Tracking Infection Diffusion in Social Networks: Filtering Algorithms and Threshold Bounds

Vikram Krishnamurthy, *Fellow, IEEE*, Sujay Bhatt, and Tavis Pedersen

***Abstract*—This paper deals with the statistical signal processing over graphs for tracking infection diffusion in social networks. Infection (or Information) diffusion is modeled using the susceptible-infected-susceptible (SIS) model. Mean field approximation is employed to approximate the discrete valued infection dynamics by a deterministic difference equation, thereby yielding a generative model for the infection diffusion. The infection is shown to follow polynomial dynamics and is estimated using an exact nonlinear Bayesian filter. We compute posterior Cramér-Rao bounds to obtain the fundamental limits of the filter that depend on the structure of the network. The SIS model is extended to include homophily, and filtering on these networks is illustrated. Considering the randomly evolving nature of real world networks, a filtering algorithm for estimating the underlying degree distribution is also investigated using generative models for the time evolution of the network. We validate the efficacy of the proposed models and algorithms with synthetic data and Twitter datasets. We find that the SIS model is a satisfactory fit for the information diffusion, and the nonlinear filter effectively tracks the information diffusion.**

***Index Terms*—Cramér-Rao bounds, diffusion threshold, homophily, mean-field dynamics, non-linear Bayesian filter, social Networks, stochastic dominance, twitter dataset.**

I. Introduction

STATISTICAL signal processing on graphs is an emerging field in which the structural properties of the graph are utilized to derive statistical inference algorithms. As described in [1], there is a wide range of social phenomena such as diffusion of technological innovations, cultural fads, and economic conventions [2] where individual decisions are influenced by the decisions of others. In this paper, we consider social networks represented as graphs and we are interested in analyzing the manner in which information (or infection) spreads through the network. A large body of research on social networks has been devoted to the diffusion of information (e.g., ideas, behaviors, trends) [3], and particularly on finding a set of target nodes so as to maximize the spread of a given product [4].

Manuscript received September 23, 2016; revised February 2, 2017 and April 3, 2017; accepted April 18, 2017. Date of publication April 25, 2017; date of current version May 17, 2017. This work was supported by an Army Research Office grant. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Edwin K. P. Chong. (*Corresponding author: Vikram Krishnamurthy.*)

V. Krishnamurthy and S. Bhatt are with the School of Electrical and Computer Engineering and Cornell Tech, Cornell University, Ithaca, NY 14850 USA (e-mail: vikramk@cornell.edu; sh2376@cornell.edu).

T. Pederson is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: tpedersen@ece.ubc.ca).

Digital Object Identifier 10.1109/TSIPN.2017.2697940

A. Organization and Main Results

Section II presents the *Susceptible Infected Susceptible* (SIS) model for infection diffusion in the network. The key result is that the mean field dynamics approximates the discrete-valued infection dynamics by a deterministic ordinary differential equation. The mean field dynamics yield a tractable model for Bayesian filtering in order to estimate the infection dynamics given a sampling procedure for the social network. From a signal processing point of view, the mean field dynamics (MFD) has an interesting interpretation: it resembles a stochastic approximation algorithm; however, in our case, it constitutes a generative model[1] instead of an algorithm.

The mean field dynamics of Section II yields a dynamical system whose state (infected population state) evolves with polynomial dynamics. Section III uses a recent result in Bayesian filtering to obtain an exact (finite dimensional) filter for the infected population state given noisy observations. We examine via numerical examples and posterior Cramér-Rao lower bounds, how state estimation over large networks is affected by the network; see [5] for posterior Cramér-Rao bounds for non-linear filters. Numerical examples illustrate the difference in performance between power law (scale free) and Erdős-Rényi graphs. Scale free networks arise in online social networks such as Twitter [6] and in the link network of the World Wide Web [7].

The classical SIS model assumes a fixed underlying social network. Section IV analyzes the diffusion threshold of a SIS model when the social network evolves over time. Since information diffusion occurs at a faster time scale compared to forming connections in social networks, we consider a two time scale formulation: the degree distribution of the underlying network changes on a slow time scale, and the infection diffuses over a faster time scale. Our results generalize the results in [1], where the network was assumed to be fixed.

Section V illustrates the SIS model and the performance of the Bayesian filter on simulated data and examines the sensitivity of the filter to the underlying graph model (Erdős-Rényi vs Scale Free). We also present the analysis in an SIS model with homophily and illustrate the improved mean field approximation and filter performance. Finally, we present a detailed example using a real Twitter dataset. It is shown via a goodness of fit test that SIS is a reasonable model for information propagation in

[1] From Theorem 1 (see Section II-B), the maximum deviation - over the entire infection dynamics trajectory - between the deterministic approximation (MFD) and the actual infection dynamics, satisfies an exponential bound.

the Twitter dataset and that the infected population state can be tracked satisfactorily over time via the Bayesian filter.

B. Related Literature

There are several models for studying the spread of infection and technology in complex networks including Susceptible-Infected-Susceptible (SIS), Susceptible-Alert-Infected-Susceptible (SAIS), and Susceptible-Exposed-Infected-Vigilant (SEIV); see [8], [9]. Susceptible-Infected-Susceptible (SIS) models have been extensively studied in [1], [10]–[13] to model information/infection diffusion, for example, the adoption of a new technology in a consumer market.

Degree-based mean field dynamics approximations for SIS models have been derived in [1], [14]. Pair approximations (PA) and approximate master equations (AME) yield more general models for the complex dynamics of large scale networks [14]. However, the resulting differential/difference equations that characterize the dynamics in PA and AME are no longer polynomial functions of the state. In this more general case, however, a suboptimal filter such as a particle filter can be used to track the infection diffusion.

In this paper, since we focus on optimal Bayesian estimation (filtering) of the infection dynamics by sampling the network, we use the SIS model. In comparison, [15] provides a stochastic approximation algorithm and analysis on a Hilbert space for tracking the degree distribution of evolving random networks with a duplication-deletion model. In small sized sensor networks represented by graphs, [16] derived an optimal estimation algorithm, based on the Kalman filter, to estimate the state at each sensor. Further, on specific structures like trees, [16] provides expressions for the steady-state covariance.

On networks having fixed degree distribution, [1] identified conditions under which a network is susceptible to an epidemic using a mean-field approach and provided a closed form solution for the infection diffusion threshold. The diffusion properties of networks was investigated using stochastic dominance of their underlying degree distributions like in [17]. We generalize these stochastic dominance results for evolving networks by considering a simple preferential attachment model as this can generate a scale-free network [18].

Finally, [11] studies the link between the power law exponent and the diffusion threshold. For the preferential attachment model, [18] studies the connection between the parameters that dictate the evolution (node and edge addition probability) and the degree distribution. We obtain similar results in this paper using stochastic dominance, but, the key emphasis is on providing a structured way to study such ordinal sensitivity relationships in large networks.

II. SIS Model and Mean Field Dynamics

This section discusses the discrete time SIS model [19], [20] and mean field dynamics for the diffusion of information in a social network. We also formulate sampling of nodes in the network. The final outcome is a state space model with polynomial dynamics and noisy observations which is amenable to Bayesian filtering.

Consider social network represented by an undirected graph G:

$$G = (V, E), \text{ where } V = \{1, \ldots, M\}, \text{ and } E \subseteq V \times V. \quad (1)$$

Here V denotes the set of M vertices (nodes) and E denotes the set of edges (relationships). The degree of a node m is its number of neighbors[2]:

$$\Xi^{(m)} = |\{g \in V : g, m \in E\}|, \ |\cdot| \text{ denotes cardinality.}$$

Let $M(d)$ denote the number of nodes in the network G with degree d, and let $\rho(d)$ denotes the degree distribution. That is, for degree $d = 0, 1, \ldots, D$,

$$M(d) = \sum_{m \in V} \mathcal{I}\big(\Xi^{(m)} = d\big), \quad \rho(d) = \frac{M(d)}{M}.$$

Here $\mathcal{I}(\cdot)$ denotes the indicator function and D denotes the maximum degree. Since $\sum_d \rho(d) = 1$, $\rho(d)$ can be viewed as the probability that a node selected randomly on V has degree d. At each time instant $n = 0, 1, \ldots$, each node m has state:

$$s_n^{(m)} \in \{1 \text{ (infected)}, 2 \text{ (susceptible)}\}.$$

Define the *infected population state* $\bar{x}_n$ at time n as the fraction of nodes with degree d at time n that are infected:

$$\bar{x}_n(d) = \frac{1}{M(d)} \sum_m \mathcal{I}\big(\Xi^{(m)} = d, s_n^{(m)} = 1\big), \quad d = 0, \ldots, D. \quad (2)$$

A. Individual Dynamics & Discrete Time SIS Model

In the SIS model, the state of nodes evolves over time as a discrete time Markov chain. The transition probabilities depend on the degree of the node and the number of infected neighbors. The infection dynamics evolves as follows[3,4]:

Step 1: A node m is chosen uniformly from the vertices in V. Suppose the node has degree $\Xi^{(m)} = d$ and the number of infected neighbors $F_n^{(m)} = a$. The state $s_n^{(m)}$ of an individual node m at time n evolves from state i to j with transition probabilities $\bar{P}_{ij}$, where $i, j \in \{1, 2\}$. $\bar{P}_{21}$ is known as the *infection* probability and $\bar{P}_{12}$ is known as the *recovery* probability. These transition probabilities of individual nodes depend on the node's degree d and its number of infected neighbors

[2]A vertex g is adjacent to a vertex m if there is an edge between them. The neighbors of a node m are all the vertices that are adjacent to m.

