Analysis, Estimation, and Validation of Discrete-Time Epidemic Processes

Philip E. Paré, Ji Liu, Carolyn L. Beck, Barrett E. Kirwan, and Tamer Başar, *Life Fellow, IEEE*

***Abstract*— Models of spreading processes over nontrivial networks are commonly motivated by modeling and analysis of biological networks, computer networks, and human contact networks. However, learning the spread parameters 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 several different spread models from the literature and explore their relationships to each other; for one of these processes, 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 necessary and sufficient conditions for estimating the spread parameters. Finally, we employ two real data sets, one from John Snow's seminal work on cholera epidemics in London in the 1850s and the other one from the United States Department of Agriculture, to validate an approximation of a well-studied network-dependent susceptible-infected-susceptible model.**

***Index Terms*— Epidemic processes, John Snow's cholera data set, networked control systems, nonlinear systems, system identification in biomedical applications, validation of networked systems.**

I. Introduction

SPREADING processes have been studied in many fields. In the systems and control community, the main interest has been on susceptible-infected-susceptible (SIS) spread models over nontrivial networks. These models have been proposed for discrete time [1]–[8] and continuous time [4], [9]–[13] and are based on an infection parameter β and a healing rate δ. 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 paper, we will focus on discrete-time SIS models, mainly for the more general, heterogeneous models, but exploring estimating the spread parameters of homogeneous models as well.

Manuscript received November 1, 2017; revised May 17, 2018 and August 5, 2018; accepted August 12, 2018. Date of publication October 10, 2018; date of current version December 27, 2019. Manuscript received in final form September 4, 2018. This work was supported in part by USDA, CA, under Grant 58-6000-4-0028 and in part by the National Science Foundation under Grant CPS 1544953 and Grant ECCS 1509302. Recommended by Associate Editor G. Mercere. *(Corresponding author: Philip E. Paré.)*

P. E. Paré, C. L. Beck, and T. Başar are with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: philip.e.pare@gmail.com; beck3@illinois.edu; basar1@illinois.edu).

J. Liu is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: ji.liu@stonybrook.edu).

B. E. Kirwan is with the Agricultural and Consumer Economics Department, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2018.2869369

Wang *et al.* [1] introduced a discrete-time homogeneous virus spread model that is dependent on a nontrivial undirected graph structure. The authors give an epidemic threshold for the model in terms of the maximum eigenvalue of the matrix depicting the graph structure in relation to the ratio of β and δ that ensures convergence to the healthy state, that is, where the virus is eradicated. Van Mieghem *et al.* [9] point out that the model in [1] is only accurate for spreading processes when it is known that the virus meets the condition to die out in exponential time. Chakrabarti *et al.* [2] explored the same model as [1] but in more detail. Ahn and Hassibi [4] studied both discrete- and continuous-time homogeneous SIS models. Both the healthy and the endemic[1] states of several models are considered, and existence, uniqueness, and stability conditions for special cases of the endemic state are established. They also provide a sufficient condition for the global stability of the endemic state for the model in [1] and [2].

In [4], they only perform a local stability analysis of the endemic state of the model of interest; the analysis here is global and for a more general model. Ahn and Hassibi [5] and Paarporn *et al.* [6] explored how an n-state discrete-time model approximates a full 2^n-state probabilistic model, and Paarporn *et al.* [6] extended the model to include human awareness. Han *et al.* [7] applied the geometric programming ideas used in [10] to the discrete-time model. Watkins *et al.* [8] studied a full probabilistic model that uses partial measurements and performed inference on the states that are not measured using a Bayesian approach and propose a feedback control technique.

These references include the basic framework and recent history of the models we consider, which we validate using real spread data. The model we focus on in this paper is similar to a special case in [4]. However, the model in [4] assumes homogeneous virus spread and an unweighted adjacency matrix. The models in Sections II and III are not limited by these assumptions.

A. Related Work

While parameter estimation of epidemic spread with real data has been carried out for some models [14]–[17], the previous work has either not included network structure or has employed a large probabilistic model. Ignoring network structure is tantamount to making a strong simplifying assumption,

[1]By "endemic" we mean relating to a virus constantly present to greater or lesser extent in a particular area.

and using a full probabilistic model can become very computationally expensive as the size of the network grows [9], [13]. 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 estimation 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 nontrivial network-dependent SIS spread models has been done. Those papers that use real data only build the network structure using real data, but do not have real spreading process data over that network. Wan *et al.* [18], [19] compared their model to a simulator of severe acute respiratory syndrome (SARS), not real data. Chakrabarti *et al.* [2] used a router network from the state of Oregon and simulated an artificial spreading process over that network. To the best of our knowledge, no work has been done that validates network-dependent SIS models using a set of real spread data. Similarly, Preciado *et al.* [10] used data from an air transportation network but simulated using arbitrarily chosen healing and infection rates. For reviews on epidemic processes, see [20] and [21].

We use two data sets to validate the spread model analyzed in this paper. The first data set is the cholera data set compiled by Snow [22]. Snow [22] 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 [23]. While now, partially due to Snow, we understand cholera, how it spreads, and how to mitigate it, and this illness is still a serious problem in poorer parts of the world today. This is highlighted by the recent outbreak in Yemen, where there have been over one million suspected cases of cholera and over 2270 cholera-related deaths since the end of April 2017 [24], [25].

John Snow's Original Spatial data set of the cholera epidemic is static and does not contain time series data. Shiode *et al.* [26] created spatial time series data, using additional sources and some statistical methods. However, Shiode *et al.* [26] did not perform any dynamic analysis on their data set and have not made the data set 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 data set. As far as we know, this is the first attempt to study Snow's cholera data set from a dynamical systems' perspective to validate the models of epidemic processes.

The second data set used herein is a record of all the payouts from the United States Department of Agriculture (USDA) to farms/farmers for all USDA-sponsored subsidy programs from 2008 to 2013. For this paper, we focus on the 2009 Average Crop Revenue Election (ACRE) Program, which was introduced in that year as an alternative to an existing program, the Direct and Counter-Cyclical Payment (DCP) Program [27], [28]. These programs are in place to reduce the risk in the U.S. farming industry, enabling the adoption of new technologies. One of the goals of this paper is to determine whether the adoption of the ACRE program followed a network-dependent discrete-time spreading consistent with the model studied herein.

A large body of literature in agricultural economics has modeled the adoption and diffusion of agricultural technology, e.g., fertilizer and new seed varieties (see [29] for a review of this literature). This literature generally models individuals' decisions to adopt new technologies or the extent of overall adoption, but the spread of information and technology is treated as a "black box." Recent work in developing countries has examined whether farmers learn about new technologies from "information neighbors." Foster and Rosenzweig examined survey data and found that farmers' adoption of high-yielding varieties during the Green Revolution depended on neighbors' experiences [30]. Recent evidence from randomized controlled trials shows that farmers learn from their neighbors' experience when the technology is novel or complex [31], but not when adjustments to current practices are minor [32]. Ghanaian farmers learned from neighbors' experience when switching from traditional crops to pineapple [31], whereas information about optimal fertilizer used for traditional crops in Kenya did not spread among neighbors [32]. We take a new approach by using virus spread models to characterize the spread of complex information among U.S. farmers.

A preliminary version of this paper is in the Proceedings of the American Control Conference [33]. However, this paper provides: 1) the complete proofs of all the results; 2) additional illustrative simulations; and 3) the validation of the model using the Snow cholera data set, which were not included in [33].

This paper is organized as follows. In Section II, the virus spread models are introduced with several remarks that provide insight into how the models are related to each other. In Section III, we analyze one of the discrete-time spreading processes from Section II that has not been explored in detail. In Section IV, we present necessary and sufficient conditions for estimating the spreading process parameters of the same model, from data produced by the models. In doing so, we establish several assumptions that need to be met by the data sets. In Section V, we validate the results from Section IV via simulation. In Section VI, we introduce Snow's foundational cholera data set from 1854 and use it to validate the spread model. In Section VII, we introduce the USDA data set and the associated subsidy programs, and we estimate the homogeneous spread parameters of the ACRE program using data from one part of the country and verify the estimated parameters by simulating the spread model over the complete contiguous United States and comparing the simulated data with the actual data. We conclude with some discussion of the results and future work in Section VIII.

B. Notation

Given a vector function of continuous time $x(t)$, we use $\dot{x}(t)$ to indicate the time-derivative. Given a vector function of discrete time x^k, the superscript indicates the time step of x. Given a vector $x \in \mathbb{R}^n$, the 2-norm is denoted by $\|x\|$

and the transpose by $x^\top$. The notation $\mathbf{0}$ denotes the vector, whose entries all equal 0. Given two vectors $x_1, x_2 \in \mathbb{R}^n$, $x_1 > x_2$ indicates each element of x_1 is greater than or equal to the corresponding element of x_2 and $x_1 \neq x_2$, and $x_1 \gg x_2$ indicates each element of x_1 is strictly greater than the corresponding element of x_2. Given a matrix $A \in \mathbb{R}^{n\times n}$, the spectral radius is $\rho(A)$ and the largest real-valued part of the eigenvalues of A is denoted by $s_1(A)$ (if the spectrum is possibly complex). In addition, a_{ij} indicates the i, jth entry of the matrix A, and $\|A\|_F$ indicates the Frobenius norm of A. The notation diag$(\cdot)$ refers to a diagonal matrix with the argument(s) on the diagonal. For any positive integer n, we define $[n] := \{1, \ldots, n\}$.

II. SIS Models

We introduce two discrete-time SIS models and discuss their relationship. For these SIS models, there are two levels of granularity for modeling the system. The state x_i can correspond to a binary classification of whether or not the ith agent is sick or healthy [4], [9], the probability of infection of the ith agent [9], or the proportion of infection of group i [34]. To estimate the spreading process parameters in Sections VI and VII, we employ the latter case.

The first discrete-time model we consider is derived from the continuous-time model

$$\dot{x}_i = (1 - x_i)\beta_i \sum_{j=1}^{n} a_{ij} x_j - \delta_i x_i \tag{1}$$

where i indicates the ith agent or group i, x_i is the infection level, $\beta_i > 0$ is the infection rate, $\delta_i > 0$ is the healing rate, and $a_{ij} \geq 0$ are edge weights between the agents/groups. Applying Euler's method [35] to (1) gives

$$x_i^{k+1} = x_i^k + h\left(\left(1 - x_i^k\right)\beta_i \sum_{j=1}^{n} a_{ij} x_j^k - \delta_i x_i^k\right) \tag{2}$$

where k is the time index and $h > 0$ is the sampling parameter. We write (2) in the matrix form as

$$x^{k+1} = x^k + h((I - X^k)BA - D)x^k \tag{3}$$

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

Remark 1: The model in (1) was derived from a mean field approximation of a 2^n state Markov chain model [9]

$$\dot{y} = Qy \tag{4}$$

where Q is the transition matrix of the Markov chain (the details of the 2^n state model are not needed for the discussion here, and hence are not included; for a more detailed discussion, see [13]). Therefore, (2) is an approximation of an approximation.

