Inferring Missing Infectious Links: A Case Study using 2003 Beijing SARS Outbreak Data

Xiaolong Zheng[1], Daniel Zeng[2,1], Aaron Sun[2], Yuan Luo[1], Quanyi Wang[3], Feiyue Wang[1]

[1] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China

[2] Department of Management Information Systems, The University of Arizona, USA

[3] Beijing Center for Disease Control and Prevention, China

Abstract

Missing data are a major practical challenge to infectious disease informatics research and practice. In particular, despite the best data collection efforts, information concerning contact-based infections can be significantly under-reported in epidemiological surveys and interviews. In this paper, we experiment with a number of computational techniques to infer missing infectious links, using the 2003 Beijing SARS outbreak data as the test case. We conclude that a Bayesian network-based approach outperforms other benchmark methods and can provide a viable solution to help alleviate the missing data challenge.

Keywords: SARS, Infectious Link, Missing Link, Transmission Characteristics

1. Introduction

SARS outbreaks have already been extensively analyzed [1-5]. Existing work has focused on identifying SARS' spreading dynamics and offering quantitative assessment of the epidemic potential of SARS, and the effectiveness of various control measures. From a research perspective, microscopic analyses studying infectious pathways and linkage among patients are of high value [6], yet suffer from the missing data challenge. In particular, when collecting and analyzing real-world infectious disease data, one is always faced with the problem that many infectious links are missing. An analysis with partial information concerning infectious links may lead to biased observations and misleading policy evaluation. One approach to potentially mitigate this problem is to inferring missing infectious links and mixing these inferred links with the observed ones for analysis purposes. This paper is motivated to compare various computational approaches to infer missing infectious links.

Our research reported in this paper is heavily relying on a Beijing SARS dataset, collected by Beijing Center for Disease Control and Prevention (CDC) during the 2003 outbreak. We first analyze transmission characteristics of the Beijing SARS outbreak. Based on our empirical findings, we attempt to infer missing infectious links based on various epidemiological features and conduct a comparative study of various computational approaches as to their inference power and accuracy.

2. 2003 Beijing SARS Outbreak Data

During the outbreak of 2003 Beijing SARS, potential cases were reported by hospitals to the Beijing CDC, which initiated epidemiologic investigations. Data sources included case report forms, epidemiologic investigation forms, and other investigation records at Beijing CDC, which recorded each patient's detail demographic information and symptoms at onset to hospitalization to identify infection sources and events of probable relevance to transmission. We have investigated extensively these report forms to obtain valuable information on infectious pathways between pairs of SARS patients. An infectious pair consists of two SARS patients between whom there exists an infectious link. Our study is based on records concerning 697 patients with 482 identifiable infectious pairs, covering the outbreak period from March 4, 2003 to May 13, 2003. The information used in our research includes age, occupation, reporting district, location (home address), and the date of onset and hospitalization.

In the next section, we investigate the transmission characteristics of Beijing SARS to gain more empirical insights. We then report our work on evaluating computational approaches for inferring missing infectious links through experiments and simulations.

3. SARS Transmission Characteristics

Based on the recorded SARS dataset described above, in this section, we mainly focus on analyzing various SARS transmission characteristics, including the impact of age difference, occupation, reporting district, onset interval, hospitalization interval, reporting interval, and transmission distance distributions.

Age difference, occupation, and reporting district distributions are shown in Fig. 1. Among the 482 infectious pairs, there are 165 pairs whose age differences are less than 10 (0-9), approximately 34.2% of all infectious pairs, as shown in Fig. 1(a). The occupation distribution (Fig. 1 (b)) indicates that not all occupations have equal probabilities to be infected with the SARS virus. The x-axis corresponds to occupations. In 21 government-classified occupations, the top-5 susceptible occupations are retiree, military personnel, governmental employees, unemployed, and working adults. The accumulated frequency of these 5 occupations is 55.4% higher than the other 16 occupations combined. Fig. 1(c) illustrates the reporting district distribution. The x-axis represents 14 reporting administrative districts. From this figure, we observe that the top-3 districts are Chaoyang, Haidian, and Dongcheng, which are labeled with 1, 2 and 3 respectively. These three districts have the accumulated frequency of 56%.