[3]The state of the network (1) at time n is given by an M dimensional vector with elements 1 or 2. The state space is given by $\{1, 2\}^M$ and the dynamics can be formulated in terms of a transition matrix between all possible states [14],[21]. In this paper, we group the 2^M states into D subsets, one for each degree, resulting in a state space of dimension $2^{M(d)}$ for each degree d.

[4]Example: Suppose there are $M = 130$ nodes, with $d \in \{1, 2, 3\}$ and $[M(1), M(2), M(3)] = [50, 50, 30]$. Suppose the fraction of infected nodes of degree 1, 2 and 3, are respectively, $\bar{x}_n(1) = 5/50$, $\bar{x}_n(2) = 20/50$ and $\bar{x}_n(3) = 6/30$. Then $\bar{x}_{n+1}(1) \in \{4/50, 5/50, 6/50\}$ if a node of degree 1 is chosen in Step 1), $\bar{x}_{n+1}(2) \in \{19/50, 20/50, 21/50\}$ if a node of degree 2 is chosen in Step 1), and $\bar{x}_{n+1}(2) \in \{5/30, 6/30, 7/30\}$ if a node of degree 3 is chosen in Step 1).

a. Therefore,

$$\bar{P}_{\mathrm{ij}}(d,a) = \mathbb{P}\left(s_{n+1}^{(m)} = j | s_n^{(m)} = i, \Xi^{(m)} = d, F_n^{(m)} = a\right). \quad (3)$$

One possible parametrization we will consider is $\bar{P}_{\mathrm{ij}}(d,a) = \mu_{\mathrm{ij}}\mathbb{F}_{\mathrm{ij}}(d,a)$, where μ_{ij} is interpreted as the spreading rate and $\mathbb{F}_{\mathrm{ij}}(d,a)$ is the neighborhood dependent diffusion function in [1]. This parametrization is used in Section IV to define the diffusion threshold.

Step 2: The population state of degree d is updated as[5]

$$\begin{bmatrix} \bar{x}_{n+1}(d) \\ 1-\bar{x}_{n+1}(d) \end{bmatrix} = \begin{bmatrix} \bar{x}_n(d) \\ 1-\bar{x}_n(d) \end{bmatrix} + \frac{1}{M(d)} \times \begin{bmatrix} \mathcal{I}\left(s_{n+1}^{(m)}=1, s_n^{(m)}=2\right) \\ -\mathcal{I}\left(s_{n+1}^{(m)}=2, s_n^{(m)}=1\right) \\ \mathcal{I}\left(s_{n+1}^{(m)}=2, s_n^{(m)}=1\right) \\ -\mathcal{I}\left(s_{n+1}^{(m)}=1, s_n^{(m)}=2\right) \end{bmatrix}. \quad (5)$$

Here $1-\bar{x}_{n+1}(d)$ denotes the fraction of susceptible nodes of degree d at time $n+1$.

The following example illustrates the SIS model. Consider a network where the nodes (users) sequentially adopt an innovation. A node m pays a cost $\mathcal{C}^m$ to adopt the innovation at time k, where the costs $\mathcal{C}^m$ for $m = 1, 2, \ldots$ are i.i.d random variables with a cumulative distribution function $\mathbb{P}_{\mathcal{C}}$. Adopting the innovation endows node m with a benefit ($\bar{b}^m$) proportional to the number of adopted neighbours a, i.e, node m adopts the innovation if $\bar{b}^m a > \mathcal{C}^m$. Hence $\bar{P}_{21}(d,a) = \mathbb{P}_{\mathcal{C}}(\bar{b}^m a)$, and $\bar{P}_{12}(d,a)$ could model an outdated innovation due to adoption of a different innovation by the neighbours.

B. Mean Field Population Dynamics

In practice even for small network size, the dimensionality of the SIS model, namely 2^M states, becomes intractable for modeling or signal processing algorithms. To obtain a useful generative model, this section presents a mean field dynamics approximation to the SIS model.

In the formulation of the mean field dynamics below, the following statistic forms a convenient parametrization of the transition probabilities of $\bar{x}_n$. Define $\alpha(\bar{x}_n)$ as the probability that a uniformly sampled link[6] in the network, at time n, points to an infected node [1], [11]. We call $\alpha(\bar{x}_n)$ as the *infected link probability*. Clearly,

$$\alpha(\bar{x}_n) = \times \frac{\sum_{d=1}^{D} (\#\text{ of links pointing to infected nodes of degree } d)}{\sum_{d=1}^{D} (\#\text{ of links pointing to nodes of degree } d)} = \frac{\sum_{d=1}^{D} d\,\rho(d)\,\bar{x}_n(d)}{\sum_{d=1}^{D} d\,\rho(d)}. \quad (6)$$

With α_n, the probability that a susceptible node of degree d, has exactly a infected neighbors is given by the binomial distribution[7] as $\binom{d}{a}\alpha_n^a(1-\alpha_n)^{d-a}$.

We can now characterize the transition probabilities of the entire population process $\bar{x}_n(d)$, whose sample path evolves according to (5). The transition probability of the population process from susceptible to infected is:

$$\mathbb{P}\left(\bar{x}_{n+1}(d) = \bar{x}_n(d) + \frac{1}{M(d)}\right) = \sum_{a=0}^{d} \bar{P}_{21}(d,a)\mathbb{P}\,(a \text{ out of } d \text{ neighbours infected}) = \sum_{a=0}^{d} \bar{P}_{21}(d,a)\binom{d}{a}\alpha_n^a(1-\alpha_n)^{d-a}. \quad (7)$$

The transition probability of the population process from infected to susceptible is evaluated similarly as:

$$\mathbb{P}\left(\bar{x}_{n+1}(d) = \bar{x}_n(d) - \frac{1}{M(d)}\right) = \sum_{a=0}^{d} \bar{P}_{12}(d,a)\mathbb{P}\,(a \text{ out of } d \text{ neighbours infected}) = \sum_{a=0}^{d} \bar{P}_{12}(d,a)\binom{d}{a}\alpha_n^a(1-\alpha_n)^{d-a}. \quad (8)$$

With the transition probabilities (7) and (8) of the population process, define the change in the fraction of infected and susceptible nodes in the population as:

$$P_{21}(d,\bar{x}_n(d)) \stackrel{\Delta}{=} (1-\bar{x}_n(d))\rho(d)\mathbb{P} \times \left(\bar{x}_{n+1}(d) = \bar{x}_n(d) + \frac{1}{M(d)}\right)$$

$$P_{12}(d,\bar{x}_n(d)) \stackrel{\Delta}{=} \bar{x}_n(d)\rho(d)\mathbb{P} \times \left(\bar{x}_{n+1}(d) = \bar{x}_n(d) - \frac{1}{M(d)}\right) \quad (9)$$

Thus the change in the fraction of population from susceptible to infected and vice versa, depends on the individual transition

[5]The stochastic difference equation (4) can be derived as follows. The martingale representation of a Markov chain $\mathcal{X}_k$ says that [22]:

$$\mathcal{X}_{k+1} = \mathcal{T}'\mathcal{X}_k + \mathcal{M}_k \quad (4)$$

where $\mathcal{T}$ is the transition matrix with elements (3), $\mathcal{M}_k$ is a martingale difference, $\mathcal{X}_k = e_i$ for some $i \in \{1, \ldots, (M(d)+1)\}$ and $e_i \in \mathbb{R}^{M(d)+1}$ is the unit indicator vector. Define

$$\mathcal{G} = \begin{bmatrix} 1 & \frac{(M(d)-1)}{M(d)} & \cdots & \frac{(M(d)-j)}{M(d)} & \cdots & \frac{1}{M(d)} & 0 \\ 0 & \frac{1}{M(d)} & \cdots & \frac{j}{M(d)} & \cdots & \frac{(M(d)-1)}{M(d)} & 1 \end{bmatrix}$$

Then from (4) it follows that $\mathcal{G}\mathcal{X}_k = \begin{bmatrix} \bar{x}_n(d) \\ 1-\bar{x}_n(d) \end{bmatrix}$ with $\bar{x}_n(d) \in \{0, \frac{1}{M(d)}, \ldots 1\}$ and so (5) follows.

[6]By link we mean a node and its associated edge. So for a graph with two nodes and one edge, there are two links: (node 1, edge), and (node 2, edge).

[7]Implicit to this binomial expression is the assumption that nodes associate randomly with other nodes in a network, *or the homogeneous mixing assumption*. This is a common assumption in modelling SIS dynamics [1].

probabilities $\bar{P}_{12}(d,a)$, $\bar{P}_{21}(d,a)$ from (7) and (8); current infected population state, $\bar{x}_n(d)$; and the degree distribution of the network, $\rho(d)$.

