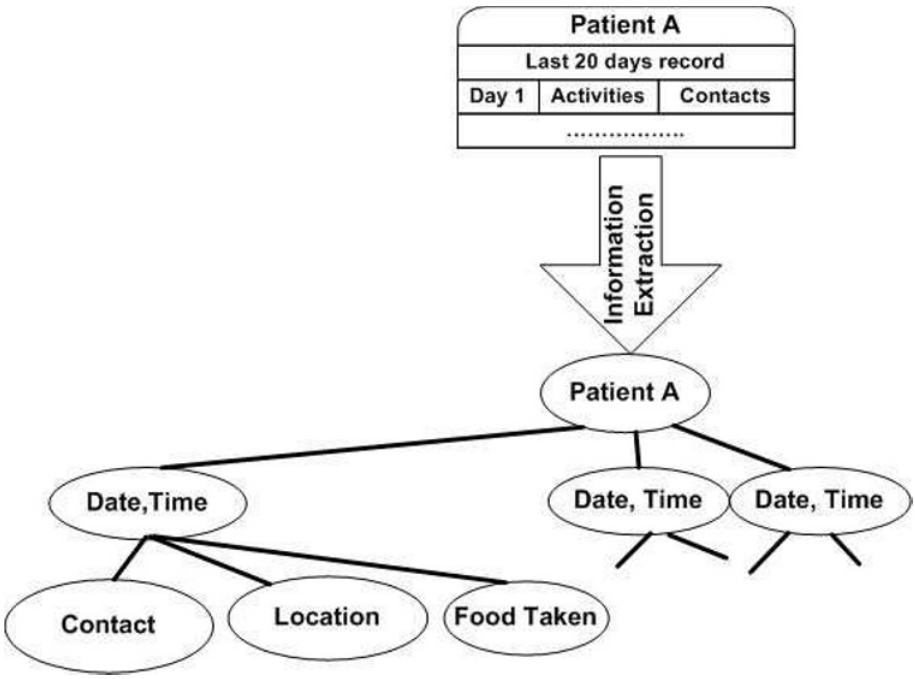


Fig. 4. Extracted episodes from patient's records.

Based on the extracted information, past activities of the patients are constructed as networks of events carrying respective time–stamps (see Figure 4). These networks are then used for visualization of the patients' past activities as well as for mining clusters of infection and identifying index cases.

Based on the extracted information, emergency medical planners may decide on the status of individuals:

- who need to be quarantined immediately as they pose the highest risk (e.g. family member and colleagues who have had direct contact with the patient),
- who should be put under surveillance as they may have been in the same enclosed space, and
- who should be put under notice as they may live in nearby apartments or may have had contact with the patient's family members (secondary infection).

5 Data Analysis for Identifying Common Attributes

The task of tracing an outbreak can be performed efficiently and promptly in combating any infectious disease by comparing elements extracted from the patients' episodes. A similar approach has been used by investigators to detect deceptive information in criminal records [11]. An example of identifying similar elements from

extracted patients' record is depicted in Figure 5. Patient 1 and patient 2 from Figure 5 are suspected of having the disease and they both have traveled on the same train (number 15) on 2 April 2006. In that case, one of them may have been carrying the disease at that time or they may have contracted the disease from a third person. Medical experts may then further examine whether patient 1 or patient 2 have contracted the disease from other sources (e.g. contaminated food) or if there was indeed a third person who may have infected them and traveled on the same train. If contraction from contaminated food is less likely, then it is crucial to identify the third person before he/she can spread the disease to the healthy population.