Discrepancies between (1), (2), and (4) are explored via simulation and discussed in Section V.

An alternative discrete-time model, studied in [4], is

$$x_i^{k+1} = x_i^k(1 - \delta_i) + \left(1 - x_i^k\right)\left(1 - \prod_{j=1}^{n}\left(1 - \beta_i a_{ij} x_j^k\right)\right). \tag{5}$$

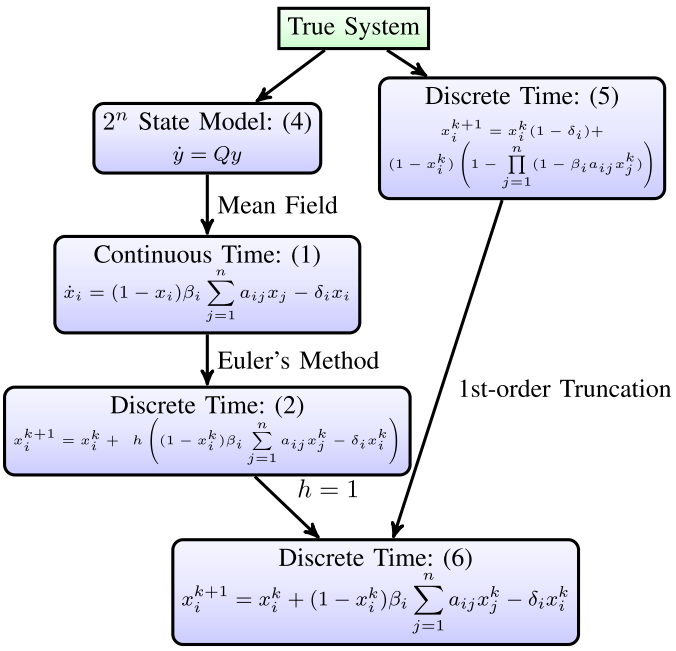


Fig. 1. Graphical illustration of the discussion in Section II and the point in Observation 1, showing how the two discrete-time spread models are related. The first modeling layer shows the 2^n state models. Arrows: different approximations taken. By "True System," we are referring to actual spreading in a real process. To derive a mathematical model of the true spreading process, some assumptions must be made; we do not assert that any of the mathematical models represent ground truth, that is, there is no completely accurate mathematical model.

By expanding the model given in (5), we obtain

$$x_i^{k+1} = x_i^k - \left(1 - x_i^k\right)\left[-\beta_i \sum_{j=1}^{n} a_{ij} x_j^k + \cdots + \beta_i^n \prod_{j=1}^{n}\left(-a_{ij} x_j^k\right)\right] - \delta_i x_i^k.$$

Remark 2: If we assume $\beta_i < 1\ \forall i$, the model in (5) can be approximated by truncating the terms with the powers of β_i greater than 1, giving

$$x_i^{k+1} = x_i^k + \left(1 - x_i^k\right)\beta_i \sum_{j=1}^{n} a_{ij} x_j^k - \delta_i x_i^k. \tag{6}$$

The preceding discussion leads us to the following observation.

Observation 1: The approximation given by (6) and the discrete approximation of the mean field approximation of the continuous 2^n state Markov model in (2) are equivalent, given $h = 1$.

The relationships between the models introduced in this section are shown in Fig. 1.

III. Analysis

In this section, a different version of the model in (2) will be analyzed as follows:

$$x^{k+1} = x^k + h((I - X^k)B - D)x^k \tag{7}$$

where $[B]_{ij} = \beta_{ij}$, capturing the infection rate and nearest-neighbor graph structure in one. Note β_{ij} could be factored into $\beta_i a_{ij}$ as in (2).

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

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

Assumption 3: For all $i \in [n]$, we have $h\delta_i \leq 1$ and $h \sum_{j \neq i} \beta_{ij} \leq 1$.

Lemma 1: For the system in (7), under the conditions of Assumptions 1–3, $x_i^k \in [0, 1]$ for all $i \in [n]$ and $k \geq 0$.

Proof: Suppose that at some time k, $x_i^k \in [0, 1]$ for all $i \in [n]$. Consider an index $i \in [n]$. Rearranging (2)

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

we see that x_i^{k+1} is a convex combination of $(1 - h\delta_i)$ and $h \sum_{j=1}^n \beta_{ij} x_j^k$. Since, by Assumptions 2 and 3, $h\delta_i, h \sum_{j=1}^n \beta_{ij} x_j^k \in [0, 1]$, we have $x_i^{k+1} \in [0, 1]$.

Furthermore, by Assumption 1, $x_i^0 \in [0, 1]$ for all $i \in [n]$, thus it follows that $x_i^k \in [0, 1]$ for all $i \in [n]$ and $k \geq 0$. □

Lemma 1 implies that the set $[0, 1]^n$ is positively invariant with respect to the system defined by (7). Since x_i denotes the fraction of group i infected or is an approximation of the probability of infection of individual i (see Fig. 1), 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 physical meaning for the epidemic model considered here. Therefore, we focus on the analysis of (7) only on the domain $[0, 1]^n$.

We also make the following assumption to ensure *nontrivial* 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 susceptible-infected (SI) models [36].

Definition 1: Consider an autonomous system

$$x^{k+1} = f(x^k) \tag{8}$$

where $f : \mathcal{X} \to \mathbb{R}^n$ is a locally Lipschitz map from a domain $\mathcal{X} \subset \mathbb{R}^n$ into $\mathbb{R}^n$. Let z be an equilibrium of (8) and $\mathcal{E} \subset \mathcal{X}$ be a domain containing z. If the equilibrium z is asymptotically stable such that for any $x^0 \in \mathcal{E}$ we have $\lim_{k\to\infty} x^k = z$, then $\mathcal{E}$ is said to be a domain of attraction for z.

Proposition 1: Let z be an equilibrium of (8) and $\mathcal{E} \subset \mathcal{X}$ be a domain containing z. Let $V : \mathcal{E} \to \mathbb{R}$ be a continuously differentiable function such that $V(z) = 0$, $V(x) > 0$ for all x in $\mathcal{E} \setminus \{z\}$, and $\Delta V^k := V(x^{k+1}) - V(x^k) < 0$ for all x^k in $\mathcal{E} \setminus \{z\}$. If $\mathcal{E}$ is a positively invariant set, then the equilibrium z is asymptotically stable with a domain of attraction $\mathcal{E}$.

This proposition is a direct consequence of Lyapunov's stability theorem for discrete-time systems, which can be found in [37], and the definition of domain of attraction.

Finally, we need an assumption on the structure of the B matrix. A square matrix is called *irreducible*, if it cannot be permuted to a block upper triangular matrix.

Assumption 5: The matrix B is irreducible.

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

Theorem 1: Suppose that Assumptions 1–5 hold for (7). If $\rho(I - hD + hB) \leq 1$, then the healthy state is asymptotically stable with the domain of attraction $[0, 1]^n$.

To prove the theorem, we need the following lemmas.

Lemma 2 [38]: Suppose that M is an irreducible nonnegative matrix such that $\rho(M) < 1$. Then, there exists a positive diagonal matrix P such that $M^\top PM - P$ is negative definite.

Lemma 3: Suppose that M is an irreducible nonnegative matrix such that $\rho(M) = 1$. Then, there exists a positive diagonal matrix P such that $M^\top PM - P$ is negative semidefinite.

Proof: From the Perron Frobenius Theorem for irreducible nonnegative matrices [39, Th. 8.4.4], there exists $v \gg 0$ such that $Mv = v$. Since $M^\top$ is also irreducible and nonnegative, there exists $u \gg 0$ such that $M^\top u = u$. Let P be a diagonal matrix, whose ith diagonal entry is equal to u_i/v_i, which gives $Pv = u$. Therefore,

$$(M^\top PM - P)v = M^\top Pv - Pv = M^\top u - u = 0.$$

Then by [40, Lemma 2.3], $\rho(M^\top PM - P) = 0$. □

Proof of Theorem 1: To simplify notation, let $M = I + hB - hD$ and $\hat{M} = I + h((I - X^k)B - D)$. By Assumptions 2–5, M is an irreducible nonnegative matrix. First, we evaluate the case where $\rho(I - hD + hB) < 1$. Therefore, by Lemma 2, there exists a positive diagonal matrix P_1 such that $M^\top P_1\ M - P_1$ is negative definite. Consider the Lyapunov function $V_1(x^k) = (x^k)^\top P_1\ x^k$. Using (7) with $x^k \neq 0$ gives

$$\begin{aligned}
\Delta V_1^k &= (x^k)^\top \hat{M}^\top P_1 \hat{M} x^k - (x^k)^\top P_1 x^k \\
&= (x^k)^\top (M^\top P_1 M - P_1)x^k - 2h(x^k)^\top B^\top X^k P_1 M x^k \\
&\quad + h^2 (x^k)^\top B^\top X^k P_1 X^k B x^k \\
&< h^2 (x^k)^\top B^\top X^k P_1 X^k B x^k \\
&\quad - 2h(x^k)^\top B^\top X^k P_1 M x^k \qquad (9) \\
&= h^2 (x^k)^\top B^\top X^k P_1 X^k B x^k \\
&\quad - 2h^2 (x^k)^\top B^\top X^k P_1 B x^k \\
&\quad - 2h(x^k)^\top B^\top X^k P_1 (I - hD) x^k \\
&\leq h^2 ((x^k)^\top B^\top X^k P_1 X^k B x^k \\
&\quad - 2(x^k)^\top B^\top X^k P_1 B x^k) \qquad (10) \\
&\leq -h^2 (x^k)^\top B^\top X^k P_1 (I - X^k) B x^k \\
&\leq 0 \qquad (11)
\end{aligned}$$

where (9) holds by Lemma 2, (10) holds by Assumptions 2 and 3, and (11) holds by Lemma 1. Therefore, by Proposition 1, the system converges asymptotically to the healthy state for this case.