Define the following 2-dimensional simplices for $d = 1,2,\ldots,D$,

$$\mathcal{S}^d = \left\{ \begin{bmatrix} \bar{x}_n(d) \\ 1-\bar{x}_n(d) \end{bmatrix} : \bar{x}_n(d) \in [0,1] \right\}$$

and the following product space,

$$\mathcal{S} = \mathcal{S}^1 \times \mathcal{S}^2 \times \ldots \times \mathcal{S}^D .$$

With $\mathbf{1}$ denoting the D dimensional vector of ones and $\bar{x}_n'$ denoting the vector transpose of $\bar{x}_n$, let

$$\begin{aligned} \bar{\mathbf{x}}_n &= [\bar{x}_n',\ (\mathbf{1}-\bar{x}_n)']', \\ \mathcal{P}_{21}(\bar{\mathbf{x}}_n) &= [P_{21}(1,\bar{x}_n(1)),\ldots,P_{21}(D,\bar{x}_n(D)), \\ &\quad \times\ P_{12}(1,\bar{x}_n(1)),\ldots,P_{12}(D,\bar{x}_n(D))]' \\ \mathcal{P}_{12}(\bar{\mathbf{x}}_n) &= [P_{12}(1,\bar{x}_n(1)),\ \ldots,P_{12}(D,\bar{x}_n(D)), \\ &\quad \times\ P_{21}(1,\bar{x}_n(1)),\ldots,P_{21}(D,\bar{x}_n(D))]'. \end{aligned} \tag{10}$$

where $\bar{x}_n$ is the infected population state defined in (2).

With the above notation, the following theorem is the main result of this section. It gives a martingale representation of the population process $\bar{\mathbf{x}}_n$ and then asserts that for large population size M, the population process converges to a deterministic difference equation with high probability.

Theorem 1 (Mean Field Dynamics): (1) The population dynamics Markov chain $\bar{\mathbf{x}}_n$ defined in (10) evolves according to the following martingale difference driven stochastic difference equation:

$$\bar{\boldsymbol{x}}_{n+1} = \bar{\boldsymbol{x}}_n + \frac{1}{M}[\mathcal{P}_{21}(\bar{\boldsymbol{x}}_n) - \mathcal{P}_{12}(\bar{\boldsymbol{x}}_n)] + \zeta_n. \tag{11}$$

Here ζ_n is a $2D$-dimensional martingale difference process[8] with $\|\zeta_n\|_2 \le \frac{\Gamma}{M}$ for some positive constant Γ.

2) Consider the following deterministic mean field dynamics process associated with the population state:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \frac{1}{M}[\mathcal{P}_{21}(\boldsymbol{x}_n) - \mathcal{P}_{12}(\boldsymbol{x}_n)] \tag{12}$$

where $\boldsymbol{x}_n \in \mathcal{S}$, $\boldsymbol{x}_n = [x_n',\ (1-x_n)']'$, $\mathcal{P}_{21}(\boldsymbol{x}_n)$ and $\mathcal{P}_{12}(\boldsymbol{x}_n)$ are defined as in (10) and $\boldsymbol{x}_0 = \bar{\boldsymbol{x}}_0$.

Then for a time horizon of T points, the deviation between the mean field dynamics $\mathbf{x}_n$ in (12) and actual population state $\bar{\mathbf{x}}_n$ in (11) satisfies[9]

$$\mathbb{P}\big\{ \max_{0\le n\le T} \|\boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}_{n+1}\|_\infty \ge \epsilon \big\} \le C_1 \exp(-C_2\epsilon^2 M)$$

for some positive constants C_1 and C_2 providing $T = O(M)$.

Theorem 1 says that the maximum deviation between the mean field dynamics (12) and actual population state (11), over the entire T point sample path, satisfies an exponential bound. Since the exponential bound $\sum_M \exp(-C\epsilon^2 M)$ is summable, the Borel-Cantelli lemma applies. This in turn implies that, if the deterministic population flow remains forever in some subset of the state space, then the stochastic process will remain in the same subset space with a probability arbitrarily close to one, provided that the population is large enough, see [23].

The proof of Theorem 1 given in the appendix, is inspired by the proof in [23] for approximating stochastic evolution in games. An important feature of the mean field dynamics is that it has a state of dimension D compared to the intractable state dimension $\prod_{d=1}^{D}(M(d)+1)$ of the infected population state $\bar{x}_n$, with $\bar{x}_n(d) \in \{0, \frac{1}{M(d)}, \frac{2}{M(d)}, \ldots, 1\}$.

For the purposes of this paper, the key outcome of Theorem 1 is that the mean field system $\mathbf{x}_n$ has polynomial dynamics, where for every $d \in \{1,2,\ldots,D\}$ we have for the infected population state,

$$x_{n+1}(d) = x_n(d) + \frac{1}{M}\left[P_{21}(d,x_n(d)) - P_{12}(d,x_n(d))\right]. \tag{13}$$

These polynomial dynamics will be exploited in Section III for estimating the infected population state given noisy measurements.

C. Sampling

We now consider sampling the social network (1). For social networks with large number of nodes, it is prohibitive to query each node. This necessitates choosing a sampling methodology to estimate the infected population state x. We assume that the degree distribution ρ of the underlying network is known[10]. Each sampled node is asked if it is infected or not and the reply (measurement) noted. Below, we consider two popular methods for sampling large networks, see [13], [24]–[29] for an overview:

1) Uniform Sampling: At each time n, $\nu(d)$ individuals are sampled[11] independently and uniformly from the population $M(d)$ comprising of agents with degree d. Thus a uniformly distributed independent sequence of nodes, denoted by $\{m_l, l \in \{1,2,\ldots,\nu(d)\}\}$, is generated from the population $M(d)$. From these independent samples, the empirical infected population state $\hat{x}_n(d)$ of degree d nodes at each time n is

$$\hat{x}_n(d) = \frac{1}{\nu(d)} \sum_{l=1}^{\nu(d)} \mathcal{I}(s_n^{(m_l)} = 1). \tag{14}$$

At each time n, $\hat{x}_n$ can be viewed as noisy observation of the infected population state x_n.

2) MCMC Based Respondent-Driven Sampling (RDS): Respondent-driven sampling (RDS) was introduced by Heckathorn [27], [28] as an approach for sampling from hidden populations in social networks and has gained enormous popularity in recent years. In RDS sampling, current sample members recruit future sample members. The RDS procedure

[8] ζ_n is a martingale difference process if $\mathbb{E}\{\zeta_n | \mathcal{F}_{n-1}\} = 0$, where $\mathcal{F}_{n-1}$ denotes the sigma-algebra generated by $\{\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{n-1}\}$.

[9] Here $\|\mathbf{x}\|_\infty = \max_i |\mathbf{x}|$ denotes the l_∞ norm of vector $\mathbf{x}$.

[10] In Section IV, an optimal filter for estimating the underlying degree distribution ρ is outlined.

[11] For large population where $M(d)$ is large, sampling with and without replacement are equivalent.

is as follows: A small number of people in the target population serve as seeds. After participating in the study, the seeds recruit other people they know through the social network in the target population. The sampling continues according to this procedure with current sample members recruiting the next wave of sample members until the desired sampling size is reached.

RDS can be viewed as a form of Markov Chain Monte Carlo (MCMC) sampling (see [30] for an excellent exposition). Let $\{m_l, l \in \{1, 2, \ldots, \nu(d)\}\}$ be the realization of an aperiodic irreducible Markov chain with state space $M(d)$ comprising of nodes of degree d. This Markov chain models the individuals of degree d that are sampled, namely, the first individual m_1 is sampled and then recruits the second individual m_2 to be sampled, who then recruits m_3 and so on. Instead of the independent sample estimator (14), an asymptotically unbiased MCMC estimate is computed as

$$\frac{\sum_{l=1}^{\nu(d)} \frac{\mathcal{I}(s_n^{(m_l)}=1)}{\pi(m_l)}}{\sum_{l=1}^{\nu(d)} \frac{1}{\pi(m_l)}} \tag{15}$$

where $\pi(m)$, $m \in M(d)$, denotes the stationary distribution of the Markov chain m_l.

In RDS, the transition matrix and, hence, the stationary distribution π in the estimate (15) is specified as follows: Assume that edges between any two nodes i and j have symmetric weights W_{ij} (i.e., $W_{\mathrm{ij}} = W_{\mathrm{ji}}$). Node i recruits node j with transition probability $W_{\mathrm{ij}} / \sum_j W_{\mathrm{ij}}$. Then, it can be easily seen that the stationary distribution is $\pi(i) = \sum_{j \in V} W_{\mathrm{ij}} / \sum_{i \in V, j \in V} W_{\mathrm{ij}}$. Using this stationary distribution along with (15) yields the RDS algorithm. Since a Markov chain over a non-bipartite connected undirected network is aperiodic, the initial seed for RDS can be picked arbitrarily, and the estimate (15) is asymptotically unbiased [30].

The key outcome of this section is that by the central limit theorem (for an irreducible aperiodic finite state Markov chain), the estimate of the probability that a node is infected in a large population (given its degree) is asymptotically Gaussian. For a sufficiently large number of samples, observation of the infected population state is approximately Gaussian, and the sample observations can be expressed as

$$y_n = C x_n + v_n \tag{16}$$

where $v_n \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ is the observation noise with the covariance matrix $\mathbf{R}$ and observation matrix C dependent on the sampling process and $x_n \in \mathbb{R}^D$ is the infected population state and evolves according to the polynomial dynamics (13).

III. Non-Linear Filter and PCRLB for Bayesian Tracking of Infected Populations

In Section II, we formulated the mean field dynamics for the infected population state as a system with polynomial dynamics (13) and linear Gaussian observations (16) obtained by sampling the network. In this section, we consider Bayesian filtering for estimating the infected population state in large networks. Then posterior Cramér-Rao lower bounds (PCRLB) are obtained for these estimates.

A. Optimal Filtering of Infected Populations

We first describe how to express the mean field dynamics (13) in a form amenable to employing the non-linear filter described in [31].