Tracing the source of infection from patients' episodes is a complex and tedious task as the information provided by the patients may contain errors and inconsistencies. For instance, the addresses extracted from two patients may not be identical and hence such overlapping depicted in Figure 5 may not be obvious. In that case, methods from identifying partial matching of strings or locating patterns in strings is employed to find any overlapping and users of the system will be prompted when similarity among elements are identified in the event data. A screen shot captured from the data analysis module of IDDQMS is depicted in Figure 6. In this module, patients' activities are visualized as segmented horizontal bars arranged along a time axis, and overlapping activities are grouped using separate colors. The scale of the horizontal bars is adjusted based on the longest duration from the recorded past activities of the patients being monitored. For the sake of simplicity, locations which appear in more than one patient's past activities are extracted and grouped for visual identification. The data analysis module in Figure 6 provides effective visualization capabilities to medical experts in identifying overlapping locations, and thus points of potential contamination among patients.

However, in the event of a large scale outbreak, pattern matching of extracted information from event data alone may not be sufficient for identifying hidden clusters of infection. For instance, the family members of patient 1 from Figure 5 may have contracted the virus and in turn, they may have passed the disease to other persons before being quarantined. In that case, comparing elements from extracted information alone may not be sufficient to identify probable clusters of infection.

6 Identifying clusters of infection

In [12], we proposed the principles and the algorithms for cluster detection in IDDQMS. The practice of cluster detection is generally referred to as contact tracing in the medical context. The proposed approach utilizes the visiting records are collected and input into the system. For finding clusters of cases and the infection tree of an infectious disease during its outbreak, three main algorithms were presented in [12]: an algorithm for detecting clusters of cases, an algorithm for detecting the infection tree of a cluster, and an algorithm for constructing the infection tree of an outbreak. In worst case scenarios, the space and time complexity of these algorithms is $O(n^2)$.

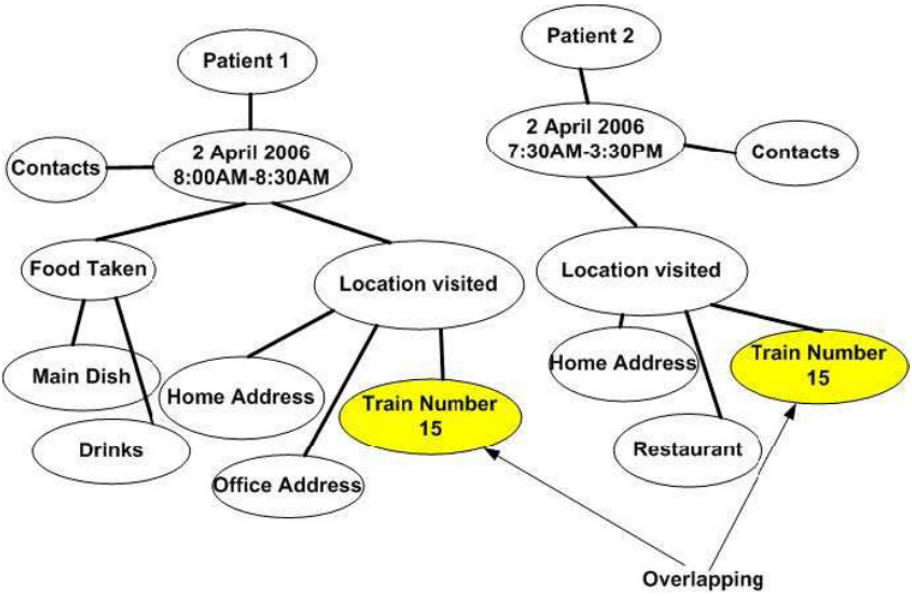


Fig. 5. Example of overlapping in patients' activities.

Based on the theoretical framework proposed in [12], in this paper we tested IDDQMS with a case study of the year 2003 Hong Kong SARS outbreak. In this case study, the real–life information, such as incubation period, onset date, confirmed date, etc, of the patients from the related medical institutions, were collected from [13–16]. To the best of our knowledge, this set of information is adequate to outline the 2003 Hong Kong SARS outbreak in its initial stage. We have used the collected information to fit in the algorithms and have deduced the infection trees (or cluster trees) of the clusters of the outbreak in its initial stage as shown in Figure 7. By following the transmission routes of the patients, we connect the separate individual infection trees, to construct a whole infection tree in Figure 8.