For the case where $\rho(I - hD + hB) = 1$, we have, by Lemma 3, that there exists a positive diagonal matrix P_2 such that $M^\top P_2\ M - P_2$ is negative semidefinite. Consider the Lyapunov function $V_2(x^k) = (x^k)^\top P_2\ x^k$. Using (7) with

$x^k \neq 0$ gives

$$\begin{aligned}\Delta V_2^k &= (x^k)^\top \hat{M}^\top P_2 \hat{M} x^k - (x^k)^\top P_2 x^k \\ &= (x^k)^\top (M^\top P_2 M - P_2) x^k - 2h(x^k)^\top B^\top X^k P_2 M x^k \\ &\quad + h^2 (x^k)^\top B^\top X^k P_2 X^k B x^k \\ &< h^2 (x^k)^\top B^\top X^k P_2 X^k B x^k - 2h(x^k)^\top B^\top X^k P_2 M x^k \\ &= h^2 (x^k)^\top B^\top X^k P_2 X^k B x^k - h(x^k)^\top B^\top X^k P_2 M x^k \\ &\quad - h^2 (x^k)^\top B^\top X^k P_2 B x^k \\ &\quad - h(x^k)^\top B^\top X^k P_2 (I - hD) x^k \\ &\leq h^2 (x^k)^\top B^\top X^k P_2 X^k B x^k - h(x^k)^\top B^\top X^k P_2 M x^k \\ &\quad - h^2 (x^k)^\top B^\top X^k P_2 B x^k \\ &\leq -h^2 (x^k)^\top B^\top X^k P_2 (I - X^k) B x^k \\ &\quad - h(x^k)^\top B^\top X^k P_2 M x^k \\ &\leq -h(x^k)^\top B^\top X^k P_2 M x^k \\ &\leq 0.\end{aligned}$$

Clearly, if $x^k = \mathbf{0}$, then $-h(x^k)^\top B^\top X^k P_2\ Mx^k = 0$. Since, by Assumptions 2 and 4, B, P_2, M are nonzero, nonnegative matrices, and if $-h(x^k)^\top B^\top X^k P_2\ Mx^k = 0$, then $x^k = \mathbf{0}$. Therefore, by Proposition 1, the healthy state is asymptotically stable with the domain of attraction $[0, 1]^n$. □

Proposition 2: Let Assumptions 1–5 hold. If $\rho(I - hD + hB) > 1$, then (7) has two equilibria, $\mathbf{0}$ and x^*, where $x^* \gg \mathbf{0}$.

Proof: Clearly, $\mathbf{0}$ is always an equilibrium of (7).

By the Perron Frobenius Theorem for irreducible nonnegative matrices [39, Th. 8.4.4], $\rho(I - hD + hB) = s_1(I - hD + hB)$ and there exists $v \gg 0$ such that

$$(I - hD + hB)v = \rho(I - hD + hB)v > v$$

since $\rho(I - hD + hB) > 1$. Therefore,

$$(-hD + hB)v = \rho(-hD + hB)v = s_1(-hD + hB)v > 0v$$

which implies

$$\rho(I - hD + hB) > 1 \Longleftrightarrow h(s_1(-D + B)) > 0.$$

This condition is the same as the condition of [41, Proposition 3] and [42], and the proof follows similarly, showing that there exists $x^* \gg \mathbf{0}$ such that

$$h((-D + B) - X^* B)x^* = \mathbf{0}.$$

Therefore, $\mathbf{0}$ and x^* are equilibria of (7). □

Theorem 1 and Proposition 2 give the following result.

Theorem 2: Under Assumptions 1–5, the healthy state is the unique equilibrium of (7) if and only if $\rho(I - hD + hB) \leq 1$.

In [4], a counterexample is provided to show that the nontrivial equilibrium of (5) is unstable. However, this example does not hold for the models in (2) and (7), because it does not meet Assumption 3. Consequently, the state of the system does not stay in the domain of interest, $[0, 1]^n$.

Remark 3: If the system has homogeneous spread parameters, the condition in Theorems 1 and 2 reduces to $\rho(A) \leq (\delta/\beta)$.

IV. Estimating Spread Parameters

In this section, we provide the assumptions and the learning techniques for several versions of the model in (2), introduced in Section II. We assume that the underlying graph structure A is known and that we have full-state measurement with no noise on the measurements, which we admit are strong assumptions. However, for the second application considered here, these assumptions are well-founded because we aggregate the data by county and the adjacency of counties is known, i.e., the graph structure is known, and any farmer that received a subsidy payout is in the data set, i.e., there are no hidden, unmeasured states.

We present several results on estimating the spread parameters of the model in (2) from data.

Theorem 3: Consider the model in (2) under Assumptions 1–5 with homogeneous virus spread, that is, β and δ are the same for all n agents, with $n > 1$. Assume that A, $x^0, \ldots, x^T$, and h are known. Then, the spread parameters can be learned uniquely if and only if $T > 0$, and there exists $l \in [T]$ such that $x^l \neq x^0$.

Proof: Since A, $x^0, \ldots, x^T$, and h are known, using the notation in (3), we can construct the matrix

$$\Phi = \begin{bmatrix} h(I - X^0)Ax^0 & -hx^0 \\ \vdots & \vdots \\ h(I - X^{T-1})Ax^{T-1} & -hx^{T-1} \end{bmatrix}. \tag{12}$$

Therefore, we can rewrite (2) as

$$\begin{bmatrix} x^1 - x^0 \\ \vdots \\ x^T - x^{T-1} \end{bmatrix} = \Phi \begin{bmatrix} \beta \\ \delta \end{bmatrix}. \tag{13}$$

By the assumption that there exists $l \in [T]$ such that $x^l \neq x^0$, the left-hand side of the equation is nonzero, and by construction, the left-hand side is in the range of Φ. This is clearly overdetermined if $T \geq 1$ and $n > 1$; therefore, it will have a unique solution using the pseudoinverse.

If $T = 0$, then there is only one data point, and learning the dynamic spread parameters is not possible. Similarly, if there does not exist $l \in [T]$ such that $x^l \neq x^0$, then

$$x^0 = \cdots = x^T. \tag{14}$$

This would only occur if x^0 were an equilibrium point of (2). So by (14), we have that the left-hand side of (13) is

$$\begin{bmatrix} x^1 - x^0 \\ \vdots \\ x^T - x^{T-1} \end{bmatrix} = \mathbf{0}. \tag{15}$$

By Proposition 2, there are two cases where (14) can occur: 1) the healthy state ($x^0 = x^* = \mathbf{0}$) or 2) the endemic state ($x^0 = x^* \gg \mathbf{0}$).

1) If $x^0 = x^* = \mathbf{0}$, then, by (12) and (14), $\Phi = 0$. Therefore, by (13) and (15), β and δ can take any values, that is, they are not unique.
2) If $x^0 = x^* \gg \mathbf{0}$, then $\Phi \neq 0$. Therefore, by (13) and (15), $[\beta\ \delta]^\top$ is in the null space of Φ. This implies that $[\beta\ \delta]^\top$ is not unique, unless the null space equals $\{\mathbf{0}\}$. If the null space equals $\{\mathbf{0}\}$, then $[\beta\ \delta]^\top = \mathbf{0}$,

which is a contradiction because if $[\beta\ \delta]^\top = \mathbf{0}$, then there is no spreading process; therefore, there is no endemic state. □

Now, we present two corollaries regarding how the ratio of the spread parameters, δ/β, can be recovered. The first covers the case for when h is unknown.

Corollary 1: Consider the model in (2) under Assumptions 1–5 with homogeneous virus spread with $n > 1$. Assume that A and $x^0, \ldots, x^T$ are known. Then, the ratio of the spread parameters can be learned uniquely if and only if $T > 0$ and there exists $l \in [T]$ such that $x^l \neq x^0$.

Proof: Since h factors out of the right-hand side of (13) and is nonzero by Assumption 4, even if h is not known, a scaled version of the pair β and δ, that is, $h\beta$ and $h\delta$, can be recovered exactly. Therefore, the proportion of the two parameters can be found. □

Corollary 2 shows that the ratio of the spread parameters can be recovered for the heterogeneous case with different δ_i and β_i values for each agent (and includes the homogeneous case as a special case) if A and the endemic state are known.

Corollary 2: Considering the model in (2) under Assumptions 1–5, if A and the endemic state, $x^* \gg 0$, are known, then

$$\frac{\delta_i}{\beta_i} = \frac{\left(1 - x_i^*\right)}{x_i^*} \sum_{j=1}^{n} a_{ij} x_j^*. \tag{16}$$

Proof: By replacing x_i^{k+1} and x_i^k in (2) with x_i^*, we have

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

Since $x_i^* > 0$, we can divide by x_i^* giving the result. □

These corollaries illustrate that under certain conditions, while the exact behavior of the system may not be recoverable, the limiting behavior of the system may be determined, by employing Theorems 1 and 2 with Remark 3.

If the assumption is made that the underlying spreading process is heterogeneous, that is, different δ_i and β_i values for each agent, a similar result to Theorem 3 can be concluded.

Theorem 4: Consider the model in (2) under Assumptions 1–5 with $n > 1$. Assume that A, $x^0, \ldots, x^T$, and h are known. Then, the spread parameters of node i can be learned uniquely if and only if $T > 0$, and there exists $l \in [T]$ such that $x_i^l \neq x_i^0$.

Proof: Since A, $x^0, \ldots, x^T$, and h are known, for each i, we can construct the matrix

$$\Phi_i = \begin{bmatrix} h\left(1 - x_i^0\right) \sum_{j=1}^{n} a_{ij} x_j^0 & -h x_i^0 \\ \vdots & \vdots \\ h\left(1 - x_i^{T-1}\right) \sum_{j=1}^{n} a_{ij} x_j^{T-1} & -h x_i^{T-1} \end{bmatrix}. \tag{17}$$

Then, we have

$$\begin{bmatrix} x_i^1 - x_i^0 \\ \vdots \\ x_i^T - x_i^{T-1} \end{bmatrix} = \Phi_i \begin{bmatrix} \beta_i \\ \delta_i \end{bmatrix}. \tag{18}$$

The remainder of the proof follows that of Theorem 3. □

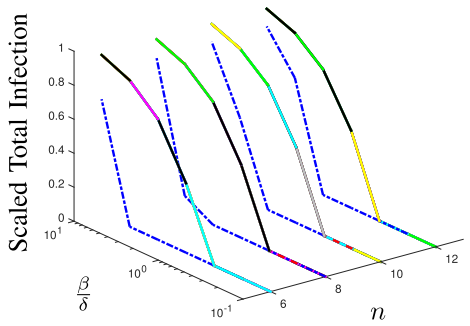


Fig. 2. Lines show the average final infection for (1), (2), and (4) with $h = 0.001, 0.01., 0.1, 0.5, 0.9, 1$ in blue, red, green, magenta, cyan, black, yellow, and gray, respectively. All models are simulated with the same graph structure of n nodes in a line, the same initial condition with every node infected, and for the same period of time, 10 000 time steps (except for $h = 0.001$, the simulations were run for 100 000 time steps, because since h was so small it took longer to reach the equilibrium).