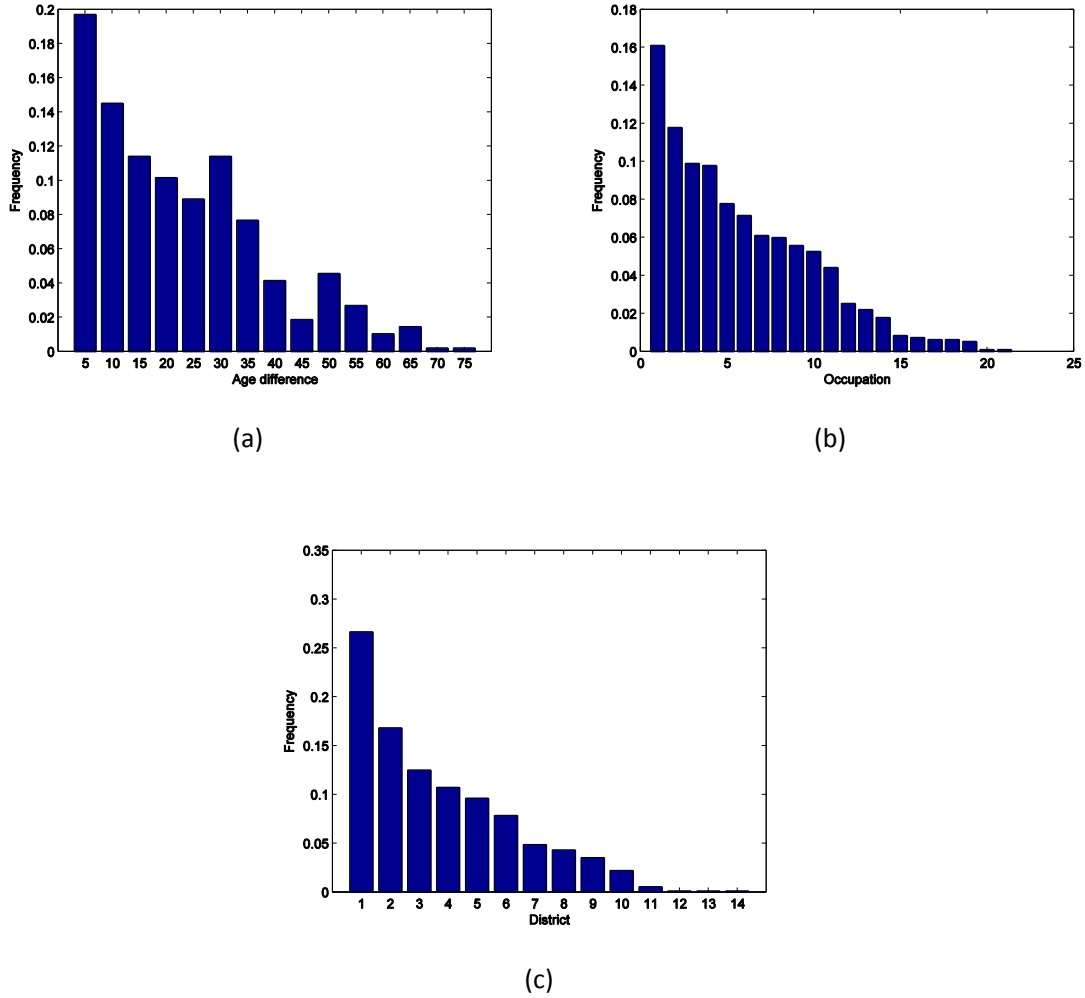
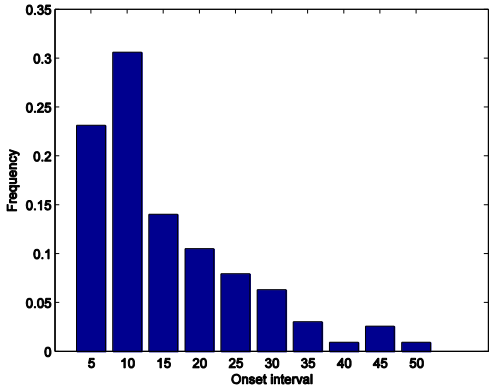
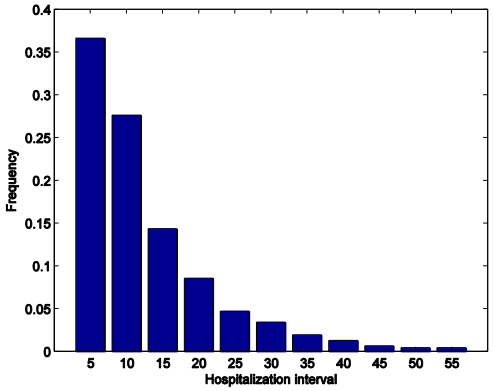


