Network-Based Analysis of Beijing SARS Data

Xiaolong Zheng[1], Daniel Zeng[1,2], Aaron Sun[2], Yuan Luo[1], Quanyi Wang[3], and 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. In this paper, we analyze Beijing SARS data using methods developed from the complex network analysis literature. Three kinds of SARS-related networks were constructed and analyzed, including the patient contact network, the weighted location (district) network, and the weighted occupation network. We demonstrate that a network-based data analysis framework can help evaluate various control strategies. For instance, in the case of SARS, a general randomized immunization control strategy may not be effective. Instead, a strategy that focuses on nodes (e.g., patients, locations, or occupations) with high degree and strength may lead to more effective outbreak control and management.

Keywords: SARS, Complex network analysis, Weighted networks.

1 Introduction

Severe Acute Respiratory Syndrome (SARS) was first found in the Guangdong Province of China in November, 2002. During its 2003 outbreak, 8,098 confirmed cases were reported in more than 30 countries within a very short period of time [1]. Among them, 2,521 cases were reported in Beijing, representing close to one third of the entire world-wide infected population.

The SARS outbreak prompted a world-wide public health response and has had a dramatic impact on the Chinese public health system as to infectious disease prevention, outbreak detection, and response. From a research perspective, significant efforts from both public health and related fields including but not limited to various subareas of informatics and computer-based modeling, have been devoted to studying the evolution and transmission patterns of SARS for future prevention and treatment purposes.

The SARS literature from the perspective of infectious disease informatics has also been growing [1, 2, 3, 4, 5, 6]. For example, several control measures have been proposed to control the outbreak of the SARS epidemic [3, 4]. Spatial analysis of SARS cases has been explored recently to reveal the associations between various related epidemic determinants [1]. Some authors have developed network-based mathematical models to analyze the transmission patterns of the SARS outbreak and to predict the outbreak diversity [2, 6]. Despite the significance and importance of using real-world SARS data to validate and evaluate

these modeling efforts, however, very limited work has been done from an empirical analysis perspective, partially due to the difficulty in accessing pertinent epidemiological data.

Our research aims to bridge some of the existing gaps in the empirical analysis line of work and to better connect the complex network analysis literature with infectious disease informatics practice. In this reported research, we used the Beijing SARS data provided by the Beijing Center for Disease Control. By modeling patients, locations (districts in Beijing), and patient occupations as nodes, respectively, and treating contacts or infections as edges, we have constructed and analyzed three kinds of SARS-related networks: the patient contact network, the district network, and the occupation network. In Section 2, we provide a brief introduction to the data and the network-based analysis methods used in our research. Section 3 presents findings based on the patient contact network. Sections 4 and 5 report findings based on the weighted district and occupation networks, respectively. We conclude the paper in Section 6 by discussing ongoing and future research.

2   Data and Analysis Methods

Our Beijing SARS data were collected from an extensive survey of 624 confirmed SARS patients from 14 administrative districts in Beijing, covering the period from March 10, 2003 to May 13, 2003. These patients worked in 21 categories of occupations. We have followed previous studies (e.g., [2, 6]) to define an "infectious link" pointing from patient A to patient B, if it is highly likely that A transmitted the SARS virus to B. In total, 447 such infectious links were identified.

In our analysis, we first constructed a patient contact network based on infectious links as typical in existing epidemiological studies, and analyzed this network. However, with SARS being a unique and highly contagious airborne epidemic disease, personal contacts uncovered in the patient surveys or interviews alone may not provide sufficient information to fully explain the transmission patterns. As such, in an exploratory attempt, we also constructed two additional networks: a location/district network and an occupation network to further illustrate the spreading of the SARS epidemic in various parts of Beijing and among different occupational categories. We study these two networks as "weighted networks," with the weight w_{ij} defined over an directed edge from node i to j given as the total number of the infectious links from i to j. Further, we study node "strength" s_i defined as $s_i = \sum_{j=1}^{N} w_{ij}$ for node i, where N is the total number of nodes in the network [7]. This strength measure can be indicative of the ability to spread the disease from a given node.