Note that (2) and (13), for $k = 0, 1, \ldots, T-1$, are an equivalent reformulation of (18).

Learning heterogeneous spread parameters, however interesting, will not help estimate the spread in other areas. Therefore, a homogeneous system should be more informative for some applications. For the Snow data set in Section VI, we will employ the heterogeneous approach, using Corollary 2 and assuming $\beta_i = 1$ for all i values. We will employ homogeneous formulation on the USDA data set in Section VII.

V. Simulations

In this section, we explore the discrepancies between (2), and (1) and (4), and we present a simulation that implements the results of Section IV.

A. Approximation Error

To quantify the error between the approximation in (2) and the full probabilistic 2^n state model in (4) and its continuous-time mean field approximation in (1), we simulate the models over a path graph with every node completely infected initially (different initial conditions performed similarly). The (β, δ) pairs used are [(.1, 1), (.215, 1), (.464, 1), (.5, .5), (1, .464), (1, .215), (1, .1)], and the number of nodes, $n = 6, 8, 10, 12$. We simulate (2) for $h = 0.001, 0.01., 0.1, 0.5, 0.9, 1$. Due to the constraints of Assumption 3 and the presence of numerical error, several of the simulations of (2) failed, namely, the tuples $(\beta, \delta, h) = (1, .464, .9), (1, .464, 1), (1, .215, 1), (1, .1, 1)$.

The comparisons of the simulation results depicted by the final average infection are in Fig. 2. As the simulations indicate, the model in (2) is very similar to (1), and, consistent with previous work [9], [13], (1) is different from (4) in the cases, where δ and β are close to each other in value. Therefore, (2) has similar disadvantages to (1) as an approximation of (4), which is logical, since (2) is an approximation of (1).

To better quantify the quality of (2) as an approximation of (1), we plot the Euclidean distance (2-norm) between the final states of (1) and (2) with all of the different h values

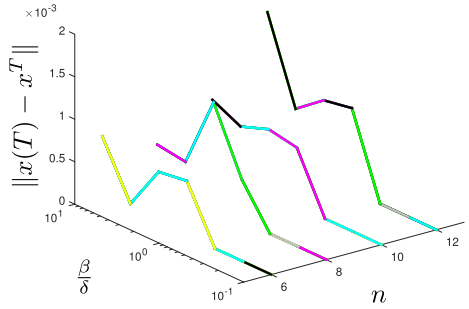


Fig. 3. Lines show the 2-norm error between the final states of (1) and (2) with the different $h = 0.001, 0.01., 0.1, 0.5, 0.9, 1$ values in green, magenta, cyan, black, yellow, and gray, respectively. All models are simulated with the same graph structure of n nodes in a line and the same initial condition with every node infected.

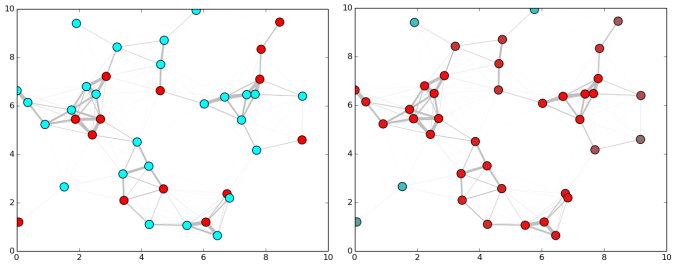


Fig. 4. This virus system follows (2) with $\beta = 1$, $\delta = 0.1$, $h = 0.1$, and A depicted by the edges. Teal indicates healthy or susceptible, while red indicates infected. For a video of this simulation please, see http://youtu.be/JhU1mEvlV-g. (a) System at time zero. (b) System at time 100.

in Fig. 3. As can be seen, the error is quite low. Therefore, we conclude that (2) is a good model when Assumption (3) is met. In addition, (2) is a decent approximation of (4) when δ and β are not too similar in magnitude and a good approximation of (1).

B. Estimation Simulation

We present here a simulation that implements the parameter estimation results from Section IV to see how they perform in practice with clean and noisy data. While the data used in this section are generated in MATLAB, the insights gained from the exercises here contribute toward our approach using the USDA data set in Section VII.

Consider a system with 40 agents, with a random set of initially infected agents, where $\beta = 1$, $\delta = 0.1$, $h = 0.1$, and the weighting matrix A is determined by the agents' relative positions given by z_i, that is, for radius $r = 2$ and $i \neq j$

$$a_{ij}(t) = \begin{cases} e^{-\|z_i - z_j\|^2}, & \text{if } \|z_i - z_j\| < r \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

See Fig. 4 for plots of the initial and final conditions. Assuming that the correct value for h and the A matrix are known, using (13) exactly recovers β and δ. If only two time-steps are used, the exact spread parameters can be recovered, consistent with Theorem 3. Using (13) with an incorrect h value to recover β and δ gives incorrect values for β and δ, but results in the right proportion between the two, consistent with Corollary 1. If the system is at the endemic state, the proportion between the spread parameters can be solved exactly using Corollary 2.

TABLE I

AVERAGES OF 1000 ESTIMATES FOR THE SPREAD PARAMETERS GIVEN ZERO-MEAN GAUSSIAN MEASUREMENT NOISE WITH STANDARD DEVIATION σ. THE PARAMETERS $\hat{\beta}$ AND $\hat{\delta}$ ARE CALCULATED USING THE NOISY DATA. THE PARAMETERS $\tilde{\beta}$ AND $\tilde{\delta}$ ARE CALCULATED USING THE NOISY DATA RESTRICTED TO THE INTERVAL [0,1]. RECALL THAT THE ORIGINAL PARAMETERS WERE $(\beta, \delta) = (1, 0.1)$

σ	$\hat{\beta}$	$\hat{\delta}$	$\tilde{\beta}$	$\tilde{\delta}$
0.01	1.0473	0.1074	1.0412	0.1064
0.02	1.1840	0.1290	1.1708	0.1269
0.03	1.3976	0.1629	1.3654	0.1587
0.04	1.6645	0.2056	1.5977	0.1984
0.05	1.9679	0.2546	1.8532	0.2441

TABLE II

AVERAGES OF 1000 ESTIMATES FOR THE SPREAD PARAMETERS USING $\hat{A}'$ IN (20). RECALL THAT THE ORIGINAL PARAMETERS WERE $(\beta, \delta) = (1, 0.1)$

σ	$\hat{\beta}$	$\hat{\delta}$
0.01	1.0721	0.0904
0.02	1.1363	0.0817
0.03	1.2012	0.0754
0.04	1.2685	0.0712
0.05	1.3336	0.0678

When we add measurement noise

$$y^{k+1} = x^{k+1} + v^{k+1}$$

where y^{k+1} is the measurement, x^{k+1} is from (2), and each v_i^{k+1} is an i.i.d. zero-mean Gaussian random variable with standard deviation σ, and use the same estimation technique from Theorem 3, the results are inaccurate; with every 0.01 increase in σ, the estimation result for β is off by 6%–7% (and always in the positive direction), and δ (always) decreases by about 10% to about 5%. See Table I for the exact values of the average values of the estimates of β and δ for a thousand runs each, for different standard deviations σ. Future work will consider maximum likelihood-type estimators.

We also added zero-mean Gaussian noise to the A matrix, therefore, allowing for uncertainty in the edge weights, as follows:

$$\hat{A} = A + \Delta$$

where Δ is a matrix of i.i.d. zero-mean Gaussian random variables. We simulated the system with $\hat{A}'$, where

$$\hat{a}'_{ij} = \begin{cases} \hat{a}_{ij}, & \text{if } \hat{a}_{ij} \in [0, 1] \\ 1, & \text{if } \hat{a}_{ij} > 1 \\ 0, & \text{if } \hat{a}_{ij} < 0. \end{cases} \tag{20}$$

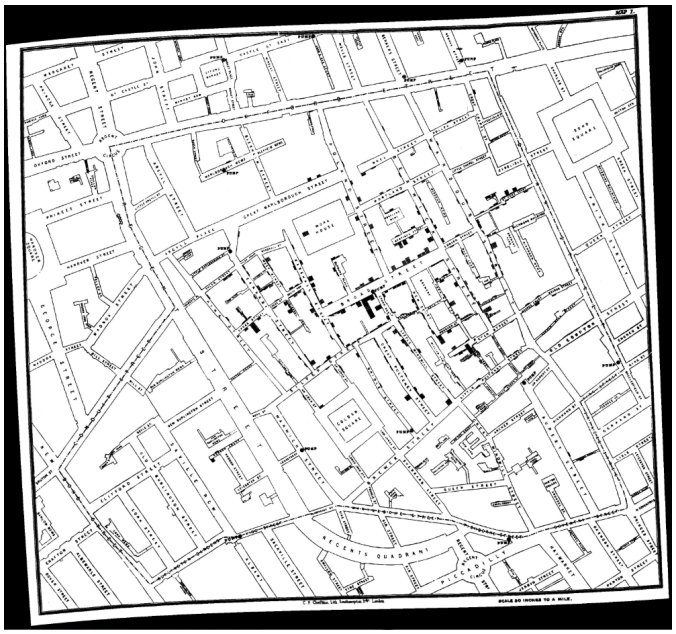


Fig. 5. This is the map of cholera spread in London in 1854 compiled by Snow [22].

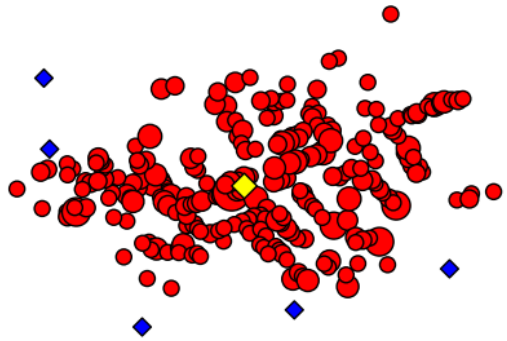


Fig. 6. Digitization of Fig. 5. The uncontaminated water pumps are depicted by blue diamonds, the contaminated water pump by the yellow diamond, and household deaths by red dots with the diameters scaled by the number of deaths, respectively.

The restrictions on $\hat{A}'$ are imposed so that Assumptions 2 and 3 are not violated, and the state does not lose its meaning of being the proportions of infected subpopulations. The average of 1000 runs of estimating the spread parameters is given in Table II.

VI. Validation: Snow Data Set

In this section, we employ the foundational cholera data set collected by Snow [22] for the validation of the model in (2).