Here we give a brief account of the outbreak as a summary from [13–16]. In Figure 7, *n1* who is a professor from Guangzhou, China is the index case (or index patient) of the outbreak. He came to Hong Kong to celebrate the wedding of his nephew. After a family dinner, he checked into the Metropole Hotel. During the night, the professor felt feverish. The next day, the professor felt so ill that he was admitted to the nearest hospital, Kwong Wah Hospital (KWH). In the hospital, he infected a nurse, *n4*.

In Figure 7, epidemiological investigation confirmed that *n2* was the index case of St Pauls Hospital (SPH) in which *n2* infected three nurses while *n3* was the index case of Prince of Wales Hospital (PWH) in which *n3* infected three doctors and three nurses. In PWH, it then caused a SARS outbreak among the healthcare workers (HCWs). Since then, SARS cases were prevalent in the community as the diseases continued to propagate progressively on their down lines. Further epidemiological investigation showed that the infection of both *n2* and *n3* took place in the Metropole Hotel while the professor was staying there, so the clusters were linked together as shown in Figure 8.

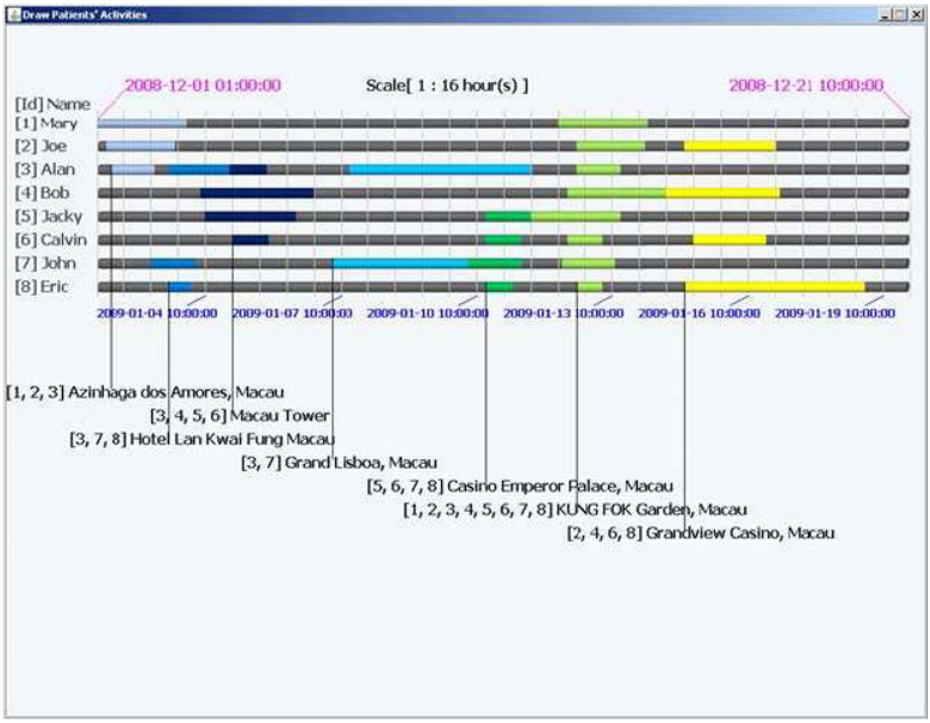


Fig. 6. Tracing/Locating source of infection.

7 Quarantine Management

During the year 2003 SARS epidemic in Beijing, more than a thousand confirmed cases and hundreds of suspected cases were identified. In the event of a major outbreak, the total number of patients who should be quarantined, monitored, or tracked may vastly increase as the disease may spread to densely populated urban areas. Therefore, managing and analyzing infection networks from all these cases and extraction of event data can easily overwhelm an existing information system. In this research, we develop an integrated system which not only includes data mining functionalities as described in the previous section but also contains tailor–made programs for managing patient records and their quarantine status. By using our integrated system, medical workers may decide appropriate quarantine actions such as admitting to intensive care units, isolating in purpose–built medical wards, and monitoring probable cases. These decision making procedures are depicted in Figure 8.