1) Mean Field Polynomial Dynamics: Consider a D-dimensional polynomial vector $f(x) \in \mathbb{R}^D$:

$$f(x) = A_0 + A_1 x + A_2 xx' + A_3 xxx' + \ldots \tag{17}$$

where the co-coefficients $A_0, A_1, \ldots, A_i$ are dimension $1, 2, \ldots, (i+1)$ tensors, respectively. Note that $A_i xx \ldots x'$ is a vector with r^{th} entry given by

$$A_i xx \ldots x'(r) = \sum_{j_1, j_2, j_3, \ldots, j_i} A_i(r, j_1, j_2, \ldots, j_i) x_{j_1} x_{j_2} \ldots x_{j_i}$$

where $A_i(r, j_1, j_2, \ldots, j_i)$ is the $r, j_1, j_2, \ldots, j_i$ entry of tensor A_i and x_j is the j^{th} entry of x.

In (13), x_{n+1} evolves according to a polynomial function of x_n. Thus it can be expressed in the form (17) by constructing the tensors A_i from $\bar{P}_{12}$, $\bar{P}_{21}$ and $\rho(d)$. We illustrate how to express the mean field dynamics of (13) in the form (17) for[12] $d = 2$. First, we note the average degree is $\sum_{d=1}^{D} d\,\rho(d)$ and the link probability given in (6) can be expressed as $\alpha_n = \phi' x_n$, where:

$$\phi = \left[\frac{\rho(1)}{\sum_{d=1}^{D} d\,\rho(d)}, \frac{2\rho(2)}{\sum_{d=1}^{D} d\,\rho(d)}, \ldots, \frac{D\rho(D)}{\sum_{d=1}^{D} d\,\rho(d)} \right]'.$$

The mean field dynamics for $d = 2$ in (13) is given as:

$$x_{n+1}(2) = x_n(2) + \frac{1}{M} \left[P_{21}(2, x_n) - P_{12}(2, x_n) \right]. \tag{18}$$

For all terms containing $\bar{P}_{12}$ there is a corresponding term containing $\bar{P}_{21}$, so for convenience we will account for all of the former with Ω and the latter with Ω^*.

$$x_{n+1}(2) = x_n(2) + \Omega - \Omega^* \tag{19}$$

where from (13), for the case $d = 2$,

$$\begin{aligned}
\Omega = \Bigg[& \frac{1}{M} \bar{P}_{12}(2,0)(\phi' x_n)^2 + \frac{2}{M} \bar{P}_{12}(2,1)(\phi' x_n) \\
& - \frac{2}{M} \bar{P}_{12}(2,1)(\phi' x_n)^2 + \frac{1}{M} \bar{P}_{12}(2,2) \\
& - \frac{2}{M} \bar{P}_{12}(2,2)(\phi' x_n) + \frac{1}{M} \bar{P}_{12}(2,2)(\phi' x_n)^2 \\
& - \frac{x_n}{M} \bar{P}_{12}(2,0)(\phi' x_n)^2 - \frac{2x_n}{M} \bar{P}_{12}(2,1)(\phi' x_n) \\
& + \frac{2x_n}{M} \bar{P}_{12}(2,0)(\phi' x_n)^2 - \frac{x_n}{M} \bar{P}_{12}(2,2) \\
& + \frac{2x_n}{M} \bar{P}_{12}(2,2)(\phi' x_n) - \frac{x_n}{M} \bar{P}_{12}(2,2)(\phi' x_n)^2 \Bigg].
\end{aligned} \tag{20}$$

By grouping terms in (19) by their powers in $x_n(d)$, we can generate the tensors of (17). The contributions to the tensors of

[12] The procedure is same for all degrees.

(17) by Ω are:

$$\begin{aligned}
A_0(2) &= \frac{\bar{P}_{12}(2,2)}{M} \\
A_1(2,:) &= \phi\left[\frac{2(\bar{P}_{12}(2,1)-\bar{P}_{12}(2,0))}{M}\right] \\
A_2(2,:,:) &= \phi\phi'\left[\frac{(\bar{P}_{12}(2,0)-2\bar{P}_{12}(2,1)+\bar{P}_{12}(2,2))}{M}\right] \\
A_2(2,2,:) &= \phi\left[\frac{2(\bar{P}_{12}(2,1)-\bar{P}_{12}(2,0))}{M}\right] \\
A_3(2,2,:,:) &= -\phi\phi'\left[\frac{(\bar{P}_{12}(2,0)-2\bar{P}_{12}(2,1)+\bar{P}_{12}(2,2))}{M}\right]
\end{aligned} \tag{21}$$

where $A_i(j_1, j_2, \ldots, j_{i-1}, :, :)$ is a submatrix of tensor A_i and $A_i(j_1, j_2, \ldots, j_i, :)$ is a subvector of tensor A_i. By following (21) for Ω and Ω^* for all d, we can generate all the coefficients in the tensors of (17) from $\bar{P}_{12}$, $\bar{P}_{21}$, and $\rho(d)$. We note that the polynomial that defines the dynamics of the network is of order D^*+1, where D^* is the highest degree node with complex dynamics, i.e: $\bar{P}_{21}(d,a) = \bar{P}_{12}(d,a) = \kappa$ for all $d > D^*$ and all a, where κ is constant with respect to d and a.

2) Optimal Filter for Polynomial Dynamics: With the above formulation, we are now ready to describe the optimal filter to estimate the infected population state. Optimal Bayesian filtering refers to recursively computing the conditional density (posterior) $p(x_k|Y_k)$, for $k = 1, 2, \cdots$, where Y_k denotes the observation sequence $y_1, \ldots, y_k$. From this posterior density, the conditional mean estimate $\mathbb{E}\{x_k|Y_k\}$ can be computed by integration. (The term optimal refers to the fact that the conditional mean estimate is the minimum variance estimate). In general for nonlinear or non-Gaussian systems, there is no finite dimensional filtering algorithm, that is, the posterior $p(x_k|Y_k)$ does not have a finite dimensional statistic. However, it is shown in [31] that for Gaussian systems with polynomial dynamics, one can devise a finite dimensional filter (based on the Kalman filter) to compute the conditional mean estimate. That is, Bayes rule can be implemented exactly (without numerical approximation) to compute the posterior, and the conditional mean can be computed from the posterior. Therefore, to estimate the infected population state using the sampled observations (16), we employ this optimal filter.

The non-linear filter prediction and update equations are given as:

Prediction step:

$$\begin{aligned}
\hat{x}_n^- &= \mathbb{E}\{x_n|Y_{n-1}\} = \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} \\
H_n^- &= \mathbb{E}\{(x_n-\hat{x}_n)(x_n-\hat{x}_n)'|Y_{n-1}\} \\
&= \mathbb{E}\{(f(x_{n-1}) - \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} + v_{n-1}) \\
&\quad \times (f(x_{n-1}) - \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} + v_{n-1})'|Y_{n-1}\} \\
&= \mathbb{E}\{f(x_{n-1})f(x_{n-1})'|Y_{n-1}\} - \mathbb{E}\{f(x_{n-1})|Y_{n-1}\} \\
&\quad \times \mathbb{E}\{f(x_{n-1})|Y_{n-1}\}' + \mathbf{Q}_{n-1}
\end{aligned} \tag{22}$$

where $Y_n = \{Y_{n-1}, y_n\}$ denotes the observation process; H_n^- denotes the priori state co-variance estimate at time n; and v_n denotes the Gaussian state noise at time n, with covariance $\mathbf{Q}_n$.

The filter is initialized with mean $\hat{x}_0$ and covariance H_0^-. The filter relies upon being able to compute the expectation $\mathbb{E}\{f(x_{n-1})f'(x_{n-1})|Y_{n-1}\}$ in terms of $\hat{x}_{n-1}$ and H_n^-. When $f(\cdot)$ is a polynomial, $f(x_{n-1})f(x_{n-1})'$ is a function of x_{n-1}, and the conditional expectations in (23) can be expressed only in terms of $\hat{x}_{n-1}$ and H_n^-, permitting a closed form[13] prediction step.

Update step:

$$\begin{aligned}
\hat{x}_n &= \mathbb{E}\{x_n|Y_n\} \\
&= \hat{x}_n^- + H_n^- C'(\mathbf{R}_n + CH_n^-C')^{-1}(y_n - C\hat{x}_n^-) \\
K_n &= H_n^- C'(\mathbf{R}_n + CH_n^-C')^{-1} \\
H_n &= (I-K_nC)H_n^-(I-K_nC)' + K_n\mathbf{R}_nK_n'
\end{aligned} \tag{23}$$

where $\hat{x}_n$ denotes the conditional mean estimate of the state and H_n the associated conditional covariance at time n. C denotes the state observation matrix; $\mathbf{R}_n$ denotes the observation noise co-variance matrix; K_n denotes the filter gain; and I denotes the identity matrix.

Since the dynamics of (13) are polynomial, the prediction and update steps of (22) and (23) can be implemented without approximation. These expressions constitute the optimal non-linear filter and can be used to track the evolving infected population state.

