Homogeneous assumption and the logistic behavior of information propagation

Michael Busch and Jeff Moehlis

Department of Mechanical Engineering, University of California, Santa Barbara, Santa Barbara, California 93106, USA

(Received 29 April 2011; revised manuscript received 15 December 2011; published 6 February 2012)

Many models of disease and rumor-spreading phenomena average the behavior of individuals in a population in order to obtain a coarse description of expected system behavior. For these types of models, we determine how close the coarse approximation is to its corresponding agent-based system. These findings lead to a general result on the logistic behavior of information propagation for networks on both connected graphs with doubly stochastic edge weights, and connected graphs with symmetric edge weights. Moreover, we discuss the appropriateness of the discrete logistic approximation for a few example heterogeneous graph topologies.

DOI: 10.1103/PhysRevE.85.026102 PACS number(s): 89.75.−k, 87.23.Ge

I. INTRODUCTION

Many different phenomena can generally be described as the exchange of information between members of a population, such as the spread of rumors [1–4], ideas [5], computer worms and viruses over the internet [6,7], and most notably the spread of infectious diseases [8–20]. These types of phenomena can be modeled at the population level, the agent level, or somewhere in between. Moreover, the popular work on small-world [21] and scale-free [22] networks have uncovered important structural details of typical populations in these systems. Recent discoveries in the discipline of network science have motivated a second look at how structural details of the population network influence the rate and depth at which information spreads, and how agent-level interactions affect population-level interactions [13–20]. Here we continue along these lines and draw attention to novel ways of quantitatively analyzing how network structure affects the spread of information through a population.

Early descriptions of disease and rumor propagation approximate population-level behavior by low-dimensional ordinary differential equations (ODEs) [1,8–10,12]. However, it is difficult to include the network topology of the propagation medium in these low-dimensional models. A related issue is how local variations in behavior and connectivity should be averaged, which is known as the problem of heterogeneity [12,16,23]. The low-dimensional models implicitly assume that every node in the network is connected to every other node in the network equally [24]. Initial attempts at addressing this problem include proportional mixing techniques, which focused on subdividing the population into smaller homogeneous populations [11]. Recently the inclusion of network topology has been addressed using heterogeneous mean-field approaches that coarse-grain the set of nodes into various degree classes that have similar dynamical properties [13,17–19,25,26]. These methods rely on statistical properties of the network rather than the global structure of the network as a whole, and become computationally costly as the system detail is refined to the agent level.

We shall analyze the spread of information between individuals for simple contagion scenarios [27] and develop an agent-based propagation model that is similar to the probabilistic discrete-time Markov chain studied in Ref. [28]. This agent-based theoretical framework scales to subpopulations of any size [8] and has been shown to generalize heterogeneous mean-field approaches [28], where the subpopulations are typically groups of nodes that have the same out-degree. The advantage to using an agent-based theoretical framework is that it begins with the exact structure of the network and determines the dynamics rather than incorporating the detailed network properties into an already formulated coarse dynamic model, thus allowing one to accurately probe network effects at any scale. Hence, we seek to show that our agent-based contact model is consistent with the well-known low-dimensional mean-field logistic model, and discuss the implications of using a low-dimensional logistic model in place of the agent-based contact model. Because it is popularly taken as granted that low-dimensional models are not accurate representations of agent-based systems [29], our approach will attempt to rigorously quantify the differences between the two representations and uncover the network topologies where the low-dimensional and agent-based models may possibly agree.

Our findings rely on the implementation of algebraic graph theory, which has been extensively applied to the analysis of static network structure [30], particularly since adjacency matrices uniquely define a graph [30] and are computationally efficient mathematical structures [31]. From our agent-based approach to the study of information propagation, we argue that new physical insight is gained by applying the tools of algebraic graph theory to the study of the dynamics on a network. Not only will adjacency matrices allow us to rigorously attribute the size of fluctuations about the population-level solutions to finite-size effects, but they will simultaneously tell us what network structures are necessary for our assumptions to hold. Furthermore, since the mean-field population model implicitly asserts a particular graph structure, i.e., a completely connected graph, we will conclude our discussion by exploring how one can use a mean-field solution for an agent-based system, which may have an arbitrary graph structure, to assert parameter values of a corresponding coarse population model. We propose a novel way in which one can use these inferred parameter values as simple metrics for comparing the graph-dependent propagation dynamics for a set of initially informed nodes.

To communicate these ideas, this manuscript is organized as follows: In Sec. II a coarse-grain model based on the scalar logistic equation is introduced. We present the agent-based contact model in Sec. III; this describes the probability of interaction between nodes on the graph in terms of the exact network topology. Section IV describes how the models developed in Secs. II and III intersect, which leads to a general result for doubly stochastic networks and networks with interaction symmetry that is presented in Sec. V. The logistic behavior of example heterogeneous networks is discussed in Sec. VI, and a summary of our findings is given in Sec. VII.

II. LOGISTIC POPULATION MODEL

For a given population of N agents, suppose each agent can only be a member of one of two sets: the set of S-class individuals (susceptibles) or the set of I-class individuals (informeds). Agents can only go from being susceptible to being informed, and once informed remain informed for all future time. If one were to consider only the transfer events of information propagation, then the SI model is arguably the simplest population-level model that corresponds with information communication mechanisms between members. Including more complicated behavior by allowing members to leave the I-class set, by either reentering the S-class or entering an R-class subpopulation (removed/forget), is unnecessary because these are behaviors that do not typically depend on network structure; it is not necessarily true that one's neighbors' forgetfulness or recovery will directly cause one also to forget or recover. For these reasons, we will simplify our discussion by considering only SI dynamics.

The classic SI model [8,10,12] is expected to be accurate only for systems that exhibit well-mixed behavior because it assumes that (1) the population size is fixed, and (2) the members within a set are indistinguishable from every other member in that set. In discrete time, the rate of infection is proportional to the number of susceptibles and informeds at the previous time step, as well as a transmission rate β, which we shall generally treat as a time-varying expression β_t [32]. By letting I_t be the proportion of informed individuals in a population, the SI dynamics are described by the discrete scalar difference equation

$$I_{t+h} = I_t + h\beta_t I_t(1 - I_t) \tag{1}$$

and the *logistic* function that solves this equation.

This model has the additional assumption that (3) the time step h is small enough such that only one informed agent contacts all of his neighbors during that time step. As noted in Ref. [8], solutions to Eq. (1) are bounded on the unit interval for each initial value on the unit interval only if the coefficient of the $I_t(1 - I_t)$ term, say, $\alpha(t)$, is positive and satisfies the condition

$$\sup_t \alpha(t) \leqslant 1. \tag{2}$$

For Eq. (1), $\alpha(t) = h\beta_t$, and this implies that the step size of Eq. (1) must satisfy $h \leqslant 1/\sup_t \beta_t$.

III. AGENT-BASED MODEL

As an alternative, let us now consider a system of *discrete agents* that are able to share information with each other. The communication pathways between agents can be mapped as a graph $G(V,E)$ of N total agents [24,33], where each uniquely indexed node $i \in \{1,\ldots,N\} \subset V$ of the graph represents a distinct agent, and a directed edge $(i,j) \in E$ connecting two nodes i and j indicates that it is possible for agent i to transmit information to agent j. For example, in epidemiology an agent is a unique person and a parcel of information may be an infectious disease [34], and in the blogosphere an agent is a unique web user while a parcel of information may be a specific rumor about a politician or celebrity [35].

We are interested in the probability that agent i is in possession of a specific parcel of information at discrete time t, which we denote $p_t^{(i)}$. Furthermore, agent i is assumed to communicate with a neighbor, e.g., agent j with $j \neq i$, in such a way that there is a nonzero probability a_{ij} that agent j successfully communicates a parcel of information to agent i in a time period of h. In general, each a_{ij} is a time-varying expression since the graph topology of a given social network is subject to change over long enough periods of time. To keep our discussion simple, we adopt the common assumption that the information of interest spreads through the network faster than significant changes to the network are able to emerge.

Motivated by the mechanism of social media platforms such as Twitter and the Facebook newsfeed, the magnitude of each a_{ij} for a given i is the probability that, in one time step, agent i will be contacted by one of his neighbors j. In this sense the informed agents broadcast the information to their neighbors. Thus, a_{ij} is said to be an element of the weighted adjacency matrix A that uniquely defines the structure of the graph G [33], and each $a_{ij} \in \{[0,1] : \forall i \in \{1,\ldots,N\}, \sum_{j=1}^{N} a_{ij} = 1\}$. Any matrix that has this property is said to be row stochastic.