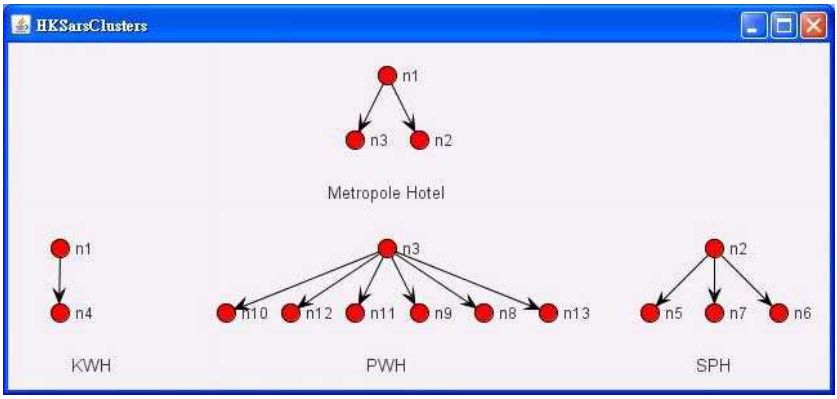


Fig. 7. The clusters shown as infection trees of the 2003 Hong Kong SARS outbreak in its initial stage visualized by JUNG..

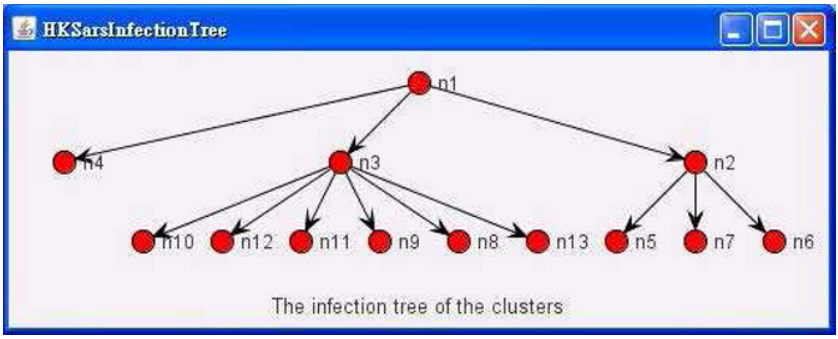


Fig. 8. The whole infection tree constructed from the infection trees in Fig 6 visualized by JUNG.

When a person is reported to have made contact with an infectious disease (through one of the possible mediums such as another person carrying the disease, a common location or contact with an object), he/she will need to be classified as either a confirmed case or a suspected case by observing his/her symptoms and recent medical records. Any suspected case has to be further classified either for quarantine or for monitoring procedure. This situation is depicted in the top part of Figure 8. The quarantine management decision making process described in Figure 8 is applied to every patient who is suspected of being exposed to the disease.

In addition, IDDQMS also provides visualization of incubation period, quarantine, and treatment status for a selected number of patients using scaled vectors. Such capabilities are useful when a number of patients within a cluster need to be compared based on their respective durations. Five time points are used in Figure 9, time of infection (confirmed), quarantined, monitored, onset of symptom, and end date of treatment.