B. Posterior Cramér-Rao Lower Bounds (PCRLB)

The PCRLB yields a useful deterministic lower bound to the covariance of the infected population state estimate computed by the optimal filter in Section III-A. In this section, we provide explicit expressions for the PCRLB. Recall that the CR bound yields a lower bound to the covariance of an unbiased estimator in the sense that the difference between the covariance matrix and the inverse of the Fisher Information Matrix is a *positive semi-definite* matrix. Here we are interested in determining lower bounds to the covariance of the conditional mean estimate of the infected population state. This lower bound is specified by the posterior CR lower bound (PCRLB). The main result [32] states that,

$$\mathbb{E}\{(\hat{X}-X)(\hat{X}-X)'\} \geq J^{-1} \tag{24}$$

where J is the Fisher Information matrix (FIM), X is the state, $\hat{X} = \mathbb{E}\{X|Y\}$ is the conditional mean state estimate, and Y is the observation. The elements of FIM matrix J are:

$$J(i,j) = \mathbb{E}\left\{\frac{\partial^2 \log p_{x,y}(X,Y)}{\partial X_i \partial X_j}\right\}.$$

Below we compute the PCRLB for the conditional mean estimate of the infected population state having polynomial dynamics (13). Recall that the population state has linear observations corrupted by Gaussian noise (16). The recursive computation

[13]For an explicit implementation of such a filter for a third order system with an exact priori update equation for H_n^- and $\hat{x}_n^-$, see [31].

of the PCRLB was first proposed by [5]. We first consider the dynamical system with mean zero Gaussian state noise having the covariance matrix $\mathbf{Q}$, then formulate the PCRLB for the polynomial dynamical system (13) without any state noise.

Let $X_n = \{X_{n-1}, x_n\}$ denote the state sequence up to time n and $J(X_n)$ denote the $((n+1)D \times (n+1)D)$ Fisher information matrix of X_n. Let $p_n = p_{x,y}(X_n, Y_n)$ denote the joint distribution at time n, and $\Delta_x^y = \nabla_x \nabla_y'$ denote the vector differential operator. Let J_n denote the $D \times D$ right lower block of $J(X_n)$. The matrix J_n^{-1} will provide a lower bound on the mean square error of estimating x_n. Based on [5], J_n can be evaluated as:

$$\begin{aligned} J_n &= \mathbb{E}\{-\Delta_{x_n}^{x_n} \log(p_n)\} \\ &- \mathbb{E}\{-\Delta_{x_n}^{X_{n-1}} \log(p_n)\}[\mathbb{E}\{-\Delta_{X_{n-1}}^{X_{n-1}} \log(p_n)\}]^{-1}\mathbb{E} \\ &\times \{-\Delta_{X_{n-1}}^{x_n} \log(p_n)\}. \end{aligned}$$

The recursion for J_n is given by:

$$J_{n+1} = \Lambda_n^{22} - \Lambda_n^{21}\left(J_n + \Lambda_n^{11}\right)^{-1}\Lambda_n^{12} \tag{25}$$

where

$$\begin{aligned} \Lambda_n^{11} &= \mathbb{E}\{(\nabla_{x_n} f_n'(x_n))\mathbf{Q}_n^{-1}(\nabla_{x_n} f_n'(x_n))'\} \\ \Lambda_n^{12} &= \mathbb{E}\{\nabla_{x_n} f_n'(x_n)\}\mathbf{Q}_n^{-1} \\ \Lambda_n^{21} &= \{\Lambda_n^{12}\}', \qquad \Lambda_n^{22} = \mathbf{Q}_n^{-1} + C\mathbf{R}_n^{-1}C', \end{aligned} \tag{26}$$

C is the linear observation matrix and $\mathbf{R}$ is the observation noise covariance matrix of (16) and

$$\begin{aligned} \nabla_{x_n} f_n'(x_n) &= \left[\frac{\partial f}{\partial x_n(0)}, \frac{\partial f}{\partial x_n(1)}, \ldots, \frac{\partial f}{\partial x_n(D)}\right]' \\ \frac{\partial f}{\partial x_n(0)} &= 0 + \frac{\partial}{\partial x_n(0)}\left[A_1 x_n\right] + \frac{\partial}{\partial x_n(0)}\left[A_2 x_n x_n'\right] \\ &\quad + \frac{\partial}{\partial x_n(0)}\left[A_3 x_n x_n x_n'\right]. \end{aligned} \tag{27}$$

Thus

$$\begin{aligned} \nabla_{x_n} f_n'(x_n) &= A_1 + (A_2 + A_2')x_n \\ &\quad + (A_{3_{\mathrm{ijk}}} + A_{3_{\mathrm{jki}}} + A_{3_{\mathrm{kji}}})x_n x_n' \end{aligned} \tag{28}$$

where $A_{3_{\mathrm{ijk}}}$ indicates the ordering of indices of the tensor A_3 and is analogous to a higher dimensional transpose.

Next consider the case when there is no state noise in the system evolution (13). Then to compute the PCRLB, we perturb the state evolution in (13) with pairwise independent Gaussian random vectors having covariance matrix $\mathbf{Q}_\epsilon = \epsilon I$, replacing the singular state evolution by a perturbed system $p_\epsilon(x_{n+1}|x_n)$. For the perturbed system we have,

$$\begin{aligned} -\log p_\epsilon(x_{n+1}|x_n) &= c + \frac{1}{2}\{x_{n+1} - f_n(x_n)\}'\mathbf{Q}_\epsilon^{-1} \\ &\quad \times \{x_{n+1} - f_n(x_n)\} \end{aligned} \tag{29}$$

where c is a constant. The recursion for J_n is then given by:

$$J_{n+1} = \Lambda_{\epsilon,n}^{22} - \Lambda_{\epsilon,n}^{21}\left(J_n + \Lambda_{\epsilon,n}^{11}\right)^{-1}\Lambda_{\epsilon,n}^{12} \tag{30}$$

where

$$\begin{aligned} \Lambda_{\epsilon,n}^{11} &= \frac{1}{\epsilon}\mathbb{E}\{(\nabla_{x_n} f_n'(x_n))(\nabla_{x_n} f_n'(x_n))'\} \\ \Lambda_{\epsilon,n}^{12} &= \frac{1}{\epsilon}\mathbb{E}\{\nabla_{x_n} f_n'(x_n)\} \\ \Lambda_{\epsilon,n}^{21} &= \{\Lambda_n^{12}\}', \quad \Lambda_{\epsilon,n}^{22} = \frac{1}{\epsilon}I + C\mathbf{R}_n^{-1}C' \end{aligned} \tag{31}$$

PCRLB: Erdős-Rényi vs Scale Free Network: With the above numerical procedure, we can compare the PCRLB for Erdős-Rényi versus scale free networks. For comparing the performance of the filter, it is convenient to consider a scalar version of the matrix inequality (24). A suitable measure is the Mean Square Error (MSE) given by the *trace* of (24),

$$\mathbb{E}\{(\hat{X} - X)'(\hat{X} - X)\} \geq \mathrm{tr}(J^{-1}). \tag{32}$$

Below we use trace of the inverse of Fisher Information Matrix (J^{-1}) in (32) as a measure of PCRLB for the mean field dynamics model (13), for two different network types:

i.) Scale-free network with degree distribution $\rho(d) \propto d^{-\gamma}$.

ii.) Erdős-Rényi network with degree distribution $\rho(d) \propto \frac{e^{-\lambda d}}{d!}$.

MFD evolution is studied under two simulation schemes:

Scheme A.) *Degree distribution based simulations*: We simulate a deterministic mean field evolution with a degree distribution corresponding to scale-free and Erdős-Rényi networks, where the observations are corrupted by Gaussian noise. Parameters for the simulating the mean field evolution: the observation noise covariance matrix $\mathbf{R}$ is a random positive definite matrix[14] with entries in $[0, 10^{-6}]$; the state transition probability matrices $\bar{P}_{12}$ and $\bar{P}_{21}$ were simulated with elements chosen uniformly at random over $[0, 1]$; the observation matrix $C = I$; and maximum degree $D = 20$.

Scheme B.) *Network based simulation*: The network based simulation involves generating a network, propagating an infection according to Section II-A over this network, and then finally sampling that network using uniform sampling described in Section II-C. Scale free networks and Erdős-Rényi networks having $M = 10000$ nodes were generated such that[15] $\gamma = 2.7$ and $\lambda = 2.7$. Parameters for the simulating the mean field evolution: the state transition probability matrices $\bar{P}_{12}$ and $\bar{P}_{21}$ were simulated with elements chosen uniformly at random over $[0, 1]$; and the state and observation noise covariance matrices $\mathbf{Q}_n$ and $\mathbf{R}_n$ for $n = 1, 2, \ldots, T$ were computed empirically. The infection was propagated over these networks for $T = 10000$ timesteps as follows: For each node

[14]One way of generating random samples of positive definite matrices is to sample from the Inverse-Wishart distribution.

[15]The value $\lambda = 2.7$ was chosen since it is similar to the out-degree of the World Wide Web

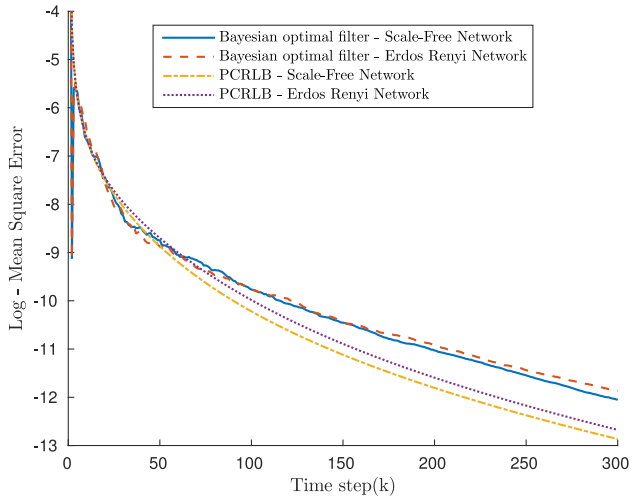


Fig. 1. Mean square error of the filtered estimate and trace of the PCRLB matrix for deterministic mean field evolution with Scale-free and Erdős-Rényi degree distributions. It can be seen that the filter (23) is insensitive to the underlying degree distribution. Both PCRLB (as in (32)) and its slope are insensitive to the underlying distribution when observation noise covariance does not depend on the network structure.