In the next section, we investigate the topological properties of the patient contact network and its evolution pattern. We then discuss findings based on the weighted district and occupation networks.
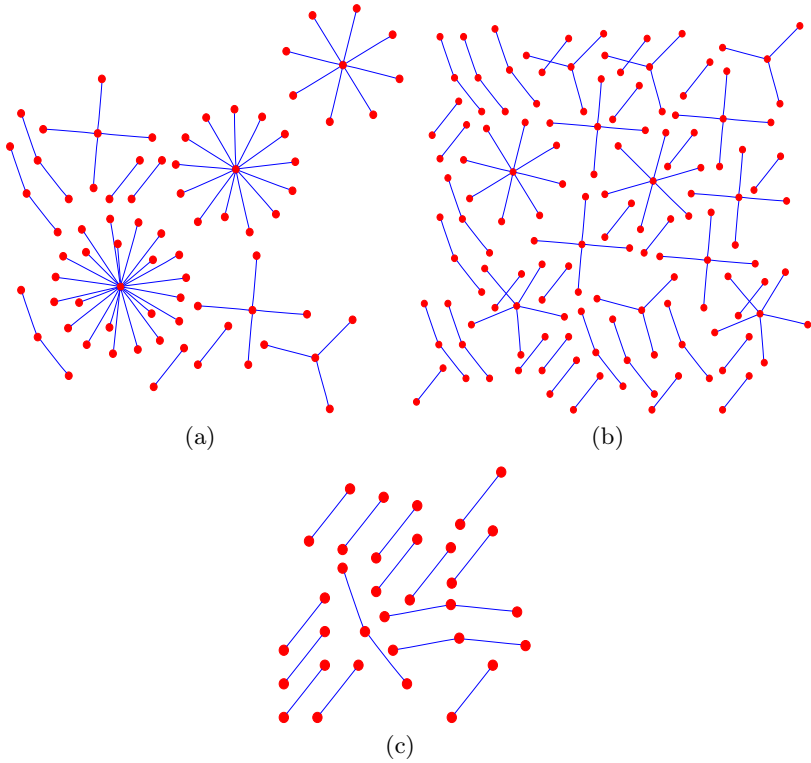
3   Patient Contact Network

Patient contact networks can provide useful information concerning disease transmission and have been studied in the existing literature in various scenarios [8, 9, 10]. Using the Beijing SARS data described in Section 2, we first study the degree distribution of the SARS contact network and then investigate its temporal evolution.

3.1   Degree Distribution

It is well-known in the complex network analysis literature that network representations of a large number of real systems can be characterized by a node degree distribution with a power-law tail [11]. This is of particular importance in epidemiology since in this case the expected reproductive number may be unbounded [12]. In epidemiology, the reproductive number is defined as the number of secondary infections generated by one patient. This concept plays a key role in understanding the dynamic process of epidemics and in evaluating impact of control measures on the spread of infection [13]. Fig. 1 shows that the SARS contact network also follows a power-law distribution. The blue line corresponds to a power-law tail $P(k) \sim k^{-\gamma}$ with $\gamma = 2.8076$. Not surprisingly, this result shows that the SARS infectious network is a scale-free network, with the implication that the expected reproductive number can be unbounded. A public health implication of this finding is that the traditional disease control approach based on random immunization (which has been shown to be effective in many epidemic outbreaks [8]) may not be effective (unless, of course, the entire population can be treated), because untreated hubs, albeit small in number, can still lead to rapid and large-scale infections [8]. Instead, an alternative control method targeting at containing highly connected nodes can be much more effective.
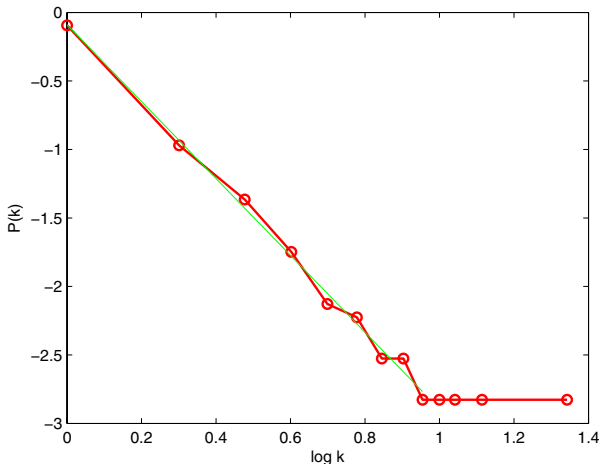
**Fig. 1.** Degree distribution of SARS patient contact network