Fig.1. Demographic characteristics. (a) Distribution of age difference. (b) Occupation distribution. (c) Reported district distribution.

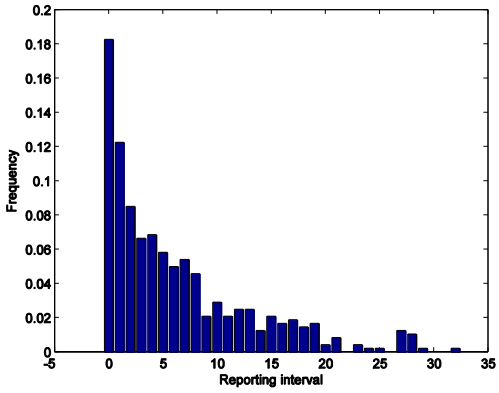
We analyze the distributions of onset, hospitalization, and reporting intervals, as shown in Fig. 2. For date of onset, 428 of 482 infectious pairs had been recorded. Among these 428 pairs, there are 230 infectious pairs whose onset interval value lies in the range of 0 to 9, accounting for 53.7 % of all the pairs. The distribution of onset interval has a mean value of 12.2 days. For the date of hospitalization, there are 468 hospitalization date records. Of these, the hospitalization interval of 300 pairs is less than 10 days, approximately 64.1% of all pairs. The distribution of hospitalization intervals has a mean value of 9.6 days. For reporting date, there are 363 pairs whose reporting intervals are less than 10, approximately accounting for 75.3 % of all 482 recorded pairs. The distribution of reporting intervals has a mean value of 6.3 days. This shows that shortening the intervals will be critical to the general population by restricting the infectious period before patients are placed in quarantine.



(a)



(b)



(c)

Fig. 2 (a) The onset interval distribution. (b) The hospitalization interval distribution. (c) The reporting interval distribution.

According to the home addresses of 404 infectious pairs, we have calculated the transmission distance of an infectious pair *i*. The transmission distance d_i is defined as follows:

$$d_i = \sqrt[2]{x_i^2 + y_i^2}$$

where x_i and y_i represent the longitude and latitude differences of pair *i* respectively. The statistical results are displayed in Fig. 3. We note that 65 cases involve roommates, 136 cases involve patients living within 0 to 0.1, and 308 less than 0.2. It is evident that most of infectious pairs lived geographically close to each other.