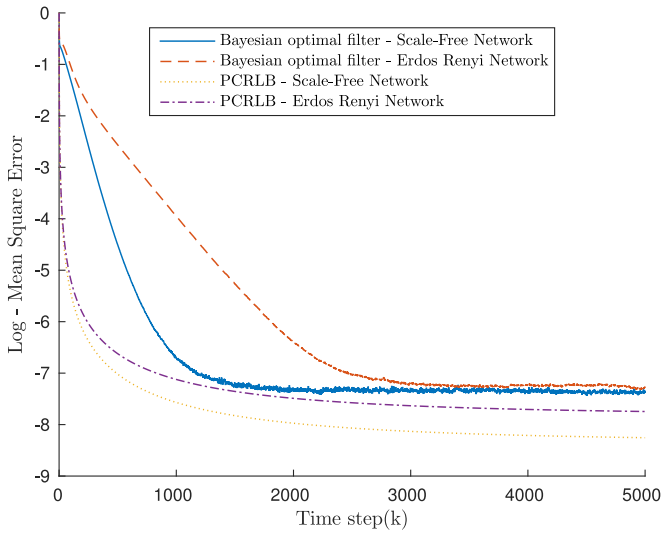


Fig. 2. Mean square error of the filtered estimate and trace of the PCRLB matrix in simulated networks - Scale-free and Erdős-Rényi. A network of 10000 nodes is generated and an infection is propagated through the network. This infection is then sampled 5000 times at each timestep to generate observations of the infected population states. State and observation covariances are computed empirically and assumed constant throughout the duration of the system dynamics. It is seen that PCRLB (as in (32)) gives a lower bound on the MSE.

m, the state was initialized by $\mathbb{P}(s_0^{(m)} = 1) = 0.01$. At each time n, a node is chosen at random from the network. Depending on its current state, the node becomes infected/susceptible with probabilities $\bar{P}_{21}(d,a)/\bar{P}_{12}(d,a)$. The time is then incremented to $n+1$ and another node is chosen at random from the network. This process repeats for T timesteps. At each timestep 5000 samples were obtained according to the uniform sampling described in Section II-C to generate an observation at that timestep.

Fig. 1 displays the PCRLB and mean square error of the infected population state estimate using the optimal filter in *degree-distribution based simulations* (Scheme A). Interestingly, it can be seen from Fig. 1 that when observation noise covariance, $\mathbf{R}$ in (16), is not network dependent both PCRLB and its slope are insensitive to the underlying network structure. The distinctions in PCRLB between the two network types can therefore be attributed to different state or observation noise covariances.

The PCRLB and mean square error of the infected population state estimate using the optimal filter, in *network based simulations* (Scheme B), are shown in Fig. 2. The slower convergence of the filtered estimate for an Erdős-Rényi compared to a Scale-Free network filter estimate is because the Erdős-Rényi network observations have a larger observation noise covariance as explained in Footnote 25 in Section V-A1.

IV. Analysis of Infection Diffusion in Evolving Social Networks

So far in this paper, we have discussed estimating infection diffusion in a fixed network. In this section, we consider networks that evolve with time, represented by time varying degree distributions, and analyse their effect on the diffusion of infection over time.

In social networks such as Twitter[16], information diffusion occurs at a faster time scale compared to underlying network evolution. Therefore, we consider a two time scale formulation: the degree distribution of the underlying network evolves on a *slow time* scale (denoted by k) and the infection diffuses over a *fast time* scale (denoted by n). There are various generative models for time evolving networks in the literature, see [33], [34], and the references therein. In this paper, we consider the preferential attachment model discussed extensively in [33], to model the time evolution of the underlying degree distribution. The primary motivation for choosing a preferential attachment graph is that it is the simplest graph whose steady state distribution obeys a power law [33], which commonly arises in several real world networks, see [7], [35], [36].

A. Preferential Attachment Model for Network Evolution

A network evolving according to the preferential attachment model is characterized by two parameters - a probability u and an initial graph G_0. The graph evolves as follows:

1) *Vertex-Step*: A vertex of the existing graph is chosen independently with probability proportional to its degree. Then a new vertex is connected to this chosen vertex.
2) *Edge-Step*: A new edge is added between two vertices of the graph chosen independently with probability proportional to their degrees.

[16]This is not true in general. Face-to-face interactions occur at a much faster time-scale than the rate at which people change their political affiliation.

At each time step, Vertex-step is realized with probability u, while the Edge-step is realized with probability $1-u$, where $u \in (0,1)$.

Let $M_k(d)$ denote the number of vertices of degree d at time k and let

$$\rho_k(d) = \frac{\mathbb{E}\{M_k(d)\}}{M_k}$$

denote the expected fraction (degree distribution) of vertices of degree d at time k, where M_k denotes the total number of nodes at time k in the network.

A vertex of degree d at time k can originate from two outcomes, either it was a vertex of degree d at time $k-1$ and had no edge added to it, or it was a vertex of degree $d-1$ at time $k-1$ and a new edge was added to it. The recursion for the degree distribution $\rho_{d,k}$ can be expressed as [33]:

$$\rho_k(d) = \left(1 - \frac{(2-u)d}{2k}\right)\rho_{k-1}(d) + \left(\frac{(d-1)(2-u)}{2k}\right)\rho_{k-1}(d-1) \tag{33}$$

Let

$$\rho_k = [\rho_k(1), \rho_k(2), \ldots, \rho_k(N), \rho_k(N^{(+)})]'$$

denote the degree distribution at time k. The state $\rho_k(N^{(+)})$ represents the fraction of nodes of degree greater than N. In matrix-vector notation, the recursion in equation (33) can be written as

$$\rho_k = L_k' \rho_{k-1}, \text{ where } L_k = I + \epsilon_k F, \quad \epsilon_k = \frac{1}{2k} \text{ and}$$

$$F = \begin{bmatrix} -(2-u) & (2-u) & 0 & \ldots 0 & 0 \\ 0 & -(2(2-u)) & (2(2-u)) & \ldots 0 & 0 \\ \vdots & \ddots & & & \\ 0 & 0 & \ldots & 0 & 0 \end{bmatrix}. \tag{34}$$

Here F is a generator matrix for a continuous time Markov chain (zero row sum and negative diagonal elements) and L_k is a stochastic matrix (row sum equal to 1) at each time k. The compound state, $\rho_k(N^{(+)})$, is modeled as an absorbing state as either edges or vertices are added during network evolution and no deletion takes place - once a node is of degree greater than d, it will continue to have degree at least d.

Equation (34) is a generative model for the evolution of the degree distribution. It is interesting to note that (34) has the form of the Chapman-Kolmogorov equation for a Markov chain η having the state space $\{1, 2, \ldots, N^+\}$. Thus, the Chapman-Kolmogorov equation is a generative model for the evolution of the network. In what follows, we will outline a filtering scheme to estimate the degree distribution as it evolves over time.

B. Filtering for Estimating the Degree Distribution

So far in this paper, we assumed that the degree distribution of the network is known. We now describe a Bayesian filter to estimate the degree distribution, when the degree distribution evolves according to the preferential attachment model[17].

In a co-evolving system, the asymptotic infected population state x_∞ can in turn influence the network rearrangement, ρ_k, at a future time $k+1$. For the preferential attachment model of Section $IV-A$, this influence can be modeled as the probability u being dependent on α_∞ (which depends on x_∞).

We exploit the Chapman-Kolmogorov generative model (34) to estimate the degree distribution by deriving a representative sample that captures the link between the degree distribution ρ and the asymptotic population state x_∞. This edge could be an important factor in determining the way connections are formed in social networks; see for example, [37]; where, similarity between individuals (homophily) breeds connection[18]. For nodes which currently are not connected, being infected increases the probability of forming an edge at a future time instant.

Below, we consider the mode of the asymptotic infected degree distribution[19] as the representative sample to estimate the degree distribution. The mode of the asymptotic infected degree distribution gives the degree with the largest fraction of infected individuals and tracking the mode can provide useful information on the nature of infection diffusion over the slow time scale k.

Let the initial estimate be $\hat{\rho}_0$, which denotes the probability distribution of the mode over the set $\{1, 2, \ldots, N^{(+)}\}$. Let $z_k \in \{1, 2, \ldots, N^{(+)}\}$ denote the mode of the infected degree distribution ρ_k at time k. Given this observation $z_k \sim \mathbb{P}(z|\rho_k)$[20] at time k, and the dynamics of the degree distribution (34), define the posterior distribution as

$$\hat{\rho}_{k+1} = \mathbb{P}(\rho_{k+1}|z_1, \ldots, z_{k+1}).$$

Then it is easily seen that the evolution of the posterior distribution is given as the Hidden Markov Model (HMM) filter [22]:

$$\hat{\rho}_{k+1} = \frac{B_{z_{k+1}} L_{k+1}' \hat{\rho}_k}{\mathbf{1}' B_{z_{k+1}} L_{k+1}' \hat{\rho}_k} \tag{35}$$

where B_{z_k} is a diagonal matrix having z_k^{th} column of the (noisy) mode observation distribution matrix[21] B as its elements.

To summarize, (35) together with the filtering algorithm in Section III-A constitutes a two time scale tracking algorithm: on the slow time scale, the degree distribution of the social network is updated based on sampling according to (35).