Fig. 2. SARS patient contact network in (a)–Week 2, (b)–Week 5, and (c)–Week 8

3.2   Network Evolution

It has been pointed out that in different phases, the transmission of epidemics may exhibit different patterns [2, 12]. Since the records in our nine-week SARS dataset are timestamped, we are able to observe the evolution of the SARS patient contact network over time. We plot three weekly "snapshots" of the contact network in Fig. 2.

The node degree in the contact network can be used to measure the node's disease spreading ability [2]. Fig. 3 (a) and (b) plot the average and maximum degrees for 9 consecutive weeks, respectively. We notice that in the first two weeks, the contact network has a relatively high average and maximum degree. These measures start to decrease with time after Week 5. This decrease can be attributed (at least partially) to several strong control measures implemented by the government after April 14, 2003 (which is in Week 5).

A connected component of the contact network is a set of nodes in which each node is connected to at least one edge. Connected components can be used to demonstrate the extent to which an epidemic can spread within a population [14]. We define the component ratio as the number of connected components divided by the total number of nodes in the network. From Fig. 4, we observe
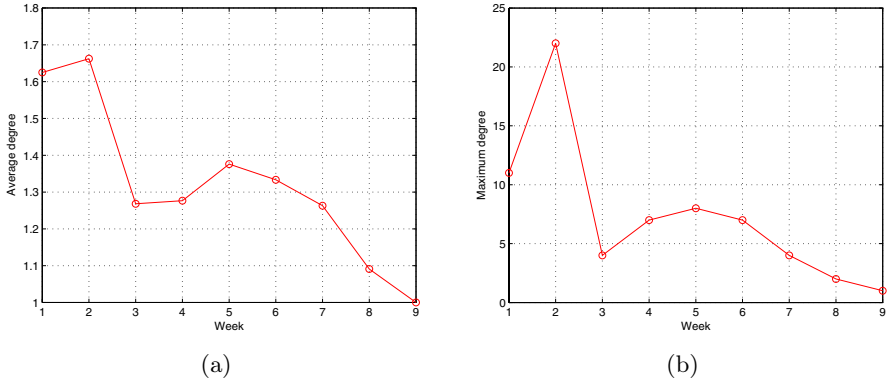
Fig. 3. Network evolution. (a)–average degree and (b)–maximum degree.

that in the first two weeks, the ratio is relatively small. After some fluctuations in the next two weeks, it starts to increase gradually. Part of this observation is due to the fact that, during the first few weeks, most SARS patients were misdiagnosed as having tuberculosis [15]. The isolation and quarantine controls were not enforced for these patients during this initial period of outbreak. After April 14, 2008, as strong control measures were taken, the epidemic was brought under control.

4   Weighted District Network

The patient contact network analyzed in the previous section can provide insights as to SARS transmission patterns among patients. However, for SARS, personal contact information available does not provide sufficient explanation for the underlying transmission patterns of this epidemic (partially due to the incomplete nature of contact information acquired through patient surveys or interviews). Geographical information is also crucial to gain a better understanding of the epidemic [16].

Fig. 5 plots the weighted district network (WDN). We analyzed the infection transmission patterns through the WDN. The results on the cumulative weight distribution are shown in Fig. 6 (a). As we can observe, the cumulative weight distribution follows a strongly right-skewed distribution, indicating a high degree of heterogeneity in the WDN.