A. Snow Data Set

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 Fig. 5, the 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, and not through the air, as was the common belief of those times. These data are plotted in Fig. 6, with diamonds

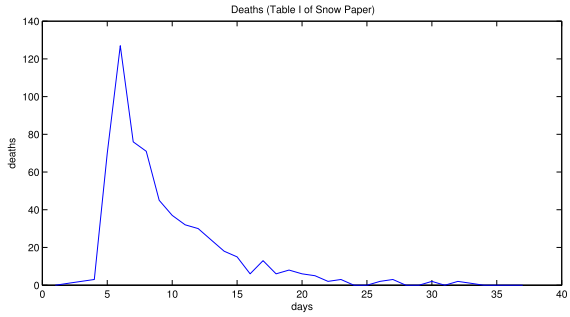


Fig. 7. Deaths per day in the Soho District of London in 1854 compiled by Snow [22].

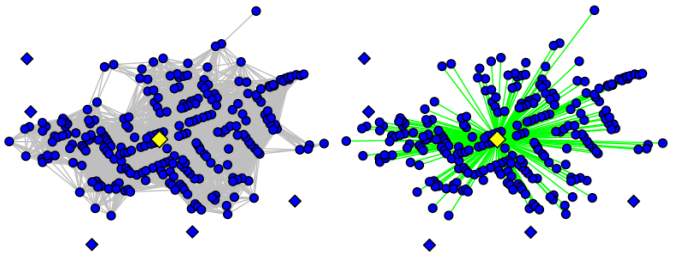
indicating the water pumps and red dots indicating deaths. The data set is comprised of 250 households with at least one death per household. Snow also documented the cumulative deaths per day in [22, Table I], plotted in Fig. 7. The exact times of the deaths for each address are not recorded. There are 616 total cumulative deaths, but the total number of deaths on the map is 489. Therefore, there is a discrepancy of 127 deaths, whose household addresses are not included in the map. For the validation of the model in (2), we use the proportion of deaths in the households as the state of the disease spread system.

B. Spread Validation

For the validation, there are three cases: 1) allowing cholera to spread through the air via nearest neighbor connection; 2) incorporating nearest neighbor connections and direct connections from the Broad Street pump; and 3) only allowing the pump to affect every relevant household. We make various assumptions in order to employ model (2). Each household with a death recorded by Snow in the map in Fig. 5 corresponds to a node in the model. The last node in the model corresponds to the contaminated pump, the one on Broad Street. The healthy water pumps are not included 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 [22]. 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. Despite these (and subsequent) relatively strong assumptions, the validation results are quite promising.

The state of the system, x^k, is the proportion of total deaths in each household up to time k. There were three attempts made to capture the behavior of the epidemic that used different graph structures and different household sizes to calculate the endemic equilibrium.

The endemic equilibrium of the system, which we call x^*, was calculated from the data in Fig. 6, for the first two attempts, 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 in any single household was 15.



(a) (b)

Fig. 8. Initial condition of simulations with graph structures. Blue circles: healthy households. Yellow diamond: infected pump. Note that $A^{(1)}$ and $A^{(3)}$ are portrayed, and that $A^{(2)}$ is the result of the union of the two sets of edges. (a) $A^{(1)}$ from (22). (b) $A^{(3)}$ from (24).

For the third attempt, we approximated the household sizes using [26, Fig. 1] (see Table III). The last element of x^*, corresponding to the pump, was set to (19/20) (alternatively, setting it to 1 makes the corresponding δ_i equal zero, which is clear by Corollary 2, and only slightly changes the rest of the estimated spread parameters).

We employed Corollary 2 to calculate the (δ_i/β_i) values. In the simulation, $\beta_i = 1$ for all i and h is chosen as large as possible while still meeting Assumption 3 (note that a larger h makes the system evolve more quickly and, therefore, renders a single time step closer to a full day). Recall that the Broad Street pump corresponds to the last agent in the model (agent n). For the initial condition in the simulations, we assume that the Broad Street pump is infected and all the households are healthy

$$x^0 = [0 \quad \cdots \quad 0 \quad 1]^\top. \tag{21}$$

This initial condition is shown in Fig. 8, where the contaminated pump is depicted as a yellow diamond. As a consequence of these assumptions, our tuning parameter for adjusting the estimated δ_i parameters, and consequently the spread behavior, is the connectivity matrix A.

For the case that captures the general belief of the era that cholera spreads through the air, we chose a graph structure that allows for local mixing. That is, we designed $A^{(1)}$ such that

$$a_{ij}^{(1)} = \begin{cases} 1, & \text{if } \|z_i - z_j\| < r \\ 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

where z_i is the location of household i and r is the smallest number such that the graph remains connected [shown in Fig. 8(a)]. Using the δ_i parameters derived from Corollary 2 using $A^{(1)}$ (again $\beta_i = 1 \ \forall i$), the system was simulated according to (2). To meet the constraints of Assumption 3, we set $h = (1/175)$. This simulation resulted in the distribution of deaths shown in Fig. 9; this plot was created by multiplying the state of the system, i.e., the proportion of deaths in each household up to that point, by the household sizes (assumed to be 20), rounding to the nearest integer, taking the difference between the states of each time step (since the state represents cumulative number of deaths up to that point), and then summing every three time series points (due to the small h value), therefore assuming that each time series point corresponds to one-third of a day. Note that for this simulation, and similarly for the subsequent simulations, zeros were added to the beginning of the simulation data to align the peaks of the simulations with the peak of the data set. As would be expected, this graph structure does not capture the behavior of the system, as shown in Fig. 10.

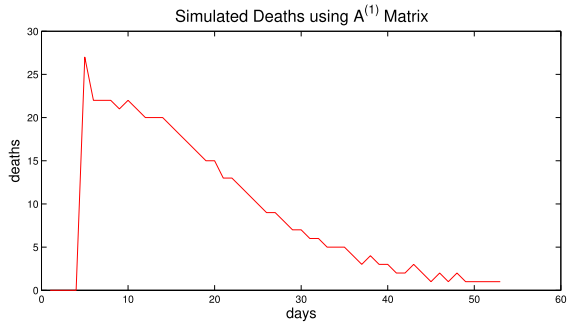


Fig. 9. Simulated data using the estimated parameters from the data in Fig. 6, employing Corollary 2 and $A^{(1)}$ from (22).

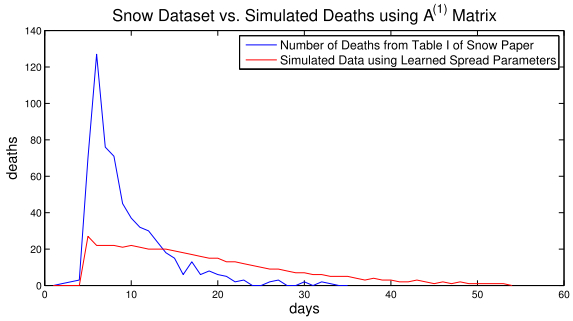


Fig. 10. Comparison of Figs. 7 and 9. Note that the model does not capture the behavior of the system well. The Euclidean distance between the two plots is 146.52, and the infinity norm is 105.

For the case that is more realistic, since it is well known (now) that cholera spreads primarily through contaminated water and that the Broad Street pump was the source of this epidemic, it was assumed that the pump affected everyone. This was done by setting

$$A^{(2)} = [\ A^{(1)}(1\!:\!n, 1\!:\!n-1)\ \upsilon\ ] \tag{23}$$

where $\upsilon = \mathbf{1} \in \mathbb{R}^n$ and the notation $A^{(1)}(1:n, 1:n-1)$ indicates all of the $A^{(1)}$ matrix except the last column. The system was simulated using the δ_i parameters derived from Corollary 2 using $A^{(2)}$, and again setting $h = (1/175)$. The resulting distribution of deaths is shown in Fig. 11 (created similar to Fig. 9). Note that the shape is very similar to the original data set from [22], shown in Fig. 7, capturing the behavior of the true epidemic.

Plotting the distributions from Figs. 7 and 11 on the same plot for comparison in Fig. 12 shows that they are not identical. One of the reasons for the discrepancy is, as noted in Section VI-A, that the total number of deaths in the map (see Fig. 5), used to derive x^* and consequently the spread parameters and the simulation, is 489, and the total number of deaths in [22, Table I], used to create the distribution

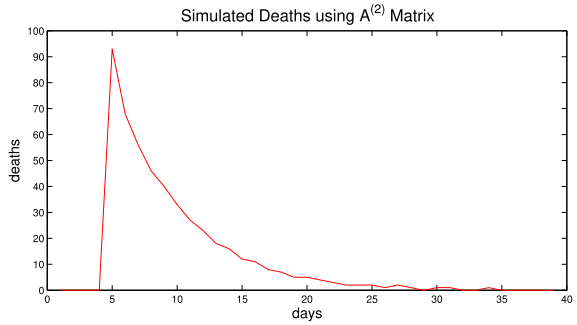


Fig. 11. Simulated data using the estimated parameters from the data in Fig. 6, employing Corollary 2 and $A^{(2)}$ from (23).

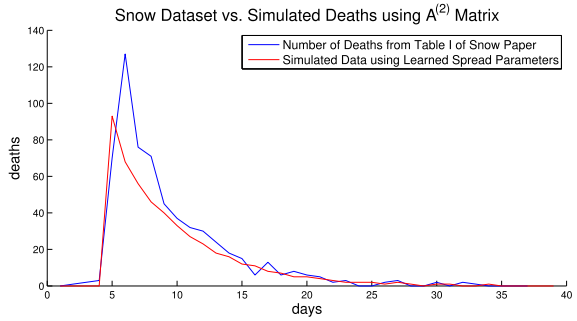


Fig. 12. Comparison of Figs. 7 and 11. Note that there is a difference in the magnitude, but the general shapes are very similar. The Euclidean distance between the two plots is 72.77, and the infinity norm is 59. One of the reasons for this discrepancy is due to the fact that we used the spatial data set in Figs. 5 and 6, which had only 489 documented deaths, while the cumulative data from [22, Table I], shown in Fig. 7 and the blue line in Fig. 14, has a total of 616 deaths. The difference of 127 has caused the discrepancy.

TABLE III

ESTIMATES FOR HOUSEHOLD SIZES FROM [26, FIG. 1] USED IN THE SIMULATION WITH $A^{(3)}$. * THE WORKHOUSE POPULATION WAS SET TO 403

Household Sizes	
Range in [26]	Estimate
0-4	4
5-9	7
10-14	12
15-24	20
24-403	25*

of deaths over days in Fig. 7, is 616. Therefore, the lack of address information for the additional 127 deaths results in this inaccuracy. However, the largest discrepancy occurs near the peak of the epidemic, when people were arriving at hospitals too sick to provide their addresses [22]. Nevertheless, the results are very promising, showing that the model in (2) captures the behavior of the cholera epidemic from John Snow's 1854 data set quite well.