Recent evidence suggests the probability that an information transfer event occurs is also dependent on the nature of the information itself [35]. For instance, two agents of a network may be in communication, but not necessarily sharing the type of information that one would like to be tracking. Thus, the average probability that the desired information is being transmitted during a given period of time h is given by $h\beta_t$, where β_t is similar to the transmission rate defined for scalar logistic models. We remark that the step size requirement (2) ensures the term $h\beta_t$ satisfies the probability axiom $h\beta_t \in [0,1]$.

It shall be assumed that once an individual is informed (infected), he does not forget (recover) or become silent (removed). Instead, we attribute any time-varying effects of the propagation dynamics to the nature of the information itself through the β_t expression. Given these conditions, the total probability of an arbitrary agent i becoming informed at a given time step $t + h$, denoted by $p_{t+h}^{(i)}$, is

$$p_{t+h}^{(i)} = p_t^{(i)} + \left(1 - p_t^{(i)}\right)\left(\sum_{j=1}^{N} h\beta_t a_{ij} p_t^{(j)}\right). \tag{3}$$

Hence, an agent is informed at time $t + h$ if he is already informed by time t, or the agent is not informed by time t and an informed neighbor successfully transmits the information. The fact that the probability at the next time step depends only on the probability at the current time step indicates that the system of equations expressed by Eq. (3) has the Markov property and is referred to as a Markov chain [36]. We remark that the diagonal elements a_{ii} of the weighted adjacency matrix are necessarily equal to zero. If this were not the case, then a contradiction would occur because then an uninformed agent would be able to spontaneously inform himself. In matrix notation, (3) becomes

$$p_{t+h} = p_t + h\beta_t(I - \mathrm{diag}\{p_t\})Ap_t, \tag{4}$$

where p_t is a column vector whose indices correspond with the agent indices.

Epidemic models of this form have been shown by Monte Carlo simulation to generalize both contact processes and reactive processes [28]. A contact process is a dynamical process where each informed agent stochastically informs just one of his neighbors per time step, while a reactive process is a dynamical process where at least one informed agent stochastically informs all of his neighbors per time step. Hence, by construction, we consider the dynamics of a reactive process.

Since we consider the dynamics of a reactive process, we constrain each S-class individual to only interact with one of his neighbors at a time. This assumption is consistent with the mechanics of simple contagions and applies to situations where information is broadcast, e.g., a radio signal, and each susceptible individual can "listen" only to one broadcasting source at a time so that the communication events are mutually exclusive. In contrast to Eq. (2) of Ref. [28], where elements of the weighted adjacency matrix describe the probability of where a random walker on the network will go next, the elements of the weighted adjacency matrix in our system describe the probability of where the random walker has come from. When comparing these two closely related frameworks, Eq. (3) of this paper can be recovered from Eq. (1) of Ref. [28] by first swapping the index of the product, and then applying De Morgan's law to obtain a series representation.

IV. COMPLETELY CONNECTED SOLUTION

Having introduced both a population-level model and an agent-based model independently, one can rigorously construct the population-level dynamics directly from the agent-level dynamics by asserting the "well-mixed" assumption that is implicit in the population-level dynamics [24] presented in Sec. II. In terms of graph topology, we argue that well-mixedness of a population corresponds to a completely connected graph. A graph is said to be completely connected if every node on the graph shares an undirected edge with every other node on the graph [33], and a network of agents on a completely connected graph is often considered to be "well-mixed" if every agent communicates with every other agent equally [24]. A graph of this type is described as being "homogeneous" because the local graph topology for each node is indistinguishable from that of every other node. For a network of N agents, this implies that elements of the weighted adjacency matrix for a completely connected graph have the following values:

$$a_{ij} = \begin{cases} \frac{1}{N-1}, & i \neq j \\ 0, & i = j \end{cases}.$$

We remark that a model in this framework, as stated, relies on the assumption that (1) the population does not change and (2) the edge weights do not change. The regularity of the adjacency matrix for the well-mixed case allows one to also find upper and lower bounding functions to the solution of Eq. (4). Since the solution to Eq. (3) is positive and monotonically increasing elementwise [8], taking the one norm is identical to summing over all of the elements:

$$\begin{aligned} |p_{t+h}|_1 &= \sum_{i=1}^{N}\left[p_t^{(i)} + h\beta_t\left(1 - p_t^{(i)}\right)\left(\sum_{j\neq i}\frac{1}{N-1}p_t^{(j)}\right)\right] \\ &= |p_t|_1 + h\beta_t|p_t|_1\left(1 - \frac{|p_t|_1}{N-1}\right) + \frac{h\beta_t}{N-1}|p_t|_2^2. \end{aligned} \tag{5}$$

We note that the maximum possible informed population—also known as carrying capacity [9]—of the model (5) is N, rather than $N-1$ as the resemblance of Eq. (5) to the discrete logistic equation might falsely suggest.

It is now possible to compare the graph-based solution of Eq. (4) to that of the traditional susceptible-infected (SI) model for systems containing an arbitrary number of agents. If the cardinalities of the susceptible and infected populations are random variables, then experimental evidence suggests that the scalar variables of the low-dimensional models represent the expected values for the sizes of those sets [8].

For comparison, Eq. (5) can be normalized with respect to the total population to obtain

$$x_{t+h} = x_t + h\beta_t x_t\left(1 - \frac{N}{N-1}x_t\right) + \frac{h\beta_t}{N(N-1)}|p_t|_2^2, \tag{6}$$

where $x_t = |p_t|_1/N$ has the usual interpretation of being the expected probability that an arbitrarily sampled agent is informed. Under this interpretation, one can think of Eq. (6) as a mean-field description of the population. Another interpretation of x_t is that it represents the proportion of informed individuals in a population.

In the thermodynamic limit where the size of the system N approaches infinity, one finds that

$$x_{t+h} = x_t + h\beta_t x_t(1 - x_t). \tag{7}$$

Hence, the dynamics of the SI model and the graph-based model are equivalent in the thermodynamic limit. Given a population of size N, the solutions to the difference equations (1) and (7) are equal when the initial concentration of Eq. (7) is taken to be the proportion of initially informed individuals of Eq. (1). For finite homogeneous populations, however, the solutions are closely bounded by the solutions to

$$x_{t+h} = x_t + h\beta_t\frac{N}{N-1}x_t(1 - x_t) \quad \text{and}$$

$$x_{t+h} = x_t + h\beta_t x_t\left(1 - \frac{N}{N-1}x_t\right),$$

when the following step-size condition is met:

$$h \leqslant \frac{N-1}{N+1}\frac{1}{\sup_t \beta_t}.$$

A proof of this claim is presented in Appendix A. Furthermore, when this step size condition is met, solutions to the discrete logistic equation of a given initial point are bounded above by solutions with greater initial values and bounded below by solutions of lesser initial values. Existence and uniqueness guarantee this feature for the continuous model, but when comparing solutions to the discrete model, this is an important feature for solutions to have because it allows one to know for certain that one solution dominates another.

V. LOGISTIC APPROXIMATION OF DYNAMICS ON CONNECTED GRAPHS

Ultimately, for a mean-field representation, one would like to find a simple scalar equation that is a close approximation to Eq. (4), and determine what structural conditions must exist to allow such a scalar reduction. By approaching this question from the point of view of algebraic graph theory, we find that, for connected graphs, if A is either a doubly stochastic or a symmetric adjacency matrix, then the largest singular value can be used to find the closest rank-one approximation to the original matrix in two-norm, $\|\cdot\|_2$. Thus, by identifying the singular values of A, one can reduce the dimensionality of the system to a simple scalar approximation to Eq. (4). In general, graph topologies that permit doubly stochastic adjacency matrices are known to be contained in the family of strongly connected graphs, and we refer the reader to Ref. [37] for a more detailed technical discussion of this topic. The study of doubly stochastic systems is relevant for engineered systems, where the graph topology is constructed to have this doubly stochastic property. The coordinated control of multi-agent systems [38] and the application of distributed consensus algorithms [39], for example, are often constructed with doubly stochastic communication topologies. Understanding the dynamics of robust information sharing amongst multi-agent systems is presently an ongoing area of research.

Though there exist matrices that are both doubly stochastic and symmetric, it is possible for a matrix to be doubly stochastic without being symmetric, or symmetric without being doubly stochastic. The latter case is more likely to occur naturally, but requires a relaxation of the row stochastic condition. Therefore, we will present the results for doubly stochastic matrices, followed by the results for symmetric matrices. We remark that the ability to utilize the matrix description of the network is critical for performing the scalar reduction in both cases, and we shall first review some important results from linear algebra that will be of use.