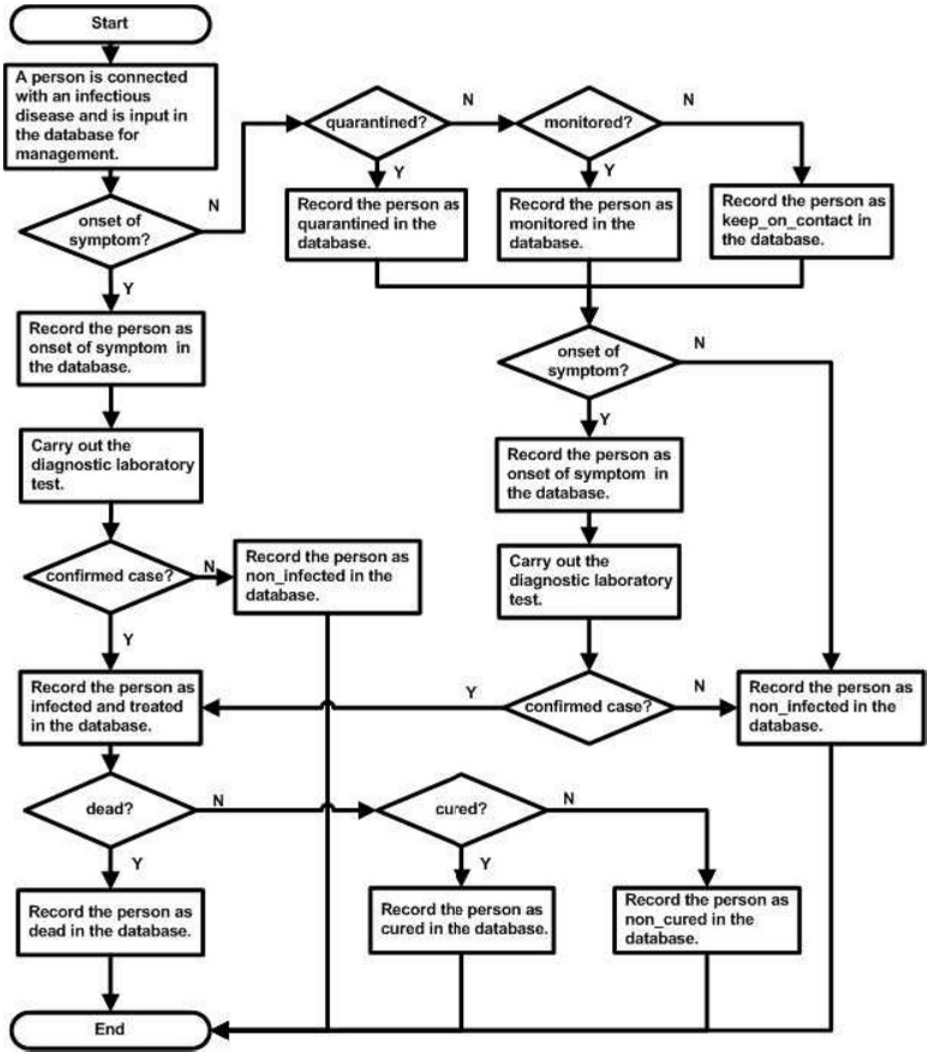


Fig. 9. Deciding patients' status for quarantine

8 Summary

In this paper, we have described our novel prototype system on Infectious Disease Detection and Quarantine Management, which can be used to identify and trace clusters of infection by mining patients' history. The system utilizes techniques from infectious disease control, information visualization, and database management systems.

As for future work, we are planning to generalize the current contact tracing algorithms and to apply them to detecting tuberculosis infections and ongoing global spread of the novel influenza A (H1N1) virus (also referred to as "swine flu" early on). Tuberculosis poses a significant threat to Macau. In 2005, Macau had the highest population density in the world at 18,428 people per square kilometer. According to the World Health Organization, there are 70 cases of tuberculosis for every 100,000 citizens in Macau compared to 50 in Europe and 39 in the Americas [17]. Since the former Portuguese enclave hosts millions of visitors annually, the task of identifying the source of any potential outbreak is even more challenging. Detecting tuberculosis clusters is far more complicated since the incubation period from exposure to developing a positive skin test is approximately 2 to 12 weeks, which is much longer than in the case of SARS. Therefore, the detection model for tuberculosis requires tracking of persons' activities for a much longer duration.