For the third case, heterogeneous household sizes were instead assumed, using [26, Fig. 1] to approximate these values (see Table III). We also 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

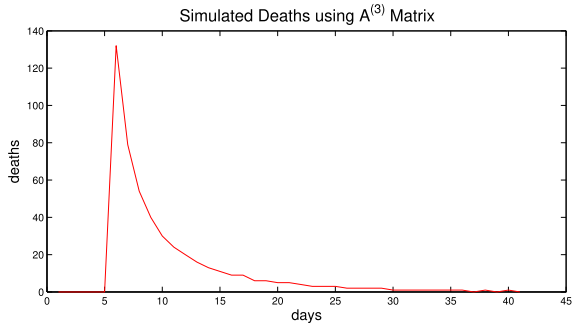


Fig. 13. Simulated data using the estimated parameters from the data in Fig. 6, employing Corollary 2 and $A^{(3)}$ from (24). A video of the spread of the simulation can be found at http://youtu.be/PXqyce7zZFM.

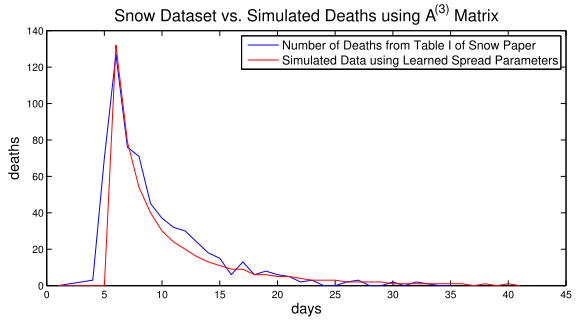


Fig. 14. Comparison of Figs. 7 and 11. Note that there is a difference in the magnitude, but the general shapes are very similar. The Euclidean distance between the two plots is 75.16, and the infinity norm is 70. One of the reasons for this discrepancy is due to the fact that we used the spatial data set in Figs. 5 and 6, which had only 489 documented deaths, while the cumulative data from [22, Table I], shown in Fig. 7 and the blue in this plot, have a total of 616 deaths. The difference of 127 has caused the discrepancy.

the pump to the workhouse was set to $(1/10)$ (corresponding to the 208th index), because the workhouse had its own well and only a small fraction of the 403 residents drank from the Broad Street pump [22] (by choosing $(1/10)$ we assume that approximately 10% of the residents drank water from the Broad Street pump). Therefore, we have

$$A^{(3)} = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & \vdots \\ 0 & 0 & \ddots & 0 & \frac{1}{10} \\ 0 & 0 & \dots & 1 & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \tag{24}$$

[shown in Fig. 8(b)]. Using the δ_i parameters derived from Corollary 2 using $A^{(3)}$, the system was simulated setting $h = (1/30)$. The distribution of the deaths is shown in Fig. 13. As a result of the larger h value, no aggregation of the data was required; the plot shows the complete simulated data set. For completeness, a link to a video of this simulation is included in the caption of Fig. 13. Simulations showed that as long as the edge weight corresponding to the workhouse was less than or equal to 0.45, the results were very similar.

Plotting the distributions from Figs. 7 and 13 on the same plot for comparison in Fig. 14 shows that we capture the

behavior of the outbreak quite well. 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 (2) captures the behavior of the cholera epidemic from John Snow's 1854 data set very well. Note that by both error metrics, the simulation from the second attempt in Fig. 12 outperforms this last attempt but recalls it required summing every three data points. Therefore, it can be argued that the last attempt captures the behavior of the data set best since no summing was required. In addition, note that the fact that $A^{(3)}$ from (24) 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 via the air, which is known to be true today.

VII. Validation: USDA Data Set

The goal of this section is to study whether variation in the spatial pattern of farmers' enrollment in ACRE during 2009–2012 follows the spreading processes presented in Section II. As we elaborate in the following, ACRE is a complex program, making the experience and knowledge of early adopters likely to spread by word of mouth through social and professional networks. For this data set, we assume homogeneous spread parameters, that is, β and δ are the same for all nodes.

A. USDA Data Set

The characteristics of the ACRE program make it a good candidate to empirically test the model of spreading. Farmers rely on the experience of neighbors in the adoption of new or complex technologies [30]–[32]. As we elaborate in the following, ACRE is a complex program. Social and professional networks will likely facilitate the spread of information about the ACRE program from the experiences of early adopters.

The ACRE program was introduced by the Food, Conservation and Energy Act of 2008 (2008 Farm Bill). Initial enrollment was unexpectedly low, in part because of the program's complexity [43]. The ACRE payment a_{ij}^k for year k is calculated by the following formula:

$$a_{ij}^k = \phi \frac{\hat{v}_{ij}^k}{\hat{v}_{\sigma j}^k} \min\left\{ \left(g_{\sigma j}^k - r_{\sigma j}^k\right), \frac{g_{\sigma j}^k}{4} \right\} \min\left\{\rho_{ij}^k, b_{ij}^k\right\} \mathbb{1}\left(r_{ij}^k < g_{ij}^k\right) \mathbb{1}\left(r_{\sigma j}^k < g_{\sigma j}^k\right) \quad (25)$$

where i is the farm index, j is the crop or commodity that subsidy corresponds to, ϕ is a constant scaling factor (equal to 0.85), σ indicates the state (e.g., Idaho), and the benchmark yield (also known as the Olympic yield) is

$$\hat{v}_{ij}^k = \frac{1}{3}\left[\sum_{l=1}^{5} v_{ij}^{k-l} - \max\{\Upsilon_{\iota j}\} - \min\{\Upsilon_{\iota j}\}\right]$$

where $v_{\iota j}^{k-l}$ is the crop yield in year $k-l$; the set $\Upsilon_{\iota j} = \{v_{ij}^{k-1}, \ldots, v_{ij}^{k-5}\}$, for $\iota \in \{i, \sigma\}$; the farm and state guaranteed revenues per acre are $g_{ij}^k = \hat{v}_{ij}^k \bar{\bar{p}}_j^k$ and $g_{\sigma j}^k = .9\hat{v}_{\sigma j}^k \bar{\bar{p}}_j^k$, respectively, with $\bar{\bar{p}}_j^k = (1/2)\sum_{l=1}^{2} \bar{p}_j^{k-l}$, where $\bar{p}_j^k$ is the National Average Market Price of crop j; actual revenue per acre is $r_{\iota j}^k = v_{\iota j}^k q_j^k$, with $q_j^k = \max\{0.7l_j^k, \bar{p}_j^k\}$, where l_j^k is the National Loan Rate, which Congress sets in the farm bill; ρ_{ij}^k is the number of acres *planted* with crop j on farm i; b_{ij}^k is the number of acres of crop j on farm i qualifying for the DCP subsidy, which are known as base acres; and $\mathbb{1}(\cdot)$ is the indicator function [27].

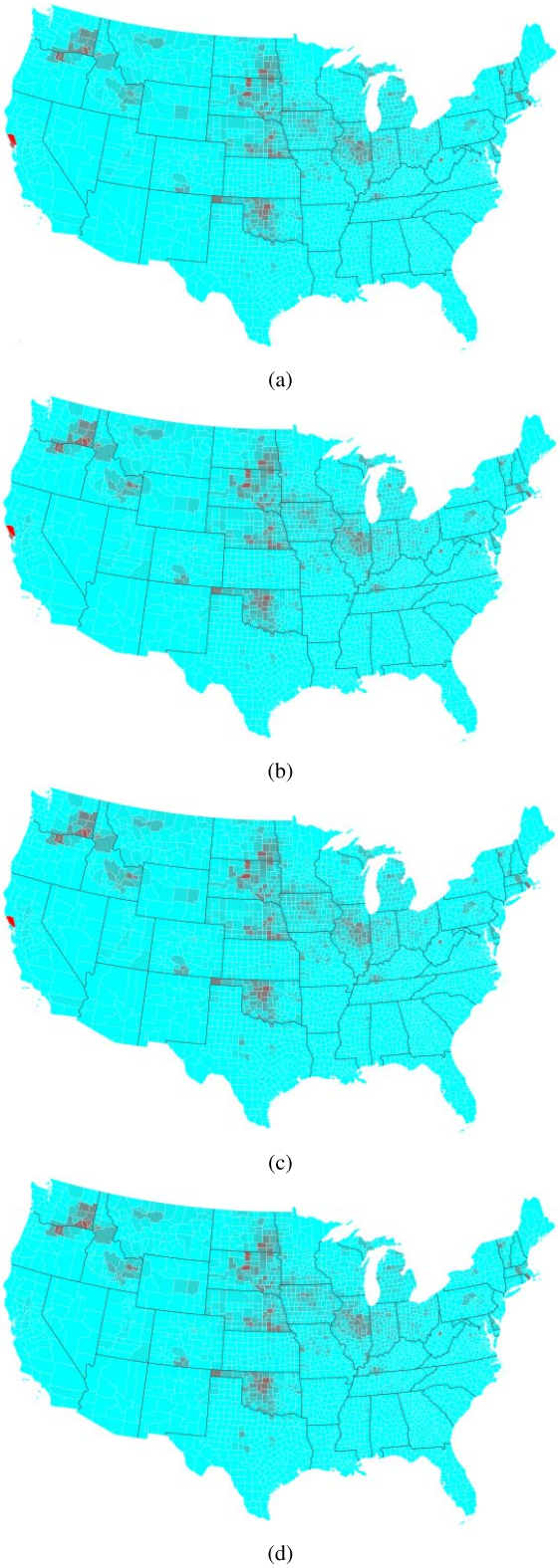
The ACRE program benefits farmers by paying out when the farmers' *actual revenue* is low. In contrast, the Countercyclical Program (CCP), which ACRE replaces, considers current prices, but the payout is determined by the subsidized land's productivity in the early 1980s.

The cost to participate in ACRE is not trivial. By choosing ACRE, farmers must forgo 20% of their annual unconditional subsidy, i.e., direct payment, and 30% of the production subsidy they would receive in the event of low crop prices. Another important consideration is that the decision to participate in ACRE is irreversible. Although farmers must re-enroll in ACRE every year, they cannot switch back to the CCP. Failure to enroll disqualifies farmers from the benefits of ACRE but not the costs. Since switching from ACRE back to CCP is not allowed, we should expect the healing rate δ to be small (or effectively zero) compared to the infection rate β, when we estimate the model parameters from the data.