Suppose a given matrix A is symmetric, that is, $A = A^T$. For $A \in R^{N\times N}$ and $A = A^T$, it is known that there exists an orthogonal matrix $W \in R^{N\times N}$ that diagonalizes A [40]:

$$A = WDW^T, \quad \text{with} \quad D = \text{diag}\{\lambda_1, \ldots, \lambda_N\}, \tag{8}$$

where $\lambda_i \in R$ is the ith eigenvalue of A such that $|\lambda_i| \geqslant |\lambda_{i+1}|$. Moreover, since the singular values of A are the positive square roots of the eigenvalues of A^TA, the result (8) and the following imply that each singular value of A is the absolute value of an eigenvalue of A:

$$A^TA = A^2 = WD^2W^T, \text{ such that } D^2 = \text{diag}\left\{\sigma_1^2, \ldots, \sigma_n^2\right\},$$

where σ_i is the ith largest singular value of A [40].

Because of the close relationship between the eigenvalues and the singular values of real symmetric matrices, the problem of identifying the largest singular value is equivalent to the identification of the spectral radius, $\rho(A)$. One can appeal to the Perron-Frobenius theorem for row stochastic matrices to determine $\rho(A) = 1$ [33]. Denoting an N-dimensional vector of ones by 1_N, the row stochasticity of A implies that $A1_N = 1_N$ is an eigenvector of A with eigenvalue 1. The normalized eigenvector $1_N/\sqrt{N}$ is then the first column, w_1, of W in Eq. (8). Thus,

$$\sigma_1 = \lambda_1 = 1, \quad \text{and} \quad w_1 = \frac{1}{\sqrt{N}}1_N. \tag{9}$$

With the largest singular value and corresponding eigenvector identified, one is able to determine the closest rank-one approximation in matrix two-norm to an arbitrary adjacency matrix that is both doubly stochastic and symmetric. To see this, suppose $A = A^T \in R^{N\times N}$ and W is an orthogonal matrix that diagonalizes A as in Eq. (8). Then A can equivalently be represented as the series

$$A = \sum_{i=1}^{n} \lambda_i w_i w_i^T, \tag{10}$$

where λ_i is the ith eigenvalue value of A and w_i is the ith column of W. Furthermore, the closest rank-k approximation to A in matrix two-norm is $X = \sum_{i=1}^{k} \lambda_i w_i w_i^T$, for $0 \leqslant k \leqslant \text{rank}(A)$. By recalling that the singular values of a symmetric matrix are the absolute value of its eigenvalues, a more general proof of this statement is provided in Ref. [41], with the symmetry requirement relaxed and replacement of the Frobenius norm by the matrix two-norm.

Using these properties of symmetric doubly stochastic matrices, one can rigorously define how well the scalar logistic model approximates the graph-based model, as proved in Appendix B. We now are able to define how well the scalar logistic model approximates the graph-based model for a doubly stochastic connected network topology from the following statement. Given a system of equations of the form (4) and defined on a doubly stochastic connected network with a reachable population of n members, the solution to the scalar logistic equation

$$\begin{aligned} x_{t+h} &= x_t + h\beta_t x_t(1 - x_t) \\ x_0 &= \frac{|p_0|_1}{N} \end{aligned} \tag{11}$$

approximates the average value of the elements of p_t to an accuracy of order $h\sigma_2$.

When comparing this result to the direct solution of the completely connected case (5), it comes as no coincidence that the second largest singular value for the completely connected adjacency matrix is $(N-1)^{-1}$. To display this fact, Eq. (6) can be written in the form

$$x_{t+h} = x_t + h\beta_t x_t(1 - x_t) + \frac{h\beta_t}{N-1}\left(\frac{|p_t|_2^2}{N} - x_t^2\right).$$

We emphasize that the important feature of the network topology that produces this result is that the weighted adjacency matrix is doubly stochastic.

The error terms in these equations define a bound on the magnitude of fluctuations of mean-field solutions about the logistic solution at each step. The dependence of the error term on the structure of the adjacency matrix alludes to the notion of structural convergence where the error converges to zero as the doubly stochastic or symmetric graph essentially becomes more completely connected in the sense of its matrix two-norm. As a practical example, take an undirected cyclic graph of N nodes, where each node has k neighbors and each edge has a weight of $1/k$. The adjacency matrix of this system is both doubly stochastic and circulant, and its second largest eigenvalue is given by Ref. [42]:

$$\lambda_2 = \sum_{m=0}^{N-1} c_m e^{-i2\pi m/N}. \tag{12}$$

Because the exponential terms of Eq. (12) are symmetric about the real axis, the sum of imaginary terms is zero, and the sum of real terms can be found by doubling the sum of the real terms over the interval $[0,\pi]$. Since $c_m = 0$ where edges do not exist and $c_m = 1/k$ where they do, one looks at the sum of the real terms to obtain the lower bound:

$$\cos\left(\pi \frac{k}{N}\right) \leqslant \lambda_2. \tag{13}$$

It is obvious that for fixed N, each edge added to the system by increasing k will make the system more completely connected. As this system grows, however, the lower bound λ_2 shows that the mean-field solutions will certainly not converge to the logistic solutions if the degree of each node does not increase at the same rate as the population.

Though one might perceive the double stochasticity requirement to be rather strict, requiring interaction symmetry between agents is quite realistic. For information-spreading phenomena that involve direct one-on-one contact between members of a population, e.g., during the spread of diseases or computer viruses, the amount of time two members spend in an interaction is symmetric. When this amount of time is scaled by the total amount of time per period of interest, then one can conceivably obtain a symmetric non-negative weighted adjacency matrix whose row sums are between zero and one. In this case the error term is defined by the largest singular value of the difference between the given adjacency matrix and the lowest rank approximation [rank$(A) = 1$] of a row stochastic matrix, denoted R_1, and whose elements are all N^{-1}. Similar to the procedure used to obtain (11) from Eq. (4), one begins with

$$p_{t+h} = p_t + h\beta_t(\mathrm{I} - \mathrm{diag}\{p_t\})(A + R_1 - R_1)p_t \tag{14}$$

to obtain

$$x_{t+h} = x_t + h\beta_t x_t(1 - x_t) + O(h\|A - R_1\|_2). \tag{15}$$

For example, let us consider chain of linked nodes arranged in a line such that each node has only two neighbors except for the nodes on the ends who each have just one neighbor. The adjacency matrix of this system will have have only nonzero elements on its upper diagonal, lower diagonal, or both. If symmetric interactions occur on this network, then one can apply (15) to this system. In this case suppose each and every interaction takes place for the same proportion of a given time step, e.g., $h/2$ so that each element along the upper and lower diagonal is $1/2$ and each row sum lies on the unit interval. As defined in Ref. [42], the structures of A and $(A - R_1)$ are both *banded Toeplitz matrices*, which are a class of matrices that asymptotically converge to their *circulant* analogs. Here, the linear chain of linked nodes yields a circulant system, e.g., A_C, by connecting the two ends of the chain. Hence, one can generally argue that for large chains of linked nodes, the results of Eqs. (12) and (13) indicate that the mean-field solution for this system does not converge to a logistic solution in the thermodynamic limit. For even small linked chains of agents, such as $N = 10$, one finds that A_C is a sufficient approximation of A for logistic dynamics since $\|A - R_1\|_2 - \|A_C - R_1\|_2 \leqslant 1.5 \times 10^{-3}$, and $\|A - R_1\|_2 = 0.9995$ indicates that the mean-field behavior of the linear chain is not expected to be logistic.

VI. MEAN-FIELD BEHAVIOR OF HETEROGENEOUS NETWORKS

Thus far we have discussed the accuracy of logistic mean-field solutions as approximations to actual solutions of information propagation on the network. Conversely, an important question to address is how well the logistic solution approximates mean-field behavior for arbitrary heterogeneous networks that perhaps do not have doubly stochastic or symmetric edge weights, or whose structure is not defined algebraically. How can one analyze the influence of a network's structural properties on the dynamics that occur on the network? One approach to answering this question involves comparing the parameters that describe the graph structure to the parameters that describe the dynamics on the the graph. For the parameters used to describe the SI-type dynamics, one can choose the transmission rate and the initial value of a scalar logistic approximation to the ensemble average of mean-field solutions.

Since a discrete logistic solution of Eq. (1) is determined by the transmission rate β_t and an initial condition that depends on population size, one should be able to deduce a β_t and initial value for a given time series that resembles a discrete logistic solution. Once these are known, one can infer a corresponding homogeneous network whose mean population behavior produces an almost identical logistic solution. Therefore, if the mean behavior of a heterogeneous network is known, then one can describe similar system dynamics in terms of a homogeneous network by fitting a discrete logistic solution to the mean heterogeneous solution. Here the error of the approximation is defined as the two-norm of the difference between the heterogeneous solution data points and points of the discrete logistic approximation for the first 100 time steps.