We are currently analyzing the transmission pattern of novel influenza A (H1N1) virus. Similar to seasonal human influenza viruses, novel influenza A (H1N1) virus is thought to spread from person to person through large–particle respiratory droplet transmission. It is currently estimated that the incubation period range from 1 to 7 days [18]. We hope that current model for detecting SARS can be extended/revised for novel influenza A (H1N1) virus infection when more information on the estimated duration of viral shedding becomes available.

We are also planning to extend current visualization functions for tracing/locating source of infections (depicted in Figure 6) and patients' status (depicted in Figure 10). For instance, segmented horizontal bars used for patients' activities can be extended with alignment functions [19] for displaying first occurrence of overlapped past locations. In addition, we are planning to provide visualization capabilities for the contact tracing algorithms proposed in [12]. For instance, one of the algorithms calculates the most relevant date of infection between two cases based on the date of mean of incubation period and the date of maximum transmission efficiency. These dates as well as the meeting dates of patients' can be presented chronologically on one or more horizontal time–lines.

We are also currently testing information extraction software LingPipe [20] to extend our information extraction phase by employing the Named Entity Recognition (NER) algorithm [21] to locate and classify atomic elements. For instance, NER can be used to arrange these elements into predefined categories such as the names of persons, organizations, locations, times, quantities, and numbers.

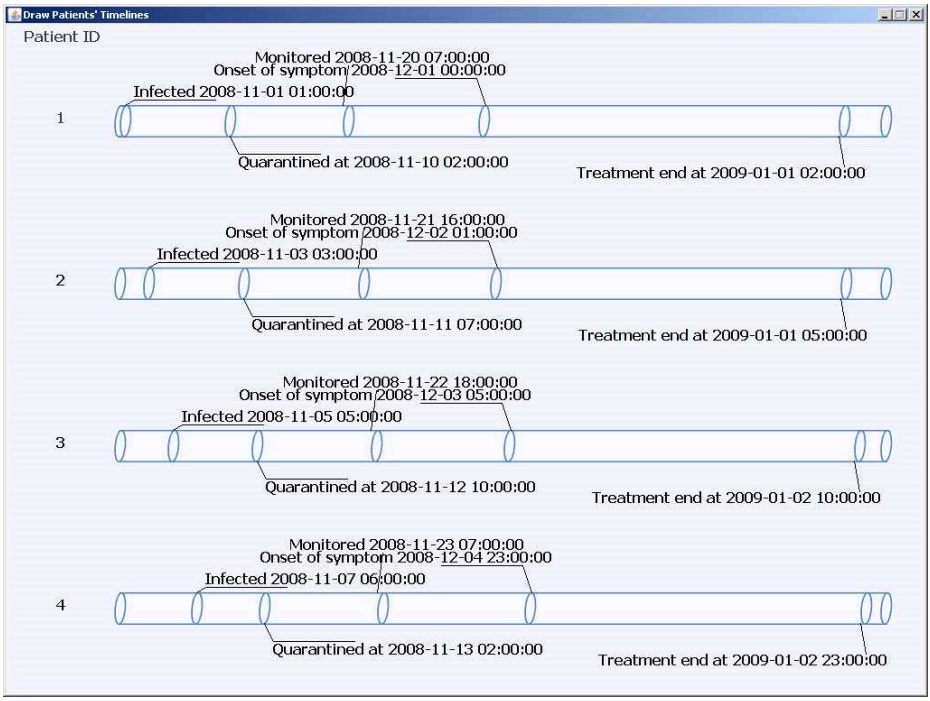


Fig. 10. Visualization of patients' status

Acknowledgments. This work is funded by the University of Macau Research Grant "Hidden Cluster Detection and Visual Data Mining Framework for Infectious Disease Control and Quarantine Management". The authors thank Dr. Lam Chong, Coordinator for Control of Communicable Disease and Surveillance of Diseases, CDC, Government of Special Administrative Region Health Bureau, Macau, for his insightful comments on the project.