To better understand the WDN, we define $s_d(k_d)$ as the average strength of nodes with degree k_d. Theoretically, if $s_d(k_d)$ and k_d are uncorrelated, then $s_d(k_d) \sim k_d^\alpha$ with $\alpha = 1$. In this case, weights cannot provide any additional information than degrees [17]. Our analysis shows that the observed $s_d(k_d)$ increases with k_d as $s_d(k_d) \sim k_d^\alpha$ with the exponent $\alpha = 1.8775$. The findings are plotted in Fig. 6 (b). Table 1 displays the top five district strengths. These results indicate that the strengths of nodes are strongly correlated to degrees in the WDN.
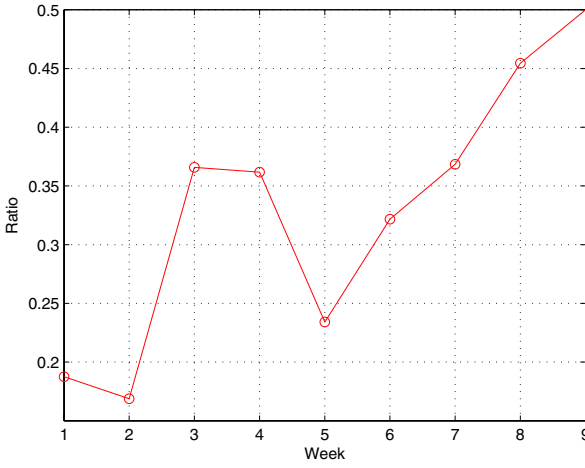
Fig. 4. Changes in network component ratio over time

Table 1. Top five district (node) strengths

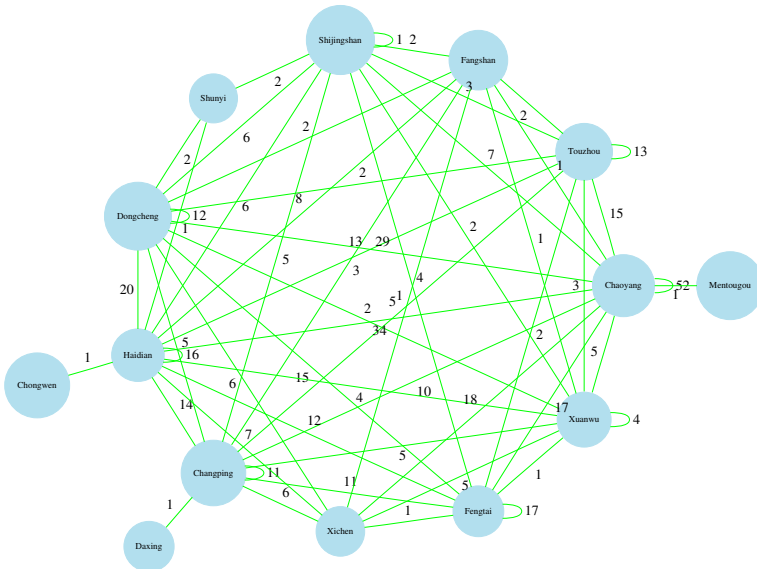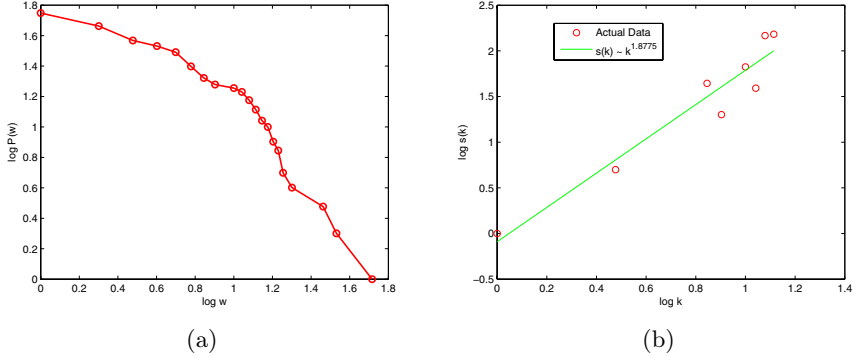| District	Chaoyang	Haidian	Dongcheng	Fengtai	Changping
Strength	241	152	113	97	87

Fig. 5. Weighted district network

(a)   (b)

Fig. 6. (a) WDN cumulative weight distribution. (b) Average node strength s_d as a function of node degree k_d.

One possible explanation is that both Chaoyang and Haidian District are major financial districts with (combined) more than 15 million permanent and temporary residents. Individuals in such densely-populated areas are more likely to be exposed to the epidemic and further spread the disease.