The mean behavior of a realization where only one agent is informed depends on the size of the reachable set for that initially informed agent. In general, the reachable set is the union of the reachable sets of all initially informed individuals. Thus, when comparing the mean behavior for different initially informed node sets, one should be sure that their reachable sets are of the same size. It is often useful to identify a set of strongly connected nodes since each node contained in a strongly connected set must necessarily have the same reachable set of nodes [33]. To keep comparisons simple, one can compute the mean population behavior with respect to time when only one node is initially informed, and repeat this computation for each node in the strongly connected set. We explore this idea for a graph topology defined by a naturally occurring scale-free graph, a family of Watts-Strogatz graphs, and a family of linked subgraphs where each subgraph is itself a Watts-Strogatz graph.

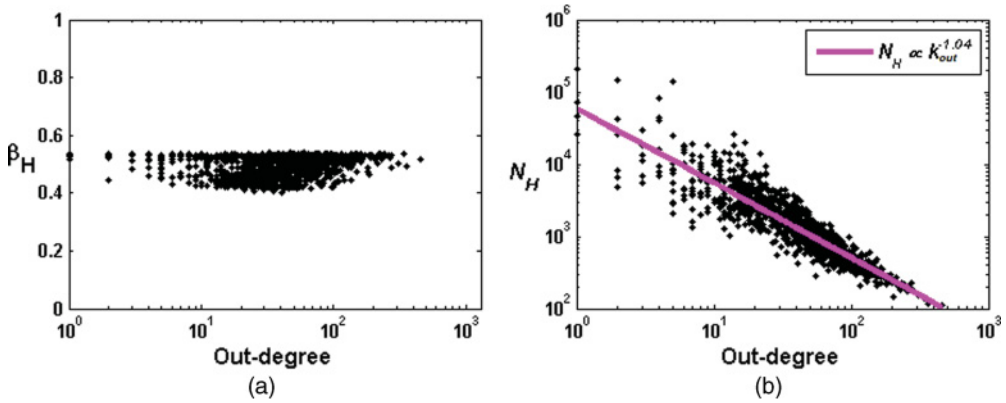


FIG. 1. (Color online) For each node contained in $\{V_{SC}\}$, data were generated according to Eq. (4), where each given node is the only one initially informed. A discrete logistic solution was then fit to each initial node's mean-field solution. (a) and (b) The optimal parameter values that minimize the two-norm difference between the original mean-field solution and the approximate logistic solution for each initial node, and plotted with respect to the initial node's out-degree (k_{out}). The logistic approximations have a mean two-norm difference of 0.0326 with a 0.0024 standard deviation, and range of [0.0277,0.0413].

A. Scale-free graph example

Here we begin with a graph topology defined by a network of Wikipedia administrator voters [43]: an example of a naturally occurring directed social network with uncorrelated degree distributions. The in-degree (k) distribution is approximately power law distributed with prob(k) $= 0.293 * k^{-1.357}$, and the sample correlation coefficient between in-degree and out-degree is $\gamma = 0.387$. Here the sample correlation coefficient γ between two finite data sets, e.g., x and y, is calculated according to Ref. [44]

$$\gamma = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{16}$$

where $\bar{x}$ and $\bar{y}$ represent the mean values of the x and y data sets, respectively.

In the context of information propagation, suppose agents i and j are neighbors. If i votes for j, then this indicates a directed relationship where we know i at least pays attention to j. In this sense, information is understood to flow from j to i and indicates the presence of a directed edge (j,i). For this study we compared the realizations for each node in the largest set $\{V_{\mathrm{SC}}\}$ defined as the set of strongly connected nodes that contains the node of greatest out-degree. The set $\{V_{\mathrm{SC}}\}$ contains 1300 nodes and a reachable set of 5158 nodes.

Denoting k_i as the number of edges directed toward agent i, each node is assumed to follow his in-neighbors equally such that each edge directed toward agent i has the value $1/k_i$, which shall be referred to as the *unbiased weighting scheme*. Using the unbiased weighting scheme allows our study to focus on network structure by controlling for edge weight. The dynamics were simulated according to Eq. (4) for the Wikipedia voting adjacency matrix, and the results are shown in Figs. 1 and 2. The realizations were generated using $\beta = 1$, to control for transmission rate, and the average probability of being informed was calculated over the entire reachable set at each time step for 100 steps with $h = 0.99$. For each node in $\{V_{\mathrm{SC}}\}$, a realization was computed where the given node has an initial probability of 1 and all other nodes have initial probabilities of 0.

Since the transmission rate used to generate each realization was a constant value, we took it as an assumption that the best fit scalar logistic approximation is generated by an unknown constant transmission rate β. Given the mean behavior of each heterogeneous realization, we first identified the best fit transmission rate (β_H) using a least-squares method because β is linear with respect to the dynamics at each time step. We then identified the optimal initial condition using an iterative process. By assuming also that only one individual is initially informed, one can deduce an effective homogeneous population size (N_H) from the initial value of the logistic fit since the initial value is the inverse of the homogeneous population size in this case.

The closest approximate logistic dynamics are depicted in Fig. 1 in terms of the β_H and N_H parameter values for the set $\{V_{\mathrm{SC}}\}$. The mean-field solutions have β_H values that appear to be rather consistent regardless of the initial node's out-degree, as shown in Fig. 1(a), while Fig. 1(b) suggests that the values of N_H depend logarithmically on initial node out-degree. Figure 2 shows how well a discrete logistic solution describes the ensemble average of population mean-field solutions for the set $\{V_{\mathrm{SC}}\}$. The best fit logistic approximations

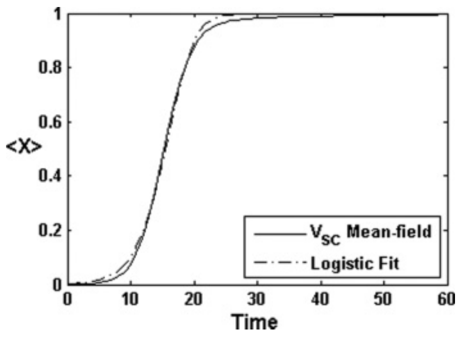


FIG. 2. Logistic approximation of the ensemble over all realizations for the set $\{V_{\mathrm{SC}}\}$. The approximation has $\beta_H = 0.4801$, $N_H = 340$, and a two-norm error of 0.0624.

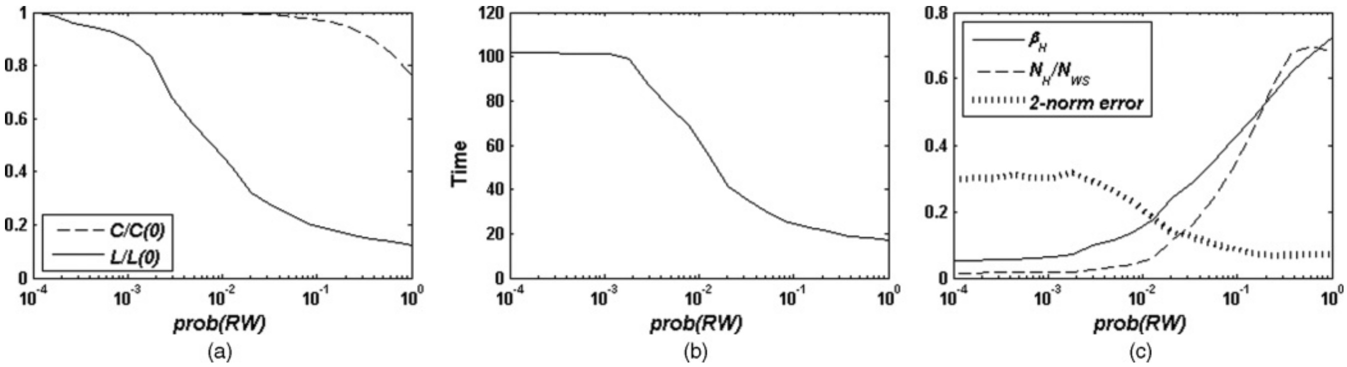


FIG. 3. Similar to the original study conducted by Watts and Strogatz [21], 20 WS graphs of size $N_{WS} = 1000$ were generated, and their graph parameters were averaged at each rewiring probability. (a) WS graph structure in terms of the average clustering coefficient (C) and average characteristic path lengths (L) over all nodes, as defined in Appendix C. Both C and L are normalized with respect to their values for zero rewiring probability. (b) Average time steps for mean-field solutions of (4) to reach $x = 0.99$. (c) Average transmission rate (β_H), homogeneous population size (N_H), and two-norm homogeneous approximation error. N_H is normalized with respect to the population size of the original WS network.

have two-norm errors that lie outside the range of two-norm errors for their individual mean-field solutions (i.e., $0.0624 \notin [0.0277, 0.0413]$). Hence, each individual mean-field solution is closer to having discrete logistic behavior than the expected average behavior of the whole system.