References

1. World Health Organization: Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. http://www.who.int/csr/sars/country/ table2004_04_21/en/
2. World Health Organization: Ten things you need to know about pandemic influenza. http://www.who.int/csr/disease/influenza/pandemic10things/en/
3. Zeng, D., Chen, H., Tseng, C., Larson, C.A., Eidson, M., Gotham, I., Lynch, C., Ascher, M.: Towards a national infectious disease information infrastructure: a case study in West Nile virus and botulism. In: 2004 Annual National Conference on Digital Government Research, Seattle, WA, IEEE (2004) 1–10
4. SAHANA, Free and open source disaster management system. http://www. sahana.lk/
5. Xu, J.J., Chen, H.: Criminal network analysis and visualization. Communications of ACM 48(6) (2005) 100–107

6. Xu, J.J., Chen, H.: CrimeNet explorer: a framework for criminal network knowledge discovery. ACM Transactions on Information Systems 23(2) (2005) 201–226
7. Torgerson, W.S.: Multidimensional scaling: I. Theory and method. Psychometrika 17 (1952)
8. JUNG, Java Universal Network/Graph Framework. http://jung.sourceforge. net/
9. MySQL homepage. http://www.mysql.com/
10. Goddard, N.L., Delpecha, V.C., Watsona, J.M., Reganb, M., , A., N.: Lessons learned from SARS: The experience of the health protection agency, England. Public Health 120(1) (2006) 27–32
11. Wang, G., Chen, H., Atabakhsh, H.: Automatically detecting deceptive criminal identities. Communication of ACM 47(3) (2004) 70–76
12. Leong, K.I., Si, Y.W., Biuk-Aghai, R.P., Fong, S.: Contact tracing in health-care digital ecosystems for infectious disease control and quarantine management. In: Third IEEE International Conference on Digital Ecosystems and Technologies (DEST 2009), IEEE (2009) 210–215
13. Tsang, K.W., Ho, P.L., Ooi, G.C., Yee, W.K., Wang, T., Chan-Yeung, M., Lam, W.K., Seto, W.H., Yam, L.Y., Cheung, T.M., Wong, P.C., Lam, B., Ip, M.S., Chan, J., Yuen, K.Y., Lai, K.N.: A cluster of cases of severe acute respiratory syndrome in Hong Kong. The New England Journal of Medicine 348(20) (2003) 1977–85
14. Abraham, T.: Twenty–First Century Plague -The Story of SARS. Johns Hopkins University Press (2005)
15. The SARS Expert Committee of Hong Kong: Report of the SARS expert committee -SARS in hong kong: from experience to action. http://www.sars-expertcom. gov.hk/english/reports/reports.html (2003)
16. The Hong Kong Legislative Council: Report of the select committee to inquire into the handling of the severe acute respiratory syndrome outbreak by the government and the hospital authority. http://www.legco.gov.hk/yr03-04/english/sc/sc_ sars/reports/sars_rpt.htm (July 2004)
17. Huxtable, N.: Population density hinders fight against TB. Macau Daily Times (28 June 2008)
18. Centers for Disease Control and Prevention: Interim guidance for clinicians on identifying and caring for patients with Swine–origin Influenza A (H1N1) virus infection. http://www.cdc.gov/h1n1flu/identifyingpatients.htm
19. Wang, T.D., Plaisant, C., Quinn, A.J., Stanchak, R., Murphy, S., Shneiderman, B.: Aligning temporal data by sentinel events: discovering patterns in electronic health records. In: The twenty-sixth annual SIGCHI conference on Human factors in computing systems, CHI '08, New York, NY, USA, ACM (2008) 457–466
20. LingPipe. http://www.alias-i.com/lingpipe/
21. Cohen, W.W., Arawagi, S.: Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. In: Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, New York, NY, ACM (2004) 89–98