Validation of Network-Dependent Epidemic Processes

A Study of Dr. Snow's Seminal Cholera Dataset[*]

Philip E. Paré
University of Illinois
Urbana, Illinois 61801
philpare@illinois.edu

Ji Liu
Stony Brook University
Stony Brook, New York 11794
ji.liu@stonybrook.edu

Carolyn L. Beck
University of Illinois
Urbana, Illinois 61801
beck3@illinois.edu

Tamer Başar
University of Illinois
Urbana, Illinois 61801
basar1@illinois.edu

Angelia Nedić
Arizona State University
Tempe, Arizona 85281
angelia.nedich@asu.edu

ABSTRACT

Models of spread processes over non-trivial networks are commonly motivated by modeling and analysis of biological networks, computer networks, and human contact networks. However, identification of such models has not yet been explored in detail, and the models have not been validated by real data. In this paper, we present a sufficient condition for asymptotic stability of the healthy equilibrium, show that the condition is necessary and sufficient for uniqueness of the healthy equilibrium, and present a result on learning the ratio of the spread parameters. Finally, we employ John Snow's seminal work on cholera epidemics in London in the 1850's to validate an approximation of a well-studied network-dependent susceptible-infected-susceptible (SIS) model.

KEYWORDS

SIS epidemic processes, model validation, data-driven analysis

ACM Reference format:
Philip E. Paré, Ji Liu, Carolyn L. Beck, Tamer Başar, and Angelia Nedić. 1997. Validation of Network-Dependent Epidemic Processes. In *Proceedings of ACM KDD epiDAMIK'18 - International Workshop on Epidemiology meets Data Mining and Knowledge Discovery, London, UK, August 2018 (epiDAMIK'18),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

Mathematical models of virus spread have been studied for centuries [2]. Recently these models have been extended to include network structure. In this work we focus on SIS models with infection parameters β_i and a healing rates δ_i. A virus model is called *homogeneous* if the infection and healing rates are the same for every agent, and *heterogeneous* if they are different for each agent. In this work, we focus on discrete-time SIS models, mainly for the more general, heterogeneous models. For reviews on epidemic processes see [8, 10].

While parameter estimation of epidemic spread with real data has been carried out for some models [6, 7, 15], the previous work has either not had network structure included or employed a large probabilistic model. Ignoring network structure is tantamount to making a strong simplifying assumption, and using a full probabilistic model can become very computationally expensive as the size of the network grows. For these reasons we focus on a nonlinear network-dependent ordinary differential equation model. To the best of our knowledge, no work has been done on the identification of spread parameters from data for these models. Many virus spread papers using these models have claimed to use real data to test their models, but no true validation of non-trivial network-dependent SIS spread models has been done. Previous work has used real data to identify underlying network structure, however there have been no prior efforts that have considered spread process data and identification over these networks. [4, 14].

We use the cholera dataset compiled by John Snow in [12] to validate the spread model analyzed in this work. Dr. Snow mapped the deaths caused by cholera in the Soho District of London in 1854 to illustrate that the infection was being spread by contaminated water via a specific pump, the Broad Street pump, and not via the air, as was the belief at the time. This seminal work by Snow has led to the modern day field of epidemiology [3]. While now, partially due to Snow, we understand cholera, how it spreads, and how to mitigate it, this illness is still a serious problem in poorer parts of the world today, highlighted by the current outbreak in Yemen where there have been over one million suspected cases of cholera and over 2,270 cholera-related deaths since the end of April 2017 [1].

John Snow's original spatial dataset of the cholera epidemic is static and does not contain time series data. Shiode *et al.* created spatial time series data, presented in [11] using additional sources and some statistical methods. However, Shiode *et al.* did not perform any dynamic analysis on their dataset, and have not made the dataset publicly available. We use a technique developed in the analysis section herein, combined with several strong but reasonable assumptions, to reproduce time series data, and in so doing, validate the model with the dataset. As far as we know, this is the first attempt to study Snow's cholera dataset from a dynamical systems' perspective to validate models of epidemic processes.