By approximating mean-field solutions on an arbitrary network by that of a homogeneous network, one can interpret the homogeneous network size N_H as being a descriptor of the ease with which information can spread through the network. This argument is particularly convincing in cases when the values of β_H are essentially the same over all mean-field solutions, because then the bounding of solutions depends only on the initial values, and thus N_H, when the step-size condition is met, as discussed in Appendix A. When only one node is initially informed, large homogeneous networks will naturally take longer for information to diffuse through than relatively smaller homogeneous networks. Therefore, one can compare a given node's effect on the the network to that of other nodes of the network. For the unbiased weighting scheme, these results do not contradict the findings of Refs. [3,13] since faster rates of information diffusion in our model are correlated with higher out-degree of the initially informed node. The correlation value between an initial node's $\log_{10}(k_{\text{out}})$ and its $\log_{10}(N_H)$ is $\gamma = -0.9001$.

B. Watts-Strogatz graphs

One can control for structural effects on the dynamics caused by the average degree of nodes on a network by analyzing a family of networks originally studied by Watts and Strogatz [21], where a regular undirected graph is constructed such that each node has the same degree and a subset of the edges are "rewired" according to a given rewiring probability. We generated, for each rewiring probability [prob(RW)], a set of 20 Watts-Strogatz (WS) graphs of 1000 nodes and average degree of 20. Each graph was given an unbiased edge weighting scheme, and the mean-field solutions were generated with a transmission rate of $\beta = 1$. The logistic solutions were approximated following the procedure described in Sec. VI A for which each node is initially informed with probability 1 and all else zero. For a set of rewiring probabilities ranging from prob$(RW) = 0$ to prob$(RW) = 1$, the average graph structural parameters are shown in Fig. 3(a), while the average homogeneous approximation parameters are shown in Figs. 3(b) and 3(c). It is noted that the trends of L and C in Fig. 3(a) suggest the presence of small-world structures logarithmically centered about prob$(RW) = 0.1$, where the ratio of C to L is greatest [21].

Figure 3(c) shows that the accuracy of the homogeneous approximation improves as the graph becomes more random. It is also noted that both β_H and N_H increase as prob(RW) increases. The increase in N_H caused by an increase in prob(RW) seems counterintuitive because one would ordinarily expect a network of shorter average path length to seem smaller from the perspective of the information diffusing on the network. However, the value of β_H also increases along with prob(RW), which likely counteracts this effect. It is also noted that the result of Eq. (13) for regular graphs indicates an order of accuracy that is proportional to $\cos[\pi(20)/(1000)]$ for this case. The average homogeneous approximation errors of the scale-free graph of Sec. VI A is an order of magnitude smaller than those of the family of WS graphs, even though the population size of the scale-free graph is almost an order of magnitude larger. It is observed that the average error decreases with the value of $k/N = 0.02$ held constant during these simulations, which indicates that the mean-field behavior of random graphs is in this sense relatively more logistic than that of regular graphs. Here we shall adopt the $\langle\cdot\rangle$ notation to denote the average value of a given parameter over all nodes at a fixed rewiring probability.

To determine which types of networks spread information the fastest, one can compare the average time it takes the system to reach 99% information saturation (i.e.,$\langle x\rangle = 0.99$) since in some cases it is possible to reach only 100% saturation in infinite time. The number of time steps needed for the system to reach 99% information saturation is depicted in Fig. 3(b). As Table I suggests, the characteristic path length is a strong indicator of the rate at which information is able to diffuse through a WS network, while β_H has more of an effect on time needed to reach saturation than N_H. It is noted that the

TABLE I. Correlation coefficients calculated according to Eq. (16) over the spectrum of rewiring probabilities, relating the average parameters in the left column to the average number of steps needed for mean-field solutions to reach $x = 0.99$.

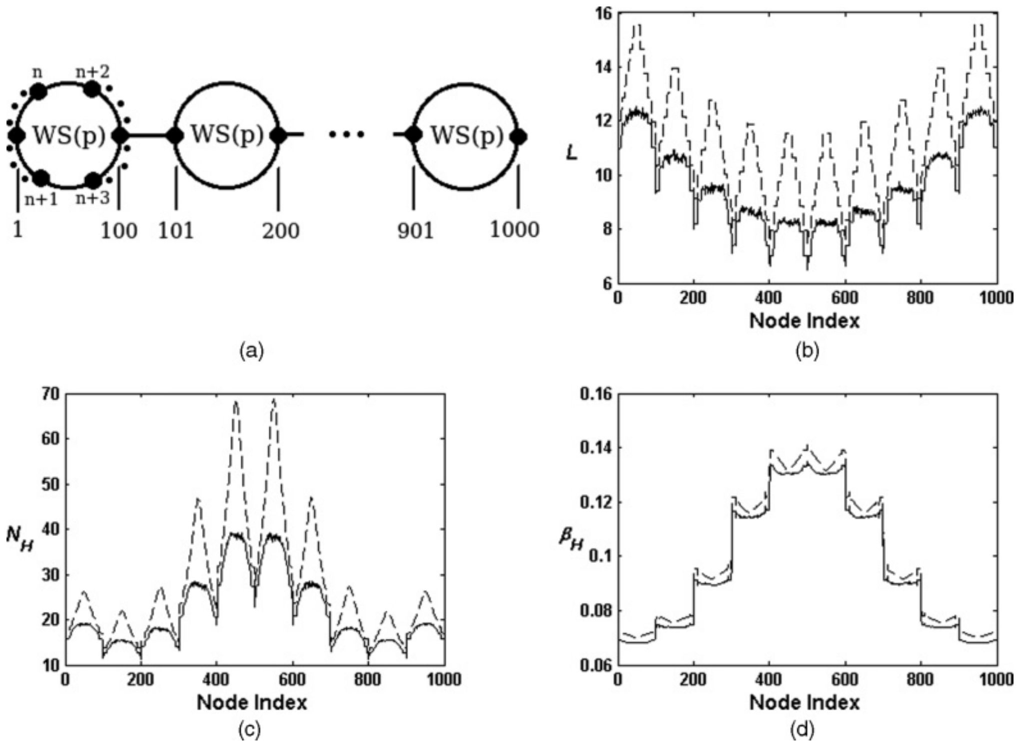
Parameter	Correlation
β_H	−0.9178
N_H	−0.8478
C	0.6304
L	0.9876

least amount of time needed to reach 99% saturation occurs for the set of graphs having rewiring probability prob(RW) = 1 (random graphs), and occurs in eight fewer time steps on average than graphs of prob(RW) = 0.1 (small-world graphs). This suggests that random networks spread information faster than small-world graphs when controlling for average node degree.

C. Chain of Watts-Strogatz graphs

To extend the analysis of WS graphs, suppose a network is constructed as a sequence of WS networks such that only one undirected edge connects two neighboring WS subnetworks, as depicted in Fig. 4(a). Applying the homogeneous approximation to this type of system allows comparison to both WS graphs and chain graphs on the macroscale, while also being able to probe the behavior of individual nodes, such as those that connect the distinct WS subgraphs, on the local scale.

For the chain of WS graphs, the average graph structural parameters are shown in Fig. 5(a), while the average homogeneous approximation parameters are shown in Fig. 5(b). Although the trend of C in Fig. 5(a) for the chain of WS graphs is almost identical to that of Fig. 3(a) for a single WS graph, there is a noticeable difference in the trends of L among Figs. 3(a) and 5(a) with respect to prob(RW). One might hypothesize that the discrepancy of L between the two systems can be attributed to the fact that each WS subgraph of the WS graph chain has 100 members instead of 1000. However, the identical behavior of C for the two systems suggests that the chain structure of the WS subgraphs has a more significant impact on L with respect to prob(RW) since rewirings were not allowed to occur between each WS subgraph. When the average homogeneous approximation parameters are compared, one finds that the data of Fig. 5(b) show trends that oppose those of Fig. 3(c): Fig. 5(b) shows an increasing error and barely decreasing β_H and N_H as prob(RW) increases. The reason for the opposing trend in the average homogeneous data for the chain of WS graphs versus the single large WS graph is that the WS subgraphs are linked as a sequential chain. As rewiring probability approaches prob(RW) = 1, each WS subgraph becomes better mixed and appears to behave more as one entity since information diffuses fastest on the single WS network level, as shown in Fig. 3(b). While the WS subgraphs become more mixed, the increase in homogeneous approximation error is likely explained by the fact that chains of nodes



(a) (b)

(c) (d)

FIG. 4. (a) Chain of 10 WS graphs with 100 nodes each are linked together with one edge connecting each WS graph. The nodes are labeled left to right, and alternating top to bottom, with increasing index. (b) Average L with respect to individual node index. (c) Average N_H with respect to individual node index. (d) Average β_H with respect to individual node index. In (b), (c), and (d) the dashed line represents data for prob(RW) = 0, and the solid line represents data for prob(RW) = 1.