The data set includes the total annual payments received by each farm in the U.S. for each USDA-sponsored program from the year 2008 to 2012. Each datapoint has a program, payment amount, payment date, contract number, commodity (usually the crop), the farm number, and the customer's (farmer's) identification number and address. The data set allows the opportunity to investigate the spread of the ACRE program through several different networks. Farmer-to-farmer networks could be created from the data by connecting farmer-nodes who receive payments on the same field or live nearby. Alternatively, farms can be aggregated to the county level. The USDA has an office in every county in the United States that distributes subsidies and administers' farm programs locally. Farmers go to these offices (not necessarily their own county's office since an adjacent county's office could be closer) to learn how the subsidy programs work. Therefore, there are strong inner county dependences, since, in addition to receiving the same information at their county offices, farmers meet each other at these offices as well. The approach of aggregating by county allows us to convert the binary decision to enroll in ACRE into a continuous measure of the proportion of eligible farms that enroll in ACRE in each county. The proportion of farms enrolled in ACRE corresponds exactly to the density of infection, facilitating our investigation of the spread of ACRE. For counties where no farms are enrolled in either, the infection state is set to zero. Alaska and Hawaii are omitted. The data for the four years considered can be found in Fig. 15(a)–(d).

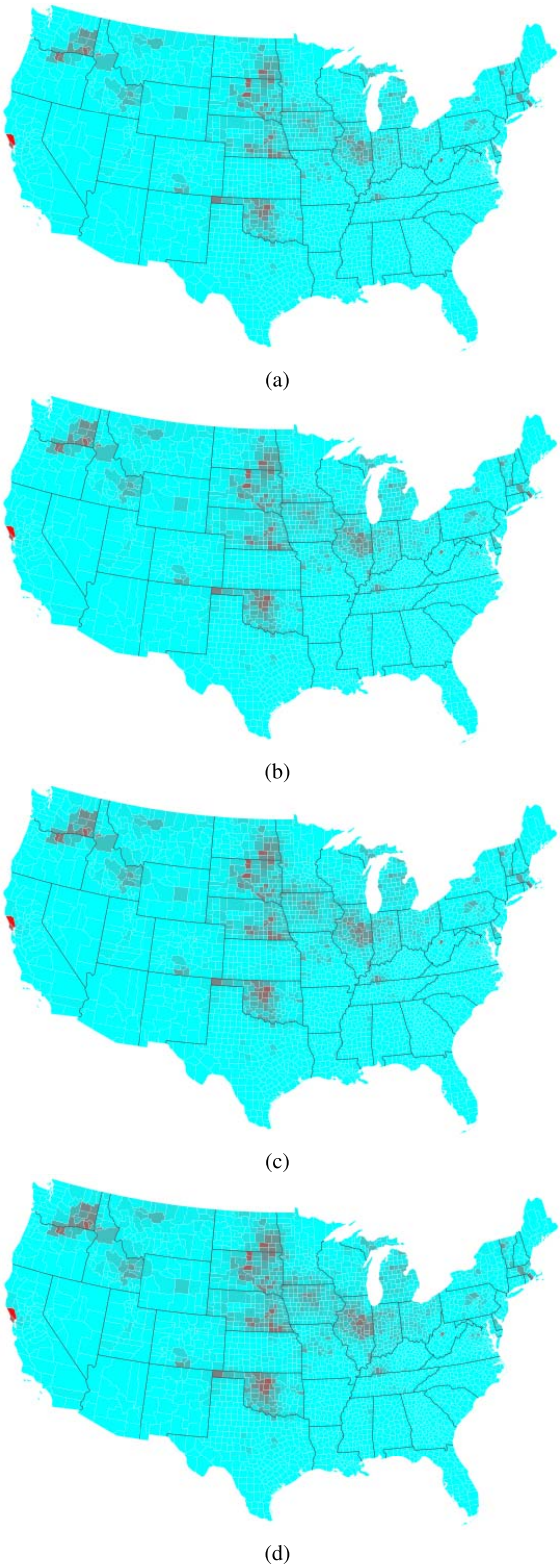
B. USDA Farm Subsidies as a Spreading Process

We apply the estimation techniques presented in Section IV and tested in Section V for the model in (2) on the data



(a) (b) (c) (d)

Fig. 15. Proportion of farms enrolled in the ACRE Program that are enrolled in either ACRE or CCP calculated from the USDA data set. (a) 2009. (b) 2010. (c) 2011. (d) 2012.



(a) (b) (c) (d)

Fig. 16. Simulated data using Fig. 15(a) as the initial condition, simulating using the model in (2) with parameters calculated using the data from Idaho, given in (27). (a) 2009. (b) 2010. (c) 2011. (d) 2012.

presented in Section VII. We estimate the homogeneous model parameters using a subset of the data set, the USDA data from Idaho, and then simulate the spread of ACRE over the whole contiguous United States using the estimated parameters. The adjacency matrices are calculated using the adjacency of counties, that is,

$$a_{ij} = \begin{cases} 1, & \text{if county } i \text{ and county } j \text{ share a border} \\ 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

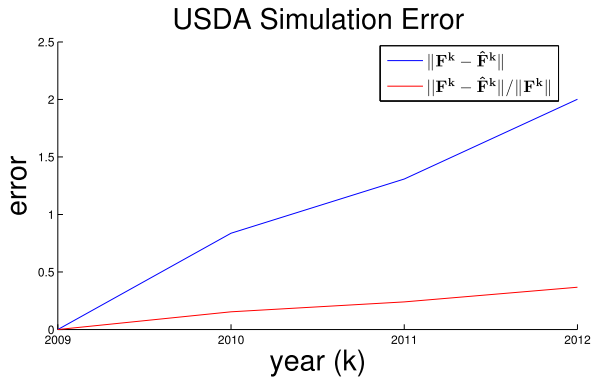


Fig. 17. Error plots for Figs. 15 and 16.

To calculate the adjacency matrix for Idaho, adjacent counties from bordering states were ignored. Substituting the Idaho data set into (13) with $h = 1$ and using the pseudoinverse give the following spread parameters:

$$\begin{bmatrix} \hat{\delta} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} 0.00909176 \\ 0.02237450 \end{bmatrix}. \tag{27}$$

As expected, switching h to the value 0.1 moves the decimal point one place to the right.

To validate the model, we simulate the spread over the contiguous United States using the model in (2) with parameters calculated using the data from Idaho, given in (27), with the data from Fig. 15(a) being used as the initial condition. The simulation results are given in Fig. 16(a)–(d). The scaled error between the data set, $\mathbb{F}$, and the simulated data, $\hat{\mathbb{F}}$, using the Frobenius norm is

$$\frac{\|\mathbb{F} - \hat{\mathbb{F}}\|_F}{\|\mathbb{F}\|_F} = \frac{2.5331}{10.8646} = 0.2332$$

showing that the system has approximately 23% error. For completeness, in Fig. 17, we include a plot of the error for each time step (year), $\|\mathbb{F}^k - \hat{\mathbb{F}}^k\|$ and $(\|\mathbb{F}^k - \hat{\mathbb{F}}^k\|/\|\mathbb{F}^k\|)$. While the model does not perfectly fit the data, it does seem to give some insight into the behavior of the system. Therefore, if the USDA wanted to test a pilot program in a certain region of the country, for example, Idaho, the resulting behavior could give some insight into how the whole country would react. The four time steps (years) do not allow the system to reach the equilibrium state, and so the behavior depends significantly on the initial condition. Therefore, given the model learned from a pilot program, the USDA could determine the best counties to target for advertising of the new subsidy programs, assuming they wanted to maximize adoption of the new program.

VIII. Conclusion

We have investigated the relationship between several different spread models. We have provided necessary and sufficient conditions for uniqueness of the healthy equilibrium and proved the existence of an endemic state under certain conditions. We have also provided a necessary condition for asymptotic stability of the healthy state. We have presented necessary and sufficient conditions for estimating discrete-time spread models from data. We have validated a discrete-time SIS virus spread model using John Snow's Seminal cholera data set with very good results. We have also used a USDA data set to validate the same model by simulating the spread of farming subsidies among farms/farmers aggregated by county.

In the future work, we would like to provide further analysis on the endemic state of the system. We would like to further study identification of the spread model allowing noise in the data. We would also like to find additional data sets to help further validate the SIS spread models. Finally, we would like to employ the results herein to develop effective control techniques to mitigate the spread of disease in real systems.

Acknowledgment

The authors would like to thank A. Shivashankar from the University of Illinois at Urbana–Champaign (UIUC) for helping with the data processing of the Snow data set and S. Gao from UIUC for helping with the data processing of the USDA data set. All material in this paper represents the position of the authors and not necessarily that of the NSF or the USDA.

References

[1] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading in real networks: An eigenvalue viewpoint," in *Proc. 22nd Int. Symp. Reliable Distrib. Syst.*, Oct. 2003, pp. 25–34.

[2] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, "Epidemic thresholds in real networks," *ACM Trans. Inf. Syst. Secur.*, vol. 10, no. 4, p. 1, 2008.

[3] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, "Discrete-time Markov chain approach to contact-based disease spreading in complex networks," *Europhys. Lett.*, vol. 89, no. 3, p. 38009, 2010.

[4] H. J. Ahn and B. Hassibi, "Global dynamics of epidemic spread over complex networks," in *Proc. 52nd IEEE Conf. Decis. Control (CDC)*, Dec. 2013, pp. 4579–4585.

[5] H. J. Ahn and B. Hassibi, "On the mixing time of the SIS Markov chain model for epidemic spread," in *Proc. 53rd Annu. Conf. Decision Control*, Dec. 2014, pp. 6221–6227.

[6] K. Paarporn, C. Eksin, J. S. Weitz, and J. S. Shamma, "Epidemic spread over networks with agent awareness and social distancing," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep./Oct. 2015, pp. 51–57.

[7] S. Han, V. M. Preciado, C. Nowzari, and G. J. Pappas, "Data-driven network resource allocation for controlling spreading processes," *IEEE Trans. Netw. Sci. Eng.*, vol. 2, no. 4, pp. 127–138, Oct./Dec. 2015.

[8] N. J. Watkins, C. Nowzari, and G. J. Pappas, "Inference, prediction and control of networked epidemics," in *Proc. Amer. Control Conf. (ACC)*, May 2017, pp. 5611–5616.

[9] P. V. Mieghem, J. Omic, and R. Kooij, "Virus spread in networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 1, pp. 1–14, Feb. 2009.

[10] V. M. Preciado, M. Zargham, C. Enyioha, A. Jadbabaie, and G. J. Pappas, "Optimal resource allocation for network protection against spreading processes," *IEEE Trans. Control Netw. Syst.*, vol. 1, no. 1, pp. 99–108, Mar. 2014.

[11] P. E. Paré, C. L. Beck, and A. Nedić, "Stability analysis and control of virus spread over time-varying networks," in *Proc. 54th IEEE Conf. Decision Control (CDC)*, Dec. 2015, pp. 3554–3559.