1 SIS MODEL

We focus on a discrete-time SIS model. The state x_i can correspond to the probability of infection of the ith agent [13] or the infected

[*] The full version of this work is available in [9].

epiDAMIK'18, London, UK
© 2016 Copyright held by the owner/author(s). 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

proportion of group i [5]. For the identification of the spread process parameters in Section 3 we employ the latter case. We model the system dynamics by

$$x_i^{k+1} = x_i^k + h\left((1-x_i^k)\beta_i \sum_{j=1}^{n} a_{ij}x_j^k - \delta_i x_i^k\right), \tag{1}$$

where k is the time index and $h > 0$ is the sampling parameter. We write (1) in matrix form as

$$x^{k+1} = x^k + h((I - X^k)BA - D)x^k, \tag{2}$$

where $X^k = diag(x^k)$, $B = diag(\beta_i)$, and $D = diag(\delta_i)$. Note that A is the matrix of a_{ij}'s and is not necessarily symmetric.

For the model to be well-defined we make several assumptions.

Assumption 1. *For all $i \in [n]$, we have $x_i^0 \in [0, 1]$.*

Assumption 2. *For all $i \in [n]$, we have $\beta_i \geq 0$, $\delta_i \geq 0$ and, for all $j \in [n]$, $a_{ij} \geq 0$.*

Assumption 3. *For all $i \in [n]$, $h\delta_i \leq 1$ and $h\beta_i \sum_{j \neq i} a_{ij} \leq 1$.*

Lemma 1.1. *For the system in* (2)*, under the conditions of Assumptions 1, 2, and 3, $x_i^k \in [0, 1]$ for all $i \in [n]$ and $k \geq 0$.*

Lemma 1.1 implies that the set $[0, 1]^n$ is positively invariant with respect to the system defined by (2). Since x_i denotes the fraction of group i infected, or is an approximation of the probability of infection of individual i and $1-x_i$ denotes the fraction of group i that is healthy, or is an approximation of the probability of individual i being healthy, it is natural to assume that their initial values are in the interval $[0, 1]$, since otherwise the values will lack any physical meaning for the epidemic model considered here. Therefore, we focus on the analysis of (2) only on the domain $[0, 1]^n$.

We also make the following assumption to ensure *non-trivial* virus spread.

Assumption 4. *We have $h \neq 0$ and $\exists i \neq j$ s.t. $\beta_{ij} > 0$.*

Note that we do not assume the healing rates to be nonzero. This allows for the possibility of SI (susceptible-infected) models.

2 ANALYSIS

For analysis purposes we need an assumption on the structure of the BA matrix. A square matrix is called *irreducible* if it cannot be permuted to a block upper triangular matrix.

Assumption 5. *The matrix BA is irreducible.*

Note that this assumption is equivalent to the underlying graph being strongly connected.

Theorem 2.1. *Suppose that Assumptions 1-5 hold for* (2)*. If $\rho(I - hD + hBA) \leq 1$, then the healthy state is asymptotically stable with domain of attraction $[0, 1]^n$.*

Proposition 1. *Suppose that Assumptions 1-5 hold. If $\rho(I - hD + hBA) > 1$, then* (2) *has two equilibria, $\mathbf{0}$ and x^*, where $x^* \gg \mathbf{0}$.*

Theorem 2.2. *Under Assumptions 1-5, the healthy state is the unique equilibrium of* (2) *if and only if $\rho(I - hD + hBA) \leq 1$.*

The following corollary shows that the ratio of the spread parameters can be recovered for the heterogeneous case with different δ_i's and β_i's for each agent (and includes the homogeneous case as a special case) if A and the endemic state are known.

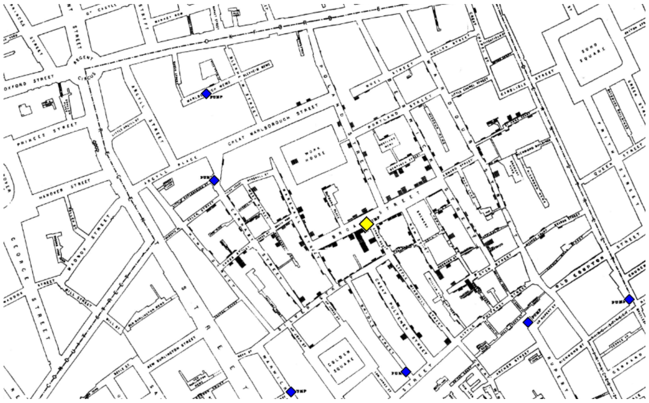


Figure 1: Map of cholera spread in London in 1854 compiled by John Snow [12]: healthy water pumps, the contaminated pump, and household deaths are depicted by blue diamonds, the yellow diamond, and black rectangles, respectively.