5   Weighted Occupation Network

Disease transmissions often happen in workplaces and in turn occupations can have an impact on the spreading patterns of infectious diseases [18]. In this section, we analyze the SARS transmission patterns based on the weighted occupation network (WON) as shown in Fig. 7. Our preliminary results show that the WON can reveal some additional insights.

The cumulative weight distribution of the WON, shown in Fig. 8 (a), follows a right-skewed distribution. Table 2 lists these five occupations with top strengths. The retiree category has the maximum strength 153, while the strengths of the other four occupations, military personnel, governmental employees, unemployed, and industry workers have relatively smaller occupation strengths.

This analysis shows that not all the occupations have equal probabilities to be infected with the SARS virus. For instance, the retiree population was more susceptible to be infected because of their lowered immune function. In the Chinese society, the retirees play an active role in family functions and child care and their working sons and daughters are in different occupations. Previous papers (e.g., [15]) have also reached similar conclusions. From a outbreak control perspective, those occupations with strong strengths need to be closely monitored.

Following an analysis procedure similar to that used for the WDN, we conclude that for the WON the average node strength s_d increases with the degree k_d and that $s_d(k_d) \sim k_d^{\beta}$, with the exponent $\beta = 1.6142$, which is larger than 1. This result is shown in Fig. 8(b), indicating that node strength is also strongly correlated to degree in the WON.
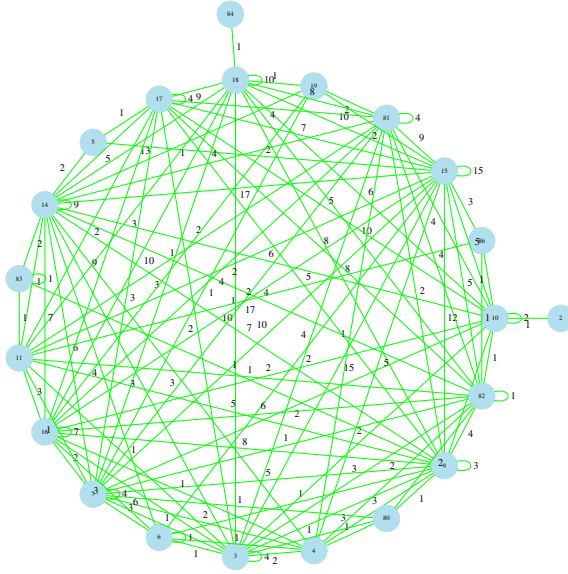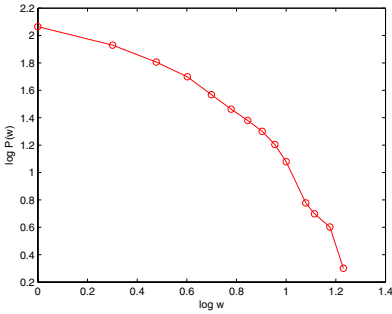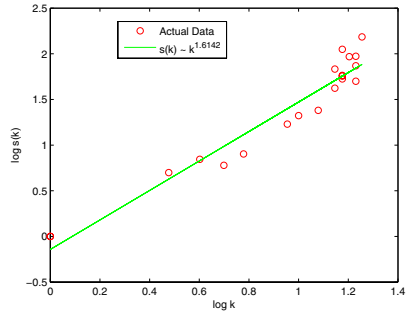
Fig. 7. Weighted occupation network

Table 2. Top five occupation (node) strengths

Occupation	Retiree	Military	Government Employee	Unemployed	Industry
Strength	153	112	94	93	74

(a)

(b)

Fig. 8. (a) WON cumulative weight distribution. (b) Average node strength s_d as a function of node degree k_d.

6 Concluding Remarks

In this paper, we analyze Beijing SARS data from a complex network analysis perspective. To better understand the SARS epidemic transmission patterns and evolution, we have studied three networks derived from the patient survey data, including a patient contact network, a weighted district network, and a weighted occupation network. The patient contact network possesses the scale-free degree distribution and its temporal evolution (as measured by average degree, maximum degree, and component ratio) exhibits some interesting patterns that can be explained by various control measures implemented during the SARS outbreak in Beijing. In both weighted district and occupation networks, the weights follow right-skewed distributions and the strengths of nodes are strongly correlated to their degrees. These observations and analysis results indicate that the traditional random isolation control method may not be effective. Instead, a more effective control program should target at nodes with high degree and strength.