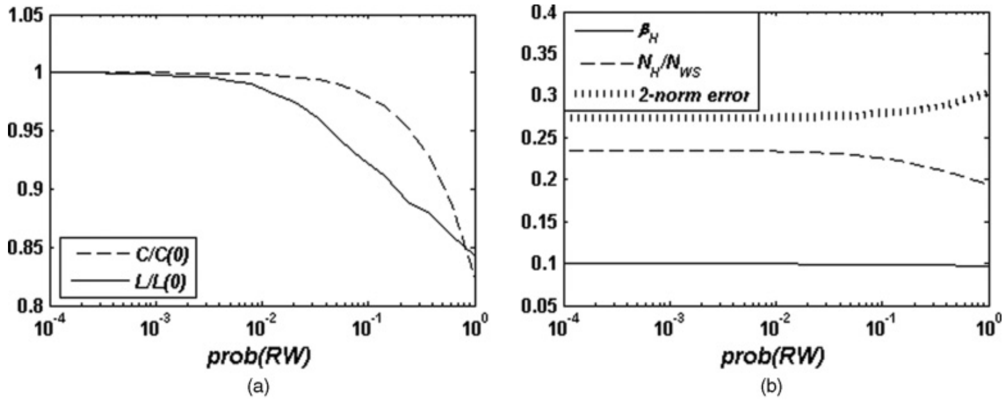


FIG. 5. Similar to the original study conducted by Watts and Strogatz [21], 20 chains of 10 WS graphs of size $N_{\text{WS}} = 100$ were generated, and their graph parameters were averaged at each rewiring probability [prob(RW)]. (a) WS graph structure in terms of the average clustering coefficient (C) and average characteristic path lengths (L) over all nodes. Both C and L are normalized with respect to their values for prob(RW) = 0. (b) Average transmission rate (β_H), homogeneous population size (N_H), and average two-norm error. N_H is normalized with respect to the population size of the constituent WS networks.

do not produce logistic mean-field solutions, as discussed in Sec. V.

Figure 6 shows that the average clustering coefficient is weakly correlated with both β_H and N_H, while the characteristic path length is somewhat correlated with β_H over all rewiring probabilities and becomes more correlated with N_H as rewiring probability increases. As the characteristic path length decreases, the negative correlation with both β_H and N_H indicates that the information not only diffuses faster, but through an effectively larger network. Hence, the structural effects that cause an increase in β_H values offset those that cause a decrease in N_H, and results in an average of 76 time steps to reach 99% of saturation for each rewiring probability. In this case one can interpret the networks as being equally capable of diffusing information.

The expected mean-field solutions are actually quite similar in performance to each other for this type of system since the collection of ensemble averages of the mean-field solutions over the spectrum of rewiring probabilities have an average two-norm error of 0.0624. Compared to the error of the homogeneous approximation to mean-field solutions, which has an average value of 0.2778 in two-norm, as deduced from the data of Fig. 5(b), the homogeneous approximation is still sensitive enough to detect subtle features in the mean-field solutions despite how nonlogistic the mean-field solutions are.

At the individual node level, Fig. 4(b) shows the characteristic path lengths at each end of the rewiring probability spectrum, along with their corresponding index labels. At this level of detail, it is easy to see the effects that the subgraph connecting nodes have on the dynamics relative to their global location on the graph. By comparing Fig. 4(b) to 4(c) and 4(d) one is able to observe how the characteristic path length of each node is reflected by its effective homogeneous network size and effective transmission rate, respectively. Figure 4(c) shows how the subgraph connecting nodes perceive the smallest effective homogeneous networks, while those toward the center of the network perceive the largest effective homogeneous networks of all. When this observation is compared to the average characteristic path length of each node, as observed in Fig. 4(b), one finds this observation to, again, be a counterintuitive result that can be explained by the opposing effect of β_H as seen in Fig. 4(d).

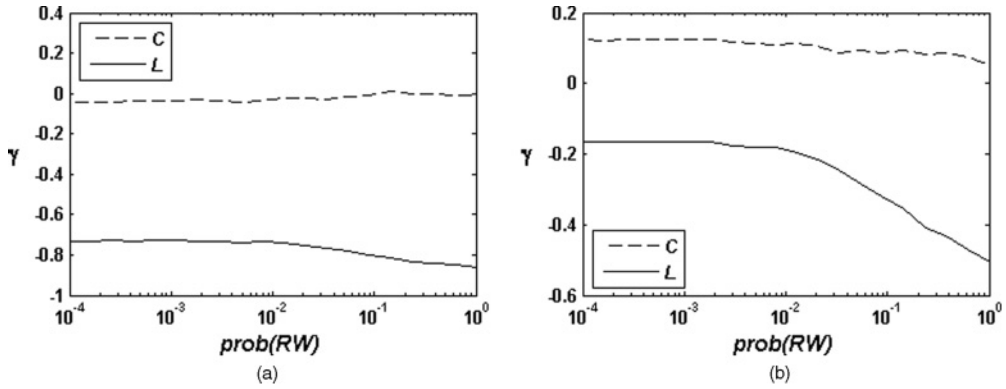


FIG. 6. (a) Average correlation values between C and L, and β_H for the chain of WS graphs. (b) Average correlation values between C and L, and N_H for the chain of WS graphs.

VII. CONCLUSIONS

By focusing our attention on SI dynamics, we have shown the importance of applying algebraic graph theory to dynamic processes in a simple information-spreading context, and the new physical insight it is able to provide to information-spreading phenomena. In contrast to the application of approximate parameter distributions to the dynamic equations, such as power-law degree distributions, adjacency matrices preserve the exact global structure of a weighted network. Here we were able to also use adjacency matrices to rigorously attribute the size of fluctuations about the population-level solutions to the structural similarity between a given graph and a completely connected graph. In the case of completely connected graphs, the fluctuations were found to be attributed to finite size effects.

Specifically, we have constructively shown that the agent-based and scalar logistic models are in exact agreement for the completely connected case in the limit as the number of agents in the system approaches infinity, as conjectured previously in different research communities. For homogeneous systems consisting of a finite number of agents, the singular values of the graph adjacency matrix produce the closest logistic approximation to the completely connected agent-based dynamics. This result was extended to connected networks that are doubly stochastic or with symmetric interactions so that systems of this type can generally be approximated by the discrete logistic equation. Although double stochasticity is typically relevant only to engineered systems, we have seen that our analytic methods are applicable to naturally occurring systems since propagation mechanisms with interaction symmetry are quite common. We also discussed how one can analyze the logistic behavior of arbitrary heterogeneous network topologies.

Moreover, by analyzing average population behavior, we found that there are instances when solutions to heterogeneous dynamics of one set of parameter values appear to be well-approximated by homogeneous dynamics for a different set of parameter values. If the entire network structure and set of parameter values are known, an implication of being able to use homogeneous systems to approximate heterogeneous systems is that it provides a standard way of comparing the dynamics of two heterogeneous systems. In general, the logistic behavior of any two homogeneous networks can be compared to each other. To avoid results that may be misleading, however, we advocate comparing only networks of the same size when making network versus network comparisons and comparing nodes with the same size reachable sets when making node versus node comparisons.

In regard to the inverse problem of using observable system behavior to infer graph features, coarse descriptions, such as mean-field behavior, are likely not to contain enough detail to distinguish one graph topology from another. Caution should be exercised when observing similar mean-field behavior of different logistic dynamical systems because uniqueness properties relating mean-field behavior to heterogeneous graph structure have yet to be established, particularly if they have different weighting schemes. Just as there can be two completely different graph structures that produce the exact same mean-field behavior, there could also be two identical graph structures of different weighting schemes that could exhibit different mean-field behaviors.

ACKNOWLEDGMENTS

We would like to thank the contributors of the MatlabBGL toolbox [45], the contributors of the CONtrolable TEST matrices toolbox [46], as well as Tapiwa Maruni for contributing some of the Matlab functions that were used during this research. We also express gratitude to the referees for their insightful and constructive feedback. This work was supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the US Army Research Office. The content of the information does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

APPENDIX A. LOGISTIC BOUNDS OF THE COMPLETELY CONNECTED SOLUTION

For systems of finite size, it is possible to bound (6) by discrete logistic functions. Since the fact that each element of $p_t \in [0,1]$ implies that $(p_t^{(i)})^2 \leqslant p_t^{(i)}$ and thus $|p_t|_2^2 \leqslant |p_t|_1$, then one obtains the upper bound:

$$x_{t+h} \leqslant x_t + h\beta_t \frac{N}{N-1} x_t(1-x_t). \tag{A1}$$

A lower bound for Eq. (6) can be obtained by simply truncating the the $|p_t|_2^2$ term:

$$x_{t+h} \geqslant x_t + h\beta_t x_t \left(1 - \frac{N}{N-1} x_t\right). \tag{A2}$$

We now compare the various logistic approximations to conclude that the single-step upper and lower logistic bounds of Eq. (6) produce upper and lower solutions for all time steps. First, we show that if given $x_t, y_t \in [0,1]$ at time t and parameters $\phi,\theta \in R_{>0}$, then $x_t \geqslant y_t$ implies $x_{t+h} \geqslant y_{t+h}$ if the two points evolve according to the same discrete logistic equation of the form $x_{t+h} = x_t + \phi h \beta_t x_t(1-\theta x_t)$.