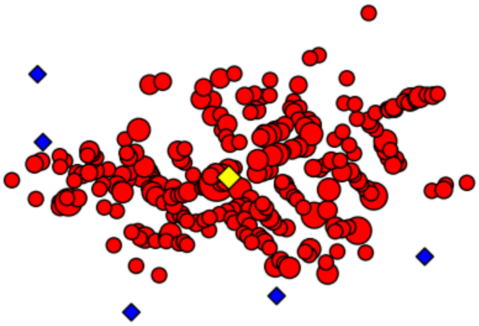


Figure 2: Digitization of Figure 1: The healthy water pumps, the contaminated pump, and the deaths are depicted by blue diamonds, the yellow diamond, and red dots with the diameters scaled by the number of deaths, respectively.

Corollary 2.3. *Considering the model in* (1) *under Assumptions 1-5, if A and the endemic state, $x^* \gg 0$, are known, then*

$$\frac{\delta_i}{\beta_i} = \frac{(1 - x_i^*)}{x_i^*} \sum_{j=1}^{n} a_{ij}x_j^*. \tag{3}$$

We will use the above corollary in the validation work that follows.

3 VALIDATION: SNOW DATASET

Now we employ the seminal cholera dataset collected by John Snow [12] for validation of the model in (1).

3.1 Snow Dataset

Snow depicted the number of deaths per household caused by cholera in the Soho District of London in 1854 on a map of the area. In Figure 1, the original map is shown, where each small rectangle corresponds to one death at that address. Snow created this map to illustrate to officials that the cholera epidemic was being spread by infected water from the Broad Street pump (the yellow diamond), and not through the air, the common belief at that time. We have plotted this data in Figure 2, with diamonds indicating the water

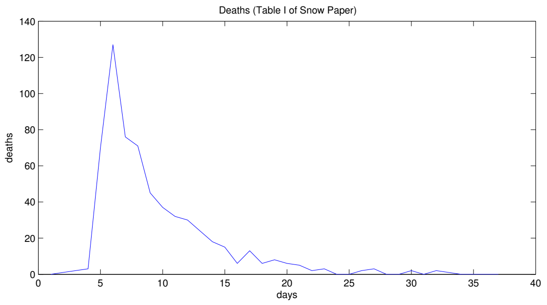


Figure 3: Total deaths per day in the Soho District of London in 1854, compiled by John Snow (from Table I in [12]).

pumps and red dots indicating deaths. The dataset is comprised of 250 households with at least one death. Snow also documented the cumulative deaths per day in Table I of [12], plotted in Figure 3. The time of deaths for each address is not recorded. The total cumulative deaths in the table is 616, but the total number of deaths on the map are 489. Therefore, there is a discrepancy of 127 deaths, whose household addresses are not included in the map.

3.2 Spread Validation

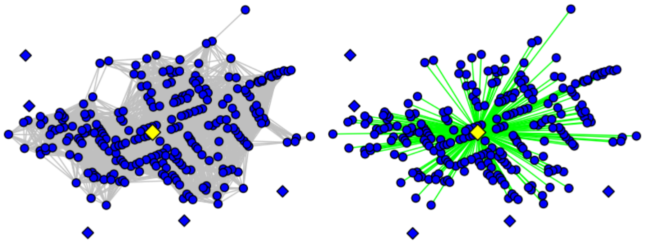
For the validation, each household with a death recorded by Snow in the map in Figure 1 corresponds to a node in the model. The last node in the model corresponds to the contaminated pump, the one on Broad Street, and we do not include the healthy water pumps in the model. We realize that ignoring the households with no recorded deaths and ignoring the healthy pumps are nontrivial assumptions. However, as was noted by Snow, many residents fled the city once they became aware of the outbreak [12]. For the households that did not flee, we assume they either had such a high healing rate that their inclusion would have been trivial and/or that these households exclusively drank from another pump and did not closely associate with neighbors who did drink from the Broad Street pump. Despite these (and subsequent) relatively strong assumptions, the validation results are quite promising.

The state of the system, x^k, is the percentage of total deaths in each household up to time k. The epidemic equilibrium of the system, which we call x^*, was calculated from the data in Figure 2, for the first attempt, by dividing the total number of deaths in each household by 20, and therefore assuming that each household has 20 members. This number was chosen because the maximum number of deaths was 15. For the last attempt we approximated the household sizes using Figure 1 in [11]; see Table 1. The last element of x^*, corresponding to the contaminated pump, was set to $\frac{19}{20}$.