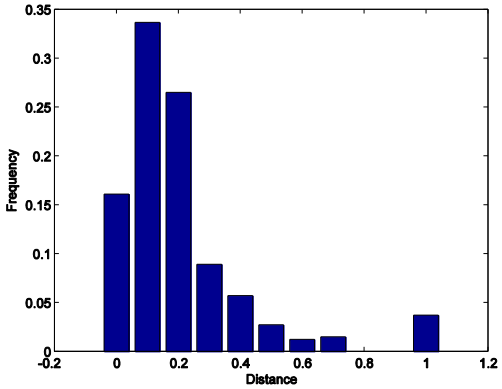


Fig.3 The transmission distance distribution

4. Inferring Missing Infectious Links

We have experimented with three imputation methods that are capable of inferring missing infectious links: Bayesian Network (Bayesnet), Logistic Regression (Logistic), and NBTree. All these methods are classical prediction models with wide applications. The prediction results can be classified into four types: true positives, true negatives, false positives and false negatives. We denote by N_{tp} the number of true positives, N_{fp} the number of false positives, N_{fn} the number of false negatives, and N_{tn} the number of true negatives. The standard metrics to evaluate the performance of these prediction models include precision *P*, recall *R*, and F1-measure $F1$, which are defined as:

$$A = \frac{N_{tp} + N_{tn}}{N_{tp} + N_{fp} + N_{tn} + N_{fn}} \quad (1)$$

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}} \quad (2)$$

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}} \quad (3)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4)$$

We have conducted an experimental study using the real-world Beijing SARS infection data to evaluate these prediction methods. In our experiment, we used 447 infectious links as positive data and randomly selected 555 non-infectious links as negative data. Table 2 shows the prediction results of these models for comparison purposes. From this table, we can observe that Bayesnet achieves highest accuracy, precision, recall and F1-measure. Applying the trained Bayesnet model, we were able to infer 213 links from all possible 83700 potential links. Epidemiological analyses using these inferred links seem to deliver findings consistent with expectations (reported elsewhere).

Measure	Bayesnet	Logistic	NBTree
A	**76.6635 %**	67.9846 %	71.2633 %
P	**0.754**	0.665	0.693
R	**0.739**	0.627	0.685
F1	**0.746**	0.645	0.689

Table 2. Performance of Three Classification Algorithms

5. Concluding Remarks

In this report, we report an empirical analysis of 2003 Beijing SARS outbreak. To deal with the significant problem of missing useful infectious link information, we have computationally evaluated several classification methods to predict the missing infectious links. Our findings indicate that a model based on Bayesian network seems to deliver satisfactory performance. Our current work is focused on further improving the performance of the missing link inference mechanism and deriving at a more general framework that can help choose the appropriate method based on the characteristics of the infectious disease under study and the available epidemiological dataset.

Acknowledgements

We would like to thank Ping Yan, Jian Ma, Zhidong Cao, Su Li, Xiaoli Wu, and Hao Lu, for useful discussions and helpful suggestions. This work was supported by the National Natural Science Foundation of China under Grants 60573078 and 60621001, by the Chinese Academy of Sciences Grants 2F05NO1, 2F07C012, and F08N03, by the Ministry of Science and Technology under Grant 2006AA010106, and by the Ministry of Health under Grant 2009ZX10004-315. The second author wishes to acknowledge support from the U.S. National Science Foundation through Grants IIS-0527563, IIS-0428241, and IIS-0839990, and U.S. DHS through Grant 2008-ST-061-BS0002.

References

1. Dye, C. and N. Gay, *Modeling the SARS Epidemic.* Science, 2003. **300**(5627): p. 1884-1885.
2. Liu, Y.-l., *The comparison research of Singapore and Taiwan's government SARS epidemic situation crisis management.*, in *Political Science*. 2005. p. 98.
3. Riley, S., et al., *Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions.* Science, 2003. **300**(5627): p. 1961-1966.
4. Small, M., P. Shi, and C.K. Tse, *Plausible Models for Propagation of the SARS Virus.* IEICE transactions on fundamentals of electronics, communications and computer sciences, 2004. **E87-A**(9): p. 2379-2386.
5. Wang, J.F., et al., *Data-driven exploration of 'spatial pattern-time process-driving forces' associations of SARS epidemic in Beijing, China.* Journal of Public Health, 2008. **30**(3): p. 234-244.
6. Zheng, X., et al., *Network-Based Analysis of Beijing SARS Data*, in *Proceedings of BioSecure 2008*. p. 64-73.