By assuming $x_t \geqslant y_t$, then

$$\begin{aligned} x_{t+h} - y_{t+h} &= x_t + \phi h\beta_t x_t(1-\theta x_t) - y_t - \phi h \beta_t y_t(1-\theta y_t) \\ &= (x_t - y_t) + \phi h \beta_t \big[(x_t - y_t) - \theta\big(x_t^2 - y_t^2\big)\big] \\ &= (x_t - y_t)\{1 + \phi h \beta_t [1 - \theta(x_t + y_t)]\} \\ &\geqslant (x_t - y_t)[1 + \phi h \beta_t (1 - 2\theta)] \end{aligned}$$

and is non-negative when $2\theta \leqslant 1$ or

$$h \leqslant \frac{1}{\sup_t \beta_t \phi(2\theta - 1)}. \tag{A3}$$

By imposing a requirement on the step size of the discrete logistic equation, it can be shown that one discrete logistic solution bounds another discrete logistic solution if their single step dynamics also bound each other. When applied directly to Eqs. (7), (A1), (A2), the minimum step size comes from Eq. (A2) with $\phi = 1$ and $\theta = N/(N-1)$.

Now suppose h satisfies (A3) with the ϕ and θ values from Eq. (A2) so that

$$h \leqslant \frac{N-1}{N+1} \frac{1}{\sup_t \beta_t}.$$

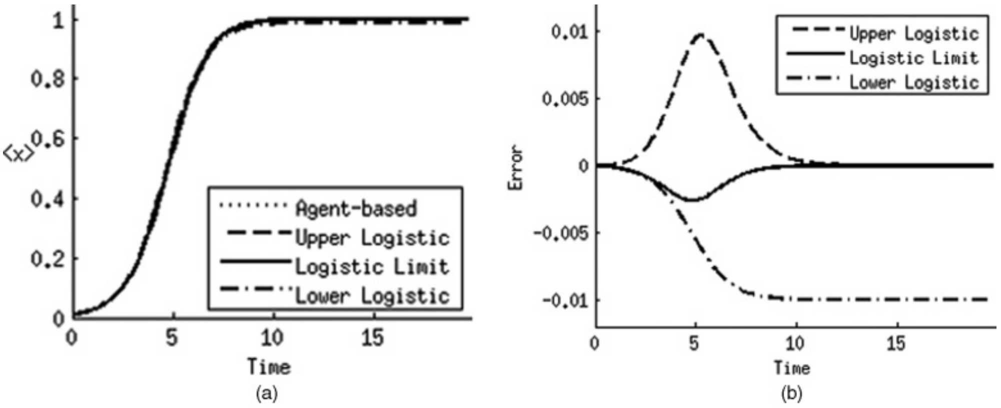


FIG. 7. (a) Comparison of the graph-based solution to the logistic equation in the thermodynamic limit, as well as the upper and lower bounding logistic solutions for the finite case. The solutions are nearly indistinguishable. (b) Pointwise error difference of the upper, lower, and thermodynamic limit logistic solutions with respect to the graph-based solution. Parameters are $\beta_t = 1$ and $N = 100$ in both plots.

The dynamic equation (7) yields a logistic approximation, x_t^*, that is bounded by Eqs. (A1) and (A2). To show this is true, let

$$x_{t+h}^{\text{Upper}} := x_t + h\beta_t \frac{N}{N-1} x_t(1-x_t),$$

$$x_{t+h}^{*} := x_t + h\beta_t x_t(1-x_t),$$

$$x_{t+h}^{\text{Lower}} := x_t + h\beta_t x_t \left(1 - \frac{N}{N-1} x_t\right).$$

Given an initial condition $|P_0|_1/N = x_0^{\text{Upper}} = x_0^* = x_0^{\text{Lower}}$, Eqs. (7), (A1), (A2) provide the respective x_1 values and their relation to each other (i.e., $x_1^{\text{Lower}} \leqslant x_1^* \leqslant x_1^{\text{Upper}}$). Let the relationship between x_0 and the x_1 terms be the base case for induction.

To show that $x_t^{\text{Upper}} \geqslant x_t^*$ implies $x_{t+h}^{\text{Upper}} \geqslant x_{t+h}^*$ as the inductive step, we begin with the hypothesis that $x_t^{\text{Lower}} \leqslant x_t^* \leqslant x_t^{\text{Upper}}$. If one were to briefly let $y_t^{\text{Lower}} = x_t^*$ and $y_t^{\text{Upper}} = x_t^*$, then the same procedure used to obtain the base cases asserts that $y_{t+h}^{\text{Lower}} \leqslant x_{t+h}^* \leqslant y_{t+h}^{\text{Upper}}$. Since $x_t^{\text{Lower}} \leqslant y_t^{\text{Lower}}$ and $y_t^{\text{Upper}} \leqslant x_t^{\text{Upper}}$, it follows that $x_{t+h}^{\text{Lower}} \leqslant y_{t+h}^{\text{Lower}}$ and $y_{t+h}^{\text{Upper}} \leqslant x_{t+h}^{\text{Upper}}$. Hence, $x_t^{\text{Lower}} \leqslant x_t^* \leqslant x_t^{\text{Upper}}$ implies $x_{t+h}^{\text{Lower}} \leqslant x_{t+h}^* \leqslant x_{t+h}^{\text{Upper}}$ so that the solution to Eq. (7) is bounded by the solutions to Eqs. (A1) and (A2). For example, Fig. 7 depicts the accuracy of solutions to Eqs. (7), (A1), and (A2) with respect to the solution generated by Eq. (6), for a completely connected network of 100 agents with one initially informed agent.

For this bounding statement to be true, the step size must be chosen appropriately based on the size on the network and the transmission rate. Thus, when comparing data to the model under the homogeneous assumption, one must consider the behavior of the information since it affects the transmission rate. The transmission rate affects the step size, which in turn affects the adjacency matrix (in the more general case).

APPENDIX B. PROOF OF GENERAL APPROXIMATION

1. Doubly stochastic matrices

To show how well the scalar logistic model approximates the graph-based model, first let A be the doubly stochastic and irreducible adjacency matrix that corresponds with the network topology. Since the elements of p_t are all non-negative, the one norm of the vector p_t is simply a sum over all of its elements (i.e., $|p_t|_1 = 1_n^T p_t$), we begin by taking the one-norm of Eq. (4):

$$\begin{aligned} |p_{t+h}|_1 &= 1^T(p_t + h\beta_t A p_t - h\beta_t \text{diag}\{p_t\} A p_t) \\ &= 1_n^T p_t + h\beta_t 1^T A p_t - h\beta_t p_t^T A p_t. \end{aligned}$$

Quadratic forms have the property that $p_t^T A p_t = p_t^T (A_S) p_t$, where $A_S = (A + A^T)/2$ is a symmetric matrix. Using the series representation $A_S = \sum_{i=1}^n \lambda_i w_i w_i^T$:

$$\begin{aligned} |p_{t+h}|_1 &= 1_n^T p_t + h\beta_t 1_n^T A p_t - h\beta_t p_t^T \sum_{i=1}^n \lambda_i w_i w_i^T p_t \\ &= |p_t|_1 + h\beta_t 1_n^T A p_t - h\beta_t \lambda_1 p_t^T w_1 w_1^T p_t \\ &\quad - h\beta_t p_t^T \sum_{i=2}^n \lambda_i w_i w_i^T p_t. \end{aligned} \tag{B1}$$

Since A is doubly stochastic, the columns of A each sum to 1 so that the second term of Eq. (B1) simplifies to $h\beta_t |p_t|_1$.