We employed Corollary 2.3 to calculate the $\frac{\delta_i}{\beta_i}$ values. Then for simulation we set $\beta_i = 1$ for all i and chose h as large as possible while still meeting Assumption 3. For the initial condition in the simulations, we began with the Broad Street pump infected and all the households healthy:

$$x^0 = \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}^\top. \tag{4}$$

This initial condition is shown in Figure 4 (as well as the two considered graph structures), where the contaminated pump is depicted as a yellow diamond. As a consequence of these assumptions, our tuning parameter for adjusting the learned δ_i parameters, and consequently the spread behavior, was the connectivity matrix A.



(a) $A^{(1)}$ **from** (5)

(b) $A^{(2)}$ **from** (6)

Figure 4: Initial condition of simulations with graph structures: blue circles indicate healthy households and the yellow diamond indicates the infected pump.

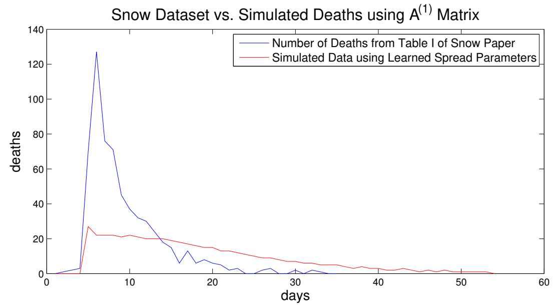


Figure 5: Comparison of Figures 3 and the simulated data using the learned parameters from the data in Figure 2, employing Corollary 2.3 and $A^{(1)}$ from (5): Note that the model does not capture the behavior of the system. The Euclidean distance between the two plots is 146.52, and the infinity norm is 105.

For the first attempt, we designed $A^{(1)}$ such that

$$a_{ij}^{(1)} = \begin{cases} 1, & \text{if } \|z_i - z_j\| < r, \\ 1, & \text{if } i = j, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where z_i is the location of household i and r was smallest number such that the graph was connected (shown in Figure 4a). Using the $\frac{\delta_i}{\beta_i}$ values derived using $A^{(1)}$, we simulated the system, using (1). To meet the constraints of Assumption 3, we had to set $h = \frac{1}{175}$. To create a plot of deaths per day, we multiplied the state of the system, i.e., the percentage of deaths in each household up to that point, by the household sizes (assumed to be 20), rounded to the nearest integer, took the difference between the states of each time step (since the state represents cumulative number of deaths up to that point), and then summed every three time series points (due to the small h value), therefore assuming that each time series point corresponds to a third of a day. Note that this approach does not capture the behavior of the system very well as it is very different than the dataset, as depicted in Figure 5. The Euclidean distance between the two plots is 146.52, and the infinity norm is 105.

Household Sizes	
Range in [11]	Estimate
0-4	4
5-9	7
10-14	12
15-24	20
24-403	25*

Table 1: Estimates for household sizes from Figure 1 in [11] used in the simulation with $A^{(2)}$: *The workhouse population was set to 403.

For the final attempt we changed to heterogeneous household sizes, using Figure 1 in [11] to approximate these values. We removed all edges except the self loops and the binary directed edges from the pump to every household with at least one death. The connection from the pump to the workhouse was set to $\frac{1}{10}$ because they had their own well and only a small fraction of the 403 residents drank from the Broad Street pump [12]. Therefore

$$A^{(2)} = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & \vdots \\ 0 & 0 & \ddots & 0 & \frac{1}{10} \\ 0 & 0 & \dots & 1 & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}. \qquad (6)$$

We found via simulation that as long as the edge weight corresponding to the workhouse was less than or equal to 0.45 then the results were very similar.

Plotting the data from Figure 3 and the simulated data using the learned parameters from the data in Figure 2, employing Corollary 2.3 and $A^{(2)}$ from (6) on the same plot for comparison in Figure 6 shows that we capture the behavior of the outbreak quite well. The Euclidean distance between the two plots is 75.16, and the infinity norm is 70. One of the reasons for this discrepancy is due to the fact that we used the spatial dataset in Figures 1-2, which had only 489 documented deaths, while the cumulative data from Table I in [12], shown in Figure 3 and the blue line in Figure 6, has a total of 616 deaths. The difference of 127 has caused the discrepancy. The lack of the address information for the additional 127 deaths is one of the reasons the plots are not identical. However, the discrepancy is distributed fairly evenly across the whole sample time. Consequently, we have shown that the model in (1) captures the behavior of the cholera epidemic from John Snow's 1854 dataset very well. Note that the fact that $A^{(2)}$ from (6) performs the best supports Snow's hypotheses that the Broad Street pump was the source of the cholera outbreak, and that cholera does not spread easily between people or the air, which is known to be true today.