[17] Recall that Sec III deals with filtering to track the infected population state in a fixed network with fixed known degree distribution.

[18] For numerical examples of homophily see Section V-B2

[19] The infected degree distribution is defined as the fraction of infected nodes of each degree given by: $\frac{\sum_m \mathcal{I}\left(\Xi^{(m)}=d, s^{(m)}=1\right)}{\sum_m \sum_d \mathcal{I}\left(\Xi^{(m)}=d, s^{(m)}=1\right)}$.

[20] As is conventional in filtering, we assume the following conditional independence: $\mathbb{P}(z|\rho_k, z_{k-1}, \cdots, z_0) = \mathbb{P}(z|\rho_k)$.

[21] The mode observation distribution matrix B is assumed to be known for estimating the degree distribution. However, it can be estimated as follows: A network having the mode (of asymptotic infected degree distribution) as z, is sampled either uniformly or using RDS. On each queried node, the state of infection $\{1,2\}$ and its degree is noted down. Using the samples, the asymptotic infected degree distribution is estimated and a noisy estimate of the mode, $\hat{z}$, is found. This is repeated many times and the fraction of the time mode is equal to $\hat{z}$ corresponds to the element $B(z, \hat{z})$: the probability of observing the mode as $\hat{z}$ given the (true) mode was z.

On the fast time scale, these estimates are used in the filter (Section III-A) to track the infected population state. We refer to [38] for a formal proof of the optimality of this two-time scale filtering algorithm.

C. Effect on Diffusion Threshold in SIS Model

On networks having fixed degree distribution, [1] identified conditions under which a network is susceptible to an epidemic using a mean-field approach and provided a closed form solution for the diffusion threshold for infection diffusion. It was shown that under reasonable conditions on the infection probabilities, the diffusion threshold decreases with the mean preserving spread. In this section, we extend this analysis of diffusion thresholds to the case of evolving networks specified by the preferential attachment model. By using results in stochastic dominance, we show below that in a preferential attachment model for a randomly evolving graph, the infection diffusion threshold decreases with the attachment probability.

We first establish a relation between the addition probability u in the preferential attachment model and the diffusion threshold θ_* (defined below) in the SIS model. While this has been explored numerically in [11], [18]; below we prove that an ordering of the transition probabilities u in (34) results in a corresponding order of the diffusion thresholds. Such a result is useful since it allows ordering preferential attachment models in terms of their diffusion threshold. The proof relies on the Chapman Kolmogorov generative model (34) for the dynamics of the preferential attachment model.

Definition 1 ([1]): The diffusion threshold[22] is

$$\theta_* = \inf\{\theta > 0 : x_\infty \in \mathbb{R}_+^D\}$$

where x_∞ denotes the asymptotic infected population state in (13) and $\theta = \frac{\mu_{21}}{\mu_{12}}$ is the ratio of μ_{12} and μ_{21} in (3).

In words, the diffusion threshold $\theta_* \in \mathbb{R}^+$ is the value of θ such that starting from a small fraction of infected agents in the network, the dynamics converges to a *positive* fraction of infected agents for all $\theta > \theta_*$. Note that Definition 1 requires the existence of asymptotic infected degree distribution x_∞. This is specified in Lemma 1 below.

Let $\varrho(\alpha) = \frac{1}{d}\sum_{d\geq 1} d\rho(d) \frac{\mathbb{P}(\bar{x}_{n+1}(d)=\bar{x}_n(d)+1/M(d))}{\mathbb{P}(\bar{x}_{n+1}(d)=\bar{x}_n(d)+1/M(d))+(1-\bar{x}(d))}$, where $\mathbb{P}\left(\bar{x}_{n+1}(d) = \bar{x}_n(d) + \frac{1}{M(d)}\right)$ is a function of α and $(1-\bar{x}(d))$ are as in (7).

Lemma 1 ([1]): x_∞ exists iff $\frac{d\varrho(0)}{d\alpha} > 1$.

Lemma 1 asserts that there exists an asymptotic infected population state if $\varrho(\alpha)$ has a slope greater than 45^0 at the origin. The asymptotic infected link probability α_∞ can be calculated from x_∞ using (6).

Theorem 2 below is the main result of this section. It characterizes of the diffusion threshold of the SIS model as a function of the addition probability u, for a preferential attachment graph[23]. It is of interest since it is a monotone comparative static result, i.e., it determines how the argument of the minimum (namely, θ_*) behaves with respect to a partial ordering of the transition matrices; we refer the reader to [39] for a comprehensive discussion of monotone statics.

Theorem 2: Consider a time evolving preferential attachment network with transition matrix L_k given by (34), where $u > 0$. For any initial degree distribution ρ_0, let ρ_k^u denote the degree distribution at time k and $\theta_*^u(k)$ denote the diffusion threshold for the network with addition probability u. Then,

1) ρ_k^u is first-order stochastically decreasing in u for every $k > 1$, where k denotes the slow-time index.
2) $\theta_*^u(k)$ is increasing in u.

The proof of Theorem 2 is in the appendix. The first part of Theorem 2 asserts that $\rho_k^{u_2} >_{sd} \rho_k^{u_1}$, [24] for $u_1 > u_2$, i.e., networks that have higher probability of edge addition always have higher degree distributions as the network evolves. The second part of Theorem 2 asserts that the diffusion threshold increases with the probability u of adding new vertices.

V. Numerical Results

Section V-A below examines the effect of sampling and model mis-specification on the performance of the non-linear filter discussed in Section III-A. This analysis is useful for selecting the sampling methodology and for assessing the performance trade-off due to imperfect knowledge of the degree distribution of the underlying network.

Section V-B presents an SIS model of networks with homophily, a numerical example of filtering on such a network, and a numerical example of the role of homophily in modeling SIS dynamics.

Section V-C presents the performance analysis on a real-world Twitter dataset. First, using a goodness-of-fit test, we validate the sufficiency of the SIS model to capture the infection propagation on the Twitter network. Second, the non-linear filter of Section III-A is shown to track the infection diffusion satisfactorily.

Section V-D illustrates the performance of the non-linear filter of Section III-A and the HMM filter of Section IV on a simulated co-evolving system. This *two-time scale* simulation illustrates the choice of time scales for the propagation and tracking of infection dynamics, and estimation of the degree distribution.

A. Factors Affecting Filter Performance

As in Section III-B, we consider ER and SF networks and analyze the effect of sampling and model mis-specification on the filter performance. To isolate these numerical examples from the approximation of $\mathbf{R}_n$, as well as state noise, we use Scheme A of Section III-B, which simulates a deterministic trajectory with noisy observations.

[22] The diffusion threshold can be evaluated explicitly as [1]:

$$\theta_* = \frac{\sum_{d=1}^D d\,\rho(d)}{\sum_{d\geq 1} d^2\rho(d)\mathbb{F}_{21}(d,1)}, \quad \text{where } \mathbb{F}_{21}(d,a) \text{ was defined in (3).}$$

[23] It should be noted that the probability u itself can be a function of α_∞ as the degree distribution and infected degree distribution are evolving on different time scales.

[24] $>_{\mathrm{sd}}$ denotes first-order stochastic dominance (see Appendix for definition)

1) Effect of Sampling on Filtering: The observation noise variance $\mathbf{R}$ in (16) depends on the sampling method employed. Using the uniform sampling mechanism outlined in Section II-C, the effect of sampling on filtering is illustrated.

Under uniform sampling, the noise variance of each observation depends upon the network degree distribution[25] and consequently parameters γ and λ. In Fig. 3, the mean square error of the filter estimate is seen to depend on the network parameters λ and γ.

In these simulations, the observation error covariance matrices were chosen as the error covariance matrices from *network based simulations*[26] (Scheme B).

2) Sensitivity of Filter Performance to Mis-Specified Model: The SIS model in Section II-A is said to be mis-specified if the degree distribution ρ is not specified. The Bayesian filter (Section III-A) implemented using a mis-specified model is referred to as a mis-specified filter. Fig. 4 compares the MSE of a Bayesian filter and a mis-specified filter, both formulated for the same underlying network. The degree distribution of the mis-specified filter is assumed to be $\rho(d) = \frac{1}{D} \quad \forall d$. For comparison, we considered a moving average (linear) estimator:

$$\hat{\hat{x}}_n = \vartheta_1 y_{n-1} + \vartheta_2 y_{n-2} + \vartheta_3 y_{n-3} + \ldots + \vartheta_\iota y_{n-\iota} + \vartheta_0. \tag{36}$$

Here $\hat{\hat{x}}_n$ is the moving average estimate at time n, the matrices ϑ_i are computed using multivariate least squares estimation, and time delay ι was chosen to be 10. It is observed in Fig. 4 that, even when the degree distribution is mis-specified, the Bayesian filters outperform the moving average filter with an MSE of the order of 10^{-6}, compared to 10^{-4} of the moving average filter.