Due to various difficulties in data collection, the Beijing SARS dataset used in our study may not be complete in that some infectious links may be missing. Our current work focuses on inferring some of these missing links for analysis purposes using methods similar to those reported in [19, 20, 21]. We are also working on analyzing various topological and distributional properties of weighted networks. The results are expected to benefit epidemiological data analysis in general.

Acknowledgments

References

[1] Wang, J.F., Christakos, G., Han, W., Meng, B.: Data-driven exploration of 'spatial pattern-time process-driving forces' associations of SARS epidemic in Beijing, China. Journal of Public Health, 1–11 (2008)

[2] Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., Brunham, R.C.: Network theory and SARS: predicting outbreak diversity. Journal of Theoretical Biology 232, 71–81 (2005)

[3] Riley, S., et al.: Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300, 1961–1966 (2003)

[4] Dye, C., Gay, N.: Modeling the SARS Epidemic. Science 300, 1884–1885 (2003)

[5] Becker, N.G., Glass, K., Li, Z., Aldis, G.K.: Controlling emerging infectious diseases like SARS. Mathematical Biosciences 193, 205–221 (2005)

[6] Small, M., Tse, C.K.: Clustering model for transmission of the SARS virus: application to epidemic control and risk assessment. Physica A: Statistical Mechanics and its Applications 351, 499–511 (2005)

[7] Barrat, A., Barthélemy, M., PastorSatorras, R., Vespignani, A.: The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America 101, 3747–3752 (2004)

[8] May, R.M., Lloyd, A.L.: Infection dynamics on scale-free networks. Physical Review E 64, 066112 (2001)

[9] Newman, M.E.J.: Spread of epidemic disease on networks. Physical Review E 66, 016128 (2002)

[10] Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86, 3200 (2001)

[11] Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47 (2002)

[12] Vazquez, A.: Causal tree of disease transmission and the spreading of infectious diseases. DIMACS Series in Discrete Mathematics and Theoretical Computer Science 70, 163–179 (2006)

[13] Haydon, D.T., Chase-Topping, M., Shaw, D.J., Matthews, L., Friar, J.K., Wilesmith, J., Woolhouse, M.E.J.: The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proceedings of Royal Society London B 270 (2003)

[14] Liljeros, F., Edling, C.R., Amaral, L.A.N.: Sexual networks: implications for the transmission of sexually transmitted infections. Microbes and Infection 5, 189–196 (2003)

[15] Shen, Z., Ning, F., Zhou, W., He, X., Liu, C., Chin, D.P., Zhu, Z., Schuchat, A.: Superspreading SARS Events, Beijing, 2003. Emerging Infectious Diseases 10, 256–260 (2004)

[16] Chen, Y., Tseng, C., King, C., Wu, T.J., Chen, H.: Incorporating Geographical Contacts into Social Network Analysis for Contact Tracing in Epidemiology: A Study on Taiwan SARS Data. In: Zeng, D., Gotham, I., Komatsu, K., Lynch, C., Thurmond, M., Madigan, D., Lober, B., Kvach, J., Chen, H. (eds.) BioSurveillance 2007. LNCS, vol. 4506, pp. 23–36. Springer, Heidelberg (2007)

[17] Bagler, G.: Analysis of the airport network of India as a complex weighted network. Physica A: Statistical Mechanics and its Applications 387, 2972–2980 (2008)

[18] Vazquez, A.: Epidemic outbreaks on structured populations. Journal of Theoretical Biology 245, 125–129 (2007)

[19] Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101 (2008)

[20] Hao, M., Irwin, K., Michael, R.L.: Effective missing data prediction for collaborative filtering. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, Amsterdam (2007)

[21] Ronald, K.P.: The problem of disguised missing data. SIGKDD Explorations Newsletters 8, 83–92 (2006)