[12] A. Khanafer, T. Başar, and B. Gharesifard, "Stability of epidemic models over directed graphs: A positive systems approach," *Automatica*, vol. 74, pp. 126–134, Dec. 2016.

[13] P. E. Paré, C. L. Beck, and A. Nedić, "Epidemic processes over time-varying networks," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1322–1334, Sep. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/7931651/, doi: 10.1109/TCNS.2017.2706138.

[14] M. J. Keeling *et al.*, "Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape," *Science*, vol. 294, no. 5543, pp. 813–817, 2001.

[15] H. Miao, X. Xia, A. S. Perelson, and H. Wu, "On identifiability of nonlinear ODE models and applications in viral dynamics," *SIAM Rev.*, vol. 53, no. 1, pp. 3–39, 2011.
[16] L. Xue, H. M. Scott, L. W. Cohnstaedt, and C. Scoglio, "A network-based meta-population approach to model Rift Valley fever epidemics," *J. Theor. Biol.*, vol. 306, pp. 129–144, Aug. 2012.
[17] A. Kolesnichenko, B. R. Haverkort, A. Remke, and P.-T. de Boer, "Fitting a code-red virus spread model: An account of putting theory into practice," in *Proc. 12th Int. Conf. Design Reliable Commun. Netw. (DRCN)*, Mar. 2016, pp. 39–46.
[18] Y. Wan, S. Roy, and A. Saberi, "Network design problems for controlling virus spread," in *Proc. 46th IEEE Conf. Decis. Control (CDC)*, Dec. 2007, pp. 3925–3932.
[19] Y. Wan, S. Roy, and A. Saberi, "Designing spatially heterogeneous strategies for control of virus spread," *IET Syst. Biol.*, vol. 2, no. 4, pp. 184–201, Jul. 2008.
[20] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Rev. Mod. Phys.*, vol. 87, no. 3, p. 925, 2015.
[21] C. Nowzari, V. M. Preciado, and G. J. Pappas, "Analysis and control of epidemics: A survey of spreading processes on complex networks," *IEEE Control Syst.*, vol. 36, no. 1, pp. 26–46, Feb. 2016.
[22] J. Snow, *On the Mode of Communication of Cholera*. Cumberland Lodge, U.K.: John Churchill, 1855.
[23] R. Bonita, R. Beaglehole, and T. Kjellström, *Basic Epidemiology*. Geneva, Switzerland: World Health Org., 2006.
[24] World Health Organization. (2018). *Weekly Epidemiological Bulletin W14 2018*. [Online]. Available: http://www.emro.who.int/images/stories/yemen/week_14.pdf?ua=1
[25] N. Einbinder. *How Yemen's Cholera Outbreak Became the Fastest Growing in Modern History*. Accessed: Oct. 20, 2017. [Online]. Available: https://www.pbs.org/wgbh/frontline/article/how-yemens-cholera-outbreak-became-the-fastest-growing-in-modern-history/
[26] N. Shiode, S. Shiode, E. Rod-Thatcher, S. Rana, and P. Vinten-Johansen, "The mortality rates and the space-time patterns of John Snow's cholera epidemic map," *Int. J. Health Geograph.*, vol. 14, no. 1, p. 21, 2015.
[27] USDA, *2009 Average Crop Revenue Election (ACRE) Program: Fact Sheet*. Accessed: Jul. 2016. [Online]. Available: https://www.fsa.usda.gov/Internet/FSA_File/acre.pdf
[28] USDA, *Direct and Counter-Cyclical Payment (DCP) Program*. Accessed: Jul. 2016. [Online]. Available: https://www.fsa.usda.gov/Internet/FSA_File/dcp2008.pdf
[29] D. Sunding and D. Zilberman, "The agricultural innovation process: Research and technology adoption in a changing agricultural sector," *Handbook Agricult. Econ.*, vol. 1, pp. 207–261, 2001.
[30] A. D. Foster and M. R. Rosenzweig, "Learning by doing and learning from others: Human capital and technical change in agriculture," *J. Political Economy*, vol. 103, no. 6, pp. 1176–1209, 1995.
[31] T. G. Conley and C. R. Udry, "Learning about a new technology: Pineapple in Ghana," *Amer. Econ. Rev.*, vol. 100, no. 1, pp. 35–69, 2010.
[32] E. Duflo, M. Kremer, and J. Robinson, "Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya," *Amer. Econ. Rev.*, vol. 101, no. 6, pp. 2350–2390, Oct. 2011.
[33] P. E. Paré, B. E. Kirwan, J. Liu, T. Başar, and C. L. Beck, "Discrete-time spread processes: Analysis, identification, and validation," in *Proc. Annu. Amer. Control Conf.*, Jun. 2018, pp. 404–409.
[34] A. Fall, A. Iggidr, G. Sallet, and J. J. Tewa, "Epidemiological models and Lyapunov functions," *Math. Model. Natural Phenomena*, vol. 2, no. 1, pp. 62–83, 2007.
[35] K. Atkinson, *An Introduction to Numerical Analysis*. Hoboken, NJ, USA: Wiley, 2008.
[36] T. Zhou, J.-G. Liu, W.-J. Bai, G. Chen, and B.-H. Wang, "Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 5, p. 056109, 2006.
[37] M. Vidyasagar, *Nonlinear Systems Analysis*. Philadelphia, PA, USA: SIAM, 2002.
[38] A. Rantzer, "Distributed control of positive systems," in *Proc. 50th IEEE Conf. Decis. Control*, Dec. 2011, pp. 6608–6611.
[39] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[40] R. S. Varga, *Matrix Iterative Analysis*. New York, NY, USA: Springer-Verlag, 2000.
[41] J. Liu, P. E. Paré, A. Nedić, C. Y. Tang, C. L. Beck, and T. Başar, "On the analysis of a continuous-time bi-virus model," in *Proc. 55th IEEE Conf. Decis. Control (CDC)*, Dec. 2016, pp. 290–295.
[42] J. Liu, P. E. Paré, A. Nedić, C. Y. Tang, C. L. Beck, and T. Başar. (Mar. 2016). "On the analysis of a continuous-time bi-virus model." [Online]. Available: https://arxiv.org/abs/1603.04098
[43] P. D. Mitchell, R. M. Rejesus, K. H. Coble, and T. O. Knight, "Analyzing farmer participation intentions and county enrollment rates for the average crop revenue election program," *Appl. Econ. Perspect. Policy*, vol. 34, no. 4, pp. 615–636, 2012.

Philip E. Paré received the B.S. degree (Hons.) in mathematics and the M.S. degree in computer science from Brigham Young University, Provo, UT, USA, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in ECE with the University of Illinois at Urbana–Champaign, Urbana, IL, USA.

His current research interests include the modeling and control of dynamic networked systems, model reduction techniques, time-varying systems, and using these ideas in conjunction with data to solve real-world problems.

Mr. Paré is a 2017–2018 College of Engineering Mavis Future Faculty Fellow. He was a recipient of the 2017–2018 Robert T. Chien Memorial Award for excellence in research and the Best Talk Award in the Decision and Control Session of the 13th CSL Student Conference, sponsored and judged by Honeywell.

Ji Liu received the B.S. degree in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2006, and the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA, in 2013.

He was a Post-Doctoral Research Associate with the Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL, USA, and the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, USA. His current research interests include distributed control and computation, distributed optimization and learning, multiagent systems, social networks, epidemic networks, and cyber-physical systems.

Carolyn L. Beck received the B.S. degree from California Polytechnic State University, San Luis Obispo, CA, USA, the M.S. degree from Carnegie Mellon University, Pittsburgh, PA, USA, and the Ph.D. degree from the California Institute of Technology, Pasadena, CA, USA, all in electrical engineering.

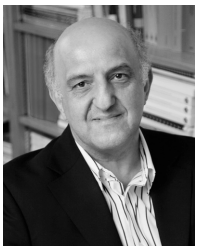
Prior to completing her Ph.D., she was with Hewlett-Packard, Santa Clara, CA, USA, for four years, designing digital hardware and software for measurement instruments. She has held visiting faculty positions at the KTH Royal Institute of Technology, Stockholm, Sweden, Stanford University, Palo Alto, CA, USA, and Lund University, Lund, Sweden. She is currently an Associate Professor with the Industrial and Enterprise Systems Engineering Department, University of Illinois at Urbana–Champaign, Urbana, IL, USA. Her current research interests range from network inference problems to control of anesthetic pharmacodynamics. Her main research interests are model reduction and approximation for the purpose of feedback control design; mathematical systems theory; and clustering and aggregation methods.

Dr. Beck has received national research awards and local teaching awards.

Barrett E. Kirwan received the B.A. degree (Hons.) in economics from Brigham Young University, Provo, UT, USA, in 1998, and the Ph.D. degree in economics from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2005.

He was an Assistant Professor with the University of Maryland at College Park, College Park, MD, USA. He is currently an Adjunct Research Assistant Professor with the Agricultural and Consumer Economics Department, University of Illinois at Urbana–Champaign, Urbana, IL, USA. His current research interests include the distribution of the benefits of agricultural policy and the behavioral effects of crop insurance.

Dr. Kirwan and his co-author received the 2017 Quality of Research Discovery Award from the Agricultural and Applied Economics Association and the 2017 Outstanding American Journal of Agricultural Economics Article Award (declined).



Tamer Başar (S'71–M'73–SM'79–F'83–LF'13) received the B.S.E.E. degree from the Robert College, Istanbul, Turkey, and the M.S., M.Phil., and Ph.D. degrees from Yale University, New Haven, CT, USA.

He is with the University of Illinois at Urbana–Champaign, Urbana, IL, USA, where he is the Swanlund Endowed Chair, a Professor of electrical and computer engineering at the Center for Advanced Study, and Research Professor at the Coordinated Science Laboratory and the Information Trust Institute. He is also the Director of the Center for Advanced Study and the Interim Dean of Engineering. He has over 900 publications in systems, control, communications, networks, and dynamic games, including books on noncooperative dynamic game theory, robust control, network security, wireless and communication networks, and stochastic networked control. His current research interests include stochastic teams, games, and networks; security; and cyber-physical systems.

Dr. Başar is a member of the U.S. National Academy of Engineering and the European Academy of Sciences and a fellow of the International Federation of Automatic Control (IFAC) and the Society for Industrial and Applied Mathematics. He has served as Presidents of the IEEE Control Systems Society (CSS), the International Society of Dynamic Games (ISDG), and the American Automatic Control Council (AACC). He has received several awards and recognitions over the years, including the highest awards of the IEEE CSS, IFAC, AACC, and ISDG, the IEEE Control Systems Award, and a number of international honorary doctorates and professorships. He was the Editor-in-Chief of *Automatica* from 2004 to 2014. He is currently editor of several book series.