4 CONCLUSION

We have provided necessary and sufficient conditions for uniqueness of the healthy equilibrium, conditions for the existence of an endemic state., and a necessary condition for asymptotic stability of the healthy state. We use a corollary of this analysis to recover the ratio of the virus spread parameters. Using this corollary we have validated a discrete-time, network-dependent SIS virus spread model using John Snow's seminal cholera dataset with very good results. In future work, would like to find other datasets to help further validate the SIS models. We would like to further study identification of the spread model accounting for noise in the data.

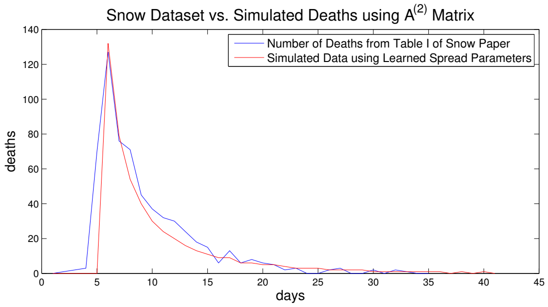


Figure 6: Comparison of Figure 3 and the simulated data using the learned parameters from the data derived using $A^{(2)}$ in (6): Note that there is a difference in the magnitude, but the general shapes are very similar.

REFERENCES

[1] Reema Al Yusfi, Malika Bouhenia, and Lauren O'Connor. 2018. Weekly Epidemiological Bulletin W14 2018. (2018). http://www.emro.who.int/images/stories/yemen/week_14.pdf?ua=1.

[2] Daniel Bernoulli. 1760. Essai dfiune nouvelle analyse de la mortalité causée par la petite vérole et des avantages de lfiinoculation pour la prévenir. *Histoire de lfiAcad. Roy. Sci.(Paris) avec Mém. des Math. et Phys. and Mém* (1760), 1–45.

[3] Ruth Bonita, Robert Beaglehole, and Tord Kjellström. 2006. *Basic epidemiology.* World Health Organization.

[4] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. 2008. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)* 10, 4 (2008), 1.

[5] A Fall, Abderrahman Iggidr, Gauthier Sallet, Jean-Jules Tewa, and others. 2007. Epidemiological models and Lyapunov functions. *Math. Model. Nat. Phenom* 2, 1 (2007), 62–68.

[6] Matt J Keeling, , and *et al.* 2001. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 294, 5543 (2001), 813–817.

[7] Hongyu Miao, Xiaohua Xia, Alan S Perelson, and Hulin Wu. 2011. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM review* 53, 1 (2011), 3–39.

[8] Cameron Nowzari, Victor M Preciado, and George J Pappas. 2016. Analysis and Control of Epidemics: A Survey of Spreading Processes on Complex Networks. *IEEE Control Systems Magazine* 36, 1 (2016), 26–46.

[9] P. E. Paré, J. Liu, C. L. Beck, B. E. Kirwan, , and T. Başar. 2018. Discrete Time Virus Spread Processes: Analysis, Identification, and Validation. conditionally accepted to *IEEE Transactions on Control Systems Technology: System Identification and Control in Biomedical Applications* (2018). arXiv:1710.11149 [math.OC].

[10] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. 2015. Epidemic processes in complex networks. *Reviews of modern physics* 87, 3 (2015), 925.

[11] Narushige Shiode, Shino Shiode, Elodie Rod-Thatcher, Sanjay Rana, and Peter Vinten-Johansen. 2015. The mortality rates and the space-time patterns of John Snow's cholera epidemic map. *International journal of health geographics* 14, 1 (2015), 21.

[12] John Snow. 1855. *On the mode of communication of cholera.* John Churchill.

[13] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. 2009. Virus spread in networks. *IEEE/ACM Transactions on Networking* 17, 1 (2009), 1–14.

[14] Yan Wan, Sandip Roy, and Ali Saberi. 2008. Designing spatially heterogeneous strategies for control of virus spread. *IET Systems Biology* 2, 4 (2008), 184–201.

[15] Ling Xue, H Morgan Scott, Lee W Cohnstaedt, and Caterina Scoglio. 2012. A network-based meta-population approach to model Rift Valley fever epidemics. *Journal of Theoretical Biology* 306 (2012), 129–144.