The matrix A_S will also be doubly stochastic, and thus row stochastic. From the Perron-Frobenius theorem [33], $\lambda_1 = 1$ and $w_1 = 1_n/\sqrt{n}$, and denoting the inner product as $\langle \cdot, \cdot \rangle$, the third term of Eq. (B1) can be simplified as follows:

$$\begin{aligned} -h\beta_t \lambda_1 p_t^T w_1 w_1^T p_t &= -h\beta_t \langle 1_n/\sqrt{n}, p_t \rangle^2 \\ &= -h\beta_t \frac{1}{n} \langle 1, p_t \rangle^2 = -h\beta_t \frac{1}{n} |p_t|_1^2. \end{aligned}$$

By applying the inner product notation to the fourth term of Eq. (B1), it can be rewritten as $-h\beta_t \sum_{i=2}^n \lambda_i \langle w_i, p_t \rangle^2$.

Upper and lower bounds on the fourth term of Eq. (B1) can be obtained by observing that

$$-|\lambda_2| \sum_{i=2}^n \langle w_i, p_t \rangle^2 \leqslant \sum_{i=2}^n \lambda_i \langle w_i, p_t \rangle^2 \leqslant |\lambda_2| \sum_{i=2}^n \langle w_i, p_t \rangle^2.$$

To further simplify this expression, we can use the submultiplicative property of matrix norms, where $\sum_{i=2}^n \langle w_i, p_t \rangle^2 \leqslant \sum_{i=1}^n \langle w_i, p_t \rangle^2 = \|W^T p_t\|_2^2 \leqslant \|W^T\|_2^2 |p_t|_2^2 = |p_t|_2^2$. By substituting $\sigma_2 = |\lambda_2|$ and applying this inequality, one obtains the following:

$$-\sigma_2 |p_t|_2^2 \leqslant \sum_{i=2}^{n} \lambda_i \langle w_i, p_t \rangle^2 \leqslant \sigma_2 |p_t|_2^2,$$

which indicates that the fourth term of Eq. (B1) is a term of order $h\sigma_2$.

Therefore, (B1) simplifies to

$$|p_{t+h}|_1 = |p_t|_1 + h\beta_t |p_t|_1 - \frac{h\beta_t}{n} |p_t|_1^2 + O(h\sigma_2). \quad \text{(B2)}$$

Finally, divide by n, and let $x_t = |p_t|_1/n$ to obtain the average probability of being informed:

$$x_{t+h} = x_t + h\beta_t x_t(1 - x_t) + O(h\sigma_2). \quad \text{(B3)}$$

2. Symmetric matrices

Beginning with expression (14), one can factor out R_1 and use the fact that $(A - R_1)$ is a symmetric matrix to obtain

$$p_{t+h} = p_t + h\beta_t (I - \text{diag}\{p_t\}) p_t + h\beta_t (I - \text{diag}\{p_t\}) W D W^T p_t,$$

where D is a matrix whose diagonal elements are the eigenvalues of $(A - R_1)$. By recognizing that $-\sigma_1 I \leqslant W D W^T \leqslant \sigma_1 I$, it follows that one obtains (15) by summing the elements and dividing by the cardinality of the population. A similar argument shows that mean-field solutions are upper-bounded by solutions to

$$x_{t+h} = x_t + \sigma_1 h\beta_t x_t(1 - x_t),$$

where $x_t = |p_t|_1/N$ and σ_1 provides a time-scaling effect on the step size when h satisfies (A3).

APPENDIX C. ADDITIONAL DEFINITIONS

As explained in Ref. [21], the characteristic path length (L) of a network is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. Similarly, one can define the characteristic path length (L_i) of a single node i as the average shortest path length between i and each other $j \in V$.

To define a network's average clustering coefficient (C), we first define the set of edges E_i that exist between a given node $i \in V$ and its neighbors as $E_i = \{(i,j) \in E, \forall j \in V\}$. If node i has k_i neighbors, then the clustering coefficient (C_i) of node i is

$$C_i = \frac{2|E_i|}{k_i(k_i - 1)}, \quad \text{(C1)}$$

where $|E_i|$ represents the cardinality of E_i, and $k_i(k_i - 1)/2$ is the maximum number of edges that can possibly exist in E_i. Hence, C is the average value of C_i over all i.

[1] D. Daley and D. Kendall, J. Inst. Math. Appl. **1**, 42 (1965).

[2] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili, Physica A **374**, 457 (2007).

[3] Y. Moreno, M. Nekovee, and A. F. Pacheco, Phys. Rev. E **69**, 066130 (2004).

[4] J. Zhou, Z. Liu, and B. Li, Phys. Lett. A **368**, 458 (2007).

[5] L. Bettencourt, A. Cintron-Arias, D. Kaiser, and C. Castillo-Chavez, Physica A **364**, 513 (2006).

[6] Y. Bulygin, in *Performance, Computing, and Communications Conference 2007, IPCCC 2007, IEEE International* (IEEE, New Orleans, LA, 2007), pp. 475–478.

[7] M. E. J. Newman, S. Forrest, and J. Balthrop, Phys. Rev. E **66**, 035101 (2002).

[8] L. Allen, Math. Biosci. **124**, 83 (1994).

[9] L. Allen and A. Burgin, Math. Biosci. **163**, 1 (2000).

[10] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, 1991).

[11] H. W. Hethcote and J. W. Van Ark, Math. Biosci. **84**, 85 (1987).

[12] H. W. Hethcote, SIAM Rev. **42**, 599 (2000).

[13] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, J. Theor. Biol. **235**, 275 (2005).

[14] C. Castellano and R. Pastor-Satorras, Phys. Rev. Lett. **105**, 218701 (2010).

[15] R. Durrett, Proc. Natl. Acad. Sci. USA **107**, 4491 (2010).

[16] J. Lloyd-Smith, S. Schreiber, and W. Getz, Contemp. Math. **410**, 235 (2006).

[17] R. M. May and A. Lloyd, Phys. Rev. E **64**, 066112 (2001).

[18] Y. Moreno, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **26**, 521 (2002).

[19] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[20] J. Saramäki and K. Kaski, J. Theor. Biol. **234**, 413 (2005).

[21] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[22] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[23] M. Morris, in D. Mollison, editor, *Epidemic Models: Their Structure and Relation to Data* (Cambridge University Press, Cambridge, 1995), pp. 302–322.

[24] A. Lloyd, S. Valeika, and A. Cintron-Arias, Contemp. Math. **410**, 209 (2006).

[25] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Physica A **346**, 34 (2005).

[26] C. Castellano and R. Pastor-Satorras, Phys. Rev. Lett. **96**, 038701 (2006).

[27] D. Centola and M. Macy, Am. J. Sociol. **113**, 702 (2007).

[28] S. Gomez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, Europhys. Lett. **89**, 38009 (2010).

[29] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[30] C. D. Godsil and G. Royle, *Algebraic Graph Theory*, *Graduate Texts in Mathematics* (Springer, New York, 2001).

[31] P. Sharma, U. Khurana, B. Shneiderman, M. Scharrenbroich, and J. Locke, in *Social Computing, Behavioral-Cultural Modeling and Prediction*, *Lecture Notes in Computer Science* Vol. 6589, edited by J. Salerno, S. Yang, D. Nau, and S.-K. Chai (Springer, Berlin, 2011), pp. 244–251.

[32] T. G. Hallam and C. E. Clark, J. Theor. Biol. **93**, 303 (1981).

[33] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks* (Princeton University Press, Princeton, 2009).

[34] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész, New J. Phys. **9**, 179 (2007).

[35] J. Leskovec, L. Backstrom, and J. Kleinberg, in *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2009), pp. 497–506.

[36] L. Allen, *An Introduction to Stochastic Processes with Applications to Biology* (Pearson Education, Upper Saddle River, NJ, 2003).

[37] B. Gharesifard and J. Cortés, in *American Control Conference (ACC), 2010* (IEEE, Baltimore, MD, 2010), p. 30.

[38] R. Ren and R. W. Beard, *Distributed Consensus in Multi-vehicle Cooperative Control Theory and Applications* (Springer, New York, 2008).

[39] R. Olfati-Saber, J. A. Fax, and R. M. Murray, Proc. IEEE **95**, 215 (2007).

[40] H. Dym, *Linear Algebra in Action* (American Mathematical Society, Providence, RI, 2007).

[41] S. J. Leon, *Linear Algebra with Applications* (Prentice Hall, Upper Saddle River, NJ, 1998).

[42] R. M. Gray, *Toeplitz and Circulant Matrices: A Review* (Now Publishers, Hanover, MA, 2006).

[43] J. Leskovec, D. Huttenlocher, and J. Kleinberg, http://snap.stanford.edu/data/wiki-Vote.html.

[44] A. L. Edwards, *An Introduction to Linear Regression and Correlation* (W. H. Freeman & Co., New York, 1984).

[45] D. Gleich, http://www.mathworks.com/matlabcentral/fileexchange/10922.

[46] A. Taylor and D. Higham, http://www.mathstat.strath.ac.uk/research/groups/numerical_analysis/contest.