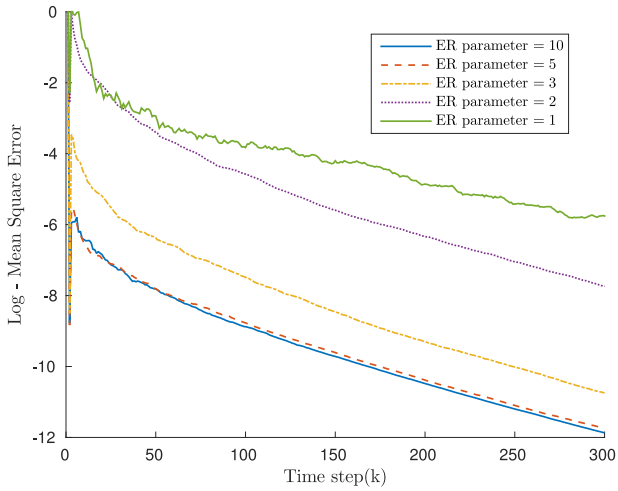
B. Infected Population Estimation in Homophily Networks

This section considers a small extension of the models considered in this paper to homophily networks. Homophily is the tendency for individuals to engage with people similar to themselves [40]. These similarities characterize distinct groups which can affect how infections spread in a network.

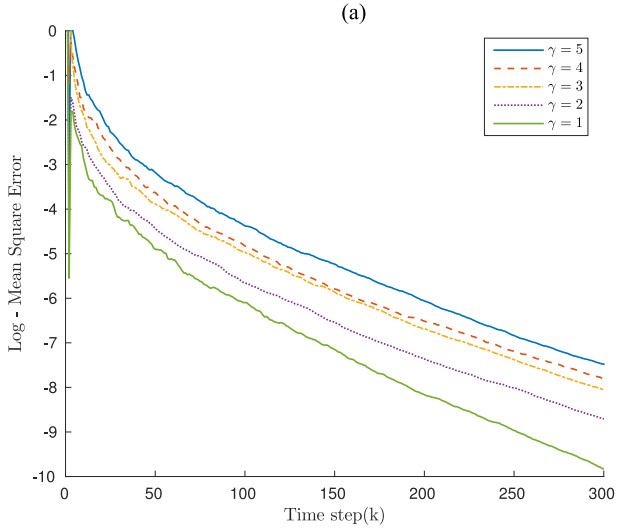
Below we show that the mean field dynamics model yields a similar polynomial structure to Section II-B implying that the optimal filers of Section III-A can be used to estimate the infected population state. Section V-B1 presents the SIS model with homophily using the approach in [41]. Section V-B2 illustrates the performance of the optimal filter for estimating the underlying infected population state in a homophily network.

[25]In the uniform sampling of Section II-C, for any given degree, an observation at time n will have a variance which corresponds to that of a scaled binomial distribution $\sigma^2(\hat{x}_n(d)) = \frac{x_n(d)(1-x_n(d))}{\nu(d)}$, where $\nu(d)$ are the number of samples of nodes of degree d and $\nu = \sum_d \nu(d)$. For a large number of independent samples, by the central limit theorem, $\nu(d) \approx \nu\rho(d)$, and the observation noise variance is inversely related to the probability that a node is of degree d.

[26]We isolate the experiment from finite network state noise, which can dominate the MSE of the filter estimate. By simulating the system with Scheme A, with observation noise informed by the synthetic network simulations of Scheme B, we can observe the effect of sampling more clearly.



(a)



(b)

Fig. 3. Figures display the log mean square error of Bayesian filter estimates for two networks: Erdős Rényi and scale free. The uniform sampling mechanism described in Section II-C depend on the degree distribution of the underlying network. These figures exhibit the effect uniform sampling has on filter performance. (a) Log mean square error of filter estimates for an Erdős Rényi network simulated using Scheme A in Section III-B. It can be seen that the MSE decreases as Erdős Rényi parameter λ increases. (b) Log mean square error of filter estimates for an scale-free network simulated using Scheme A in Section III-B. For larger scale-free parameter γ, it is observed that the MSE increases.

1) SIS Model with Homophily: Suppose the social network consists of $\mathcal{B}$ distinct types of individuals[27]. Let μ_b denote the fraction of individuals of type b, for $b \in \{1, 2 \ldots, \mathcal{B}\}$. Let $\mathbb{T}(m) \in \{1, 2, \ldots, \mathcal{B}\}$ denote the type of node m. Then the number of nodes of type b and the number of nodes of type b and degree d, respectively, are

$$M^b = \sum_{m \in V} \mathcal{I}\big(\mathbb{T}(m) = b\big),$$

$$M^b(d) = \sum_{m \in V} \mathcal{I}\big(\Xi^{(m)} = d, \mathbb{T}(m) = b\big).$$

[27]For example these types may be groups characterized by political affiliation, age, or income.

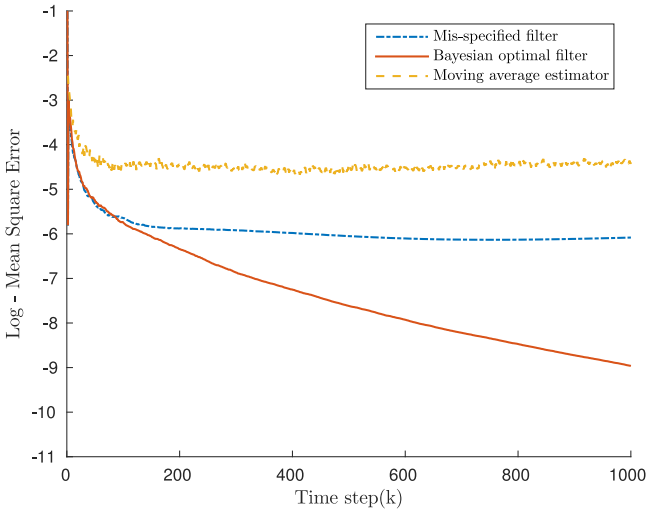


Fig. 4. Comparison of the mean square error between a Bayesian filter, a mis-specified Bayesian filter and a moving average estimator. The transition probabilities matrices $\bar{P}_{12}$ and $\bar{P}_{21}$ were simulated with elements chosen uniformly at random over $[0,1]$. The observation noise covariance matrix R is a random positive definite matrix with entries in $[0, 3\times 10^{-2}]$. The mis-specified filter, derived with $\rho(d)=\frac{1}{D}$, exhibits a plateau in performance. This plateau corresponds to the incorrect computation of the priori state estimate, $\hat{x}_n^-$, due to the misspecified filter parameters.

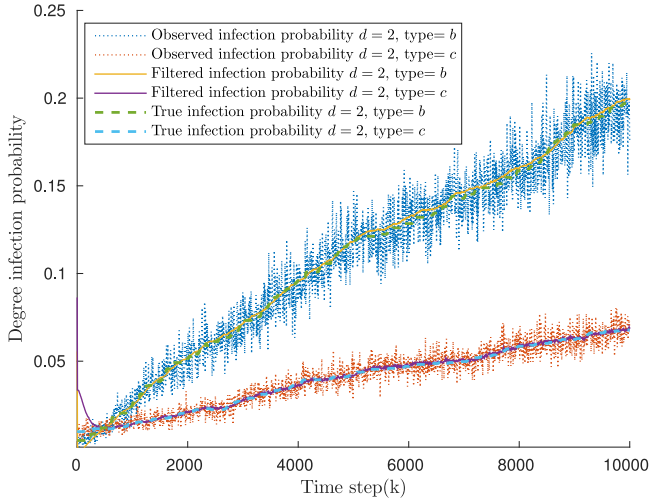


Fig. 5. Diffusion of infected population states and their corresponding filtered estimates in a scale-free network with $d=2$ in (39). The network was simulated using 'Scheme B' of Section III-B. Each node was labeled as type b or c with probability half. The infection transition probabilities of matrix $\bar{P}_{21}^b$ were chosen uniformly at random over $[0,1]$ and $\bar{P}_{21}^c=\frac{1}{4}\bar{P}_{21}^b$. $\varsigma_{\mathrm{bb}}=\varsigma_{\mathrm{cc}}=0.8$. It can be seen that the filtered estimates of the infected population state converge to the true state for both population types.

Then the degree distribution of each type b is

$$\rho^{[b]}(d)=\frac{M^b(d)}{M^b}. \tag{37}$$

These different types interact according to a transition matrix Υ, with elements[28]

$$\varsigma_{\mathrm{bc}} = P(\text{individual of type } c \text{ meets individual of type } b).$$

Let $\bar{x}_{n,b}(d)$ denote the fraction of agents of type b and degree d that are infected at time n. The main difference is that instead of a single infected link probability, used in Section II-A, we now need a vector of infected link probabilities, one for each type:

$$\alpha_b(\bar{x}_n)=\sum_{c=1}^{\mathcal{B}}\varsigma_{\mathrm{bc}}\frac{\sum_{d=1}^{D} d\,\rho^{[c]}(d)\,\bar{x}_{n,c}(d)}{\sum_{d=1}^{D} d\,\rho^{[c]}(d)}. \tag{38}$$

As in Section II-B, we can define[29] the probabilities P_{12}^b and P_{21}^b. Similar to Theorem 1, we have following mean field dynamics for the infected population state:

$$x_{n+1,b}(d)=x_{n,b}(d)+\frac{1}{M}\left[P_{21}^b(d,x_{n,b}(d)) - P_{12}^b(d,x_{n,b}(d))\right]. \tag{39}$$

Notice that the infected population state $x_{n,b}(d)$ has polynomial dynamics and is therefore amenable to the optimal filtering algorithms of Section III-A.

2) Numerical Examples of Homophily: A network was simulated according to Scheme B of Section III-B with two population types, b and c, as shown in Fig. 5. Fig. 5 shows that the filtered estimate of the infected population state converges to the true infected population state. However, the different population types (b and c) behave differently. This can be attributed to the fact that the insular structure and smaller infection probabilities of type c, result in a small fraction of type c getting infected.

The SIS model with homophily also enables the use of additional degree distribution information[30] so that the mean field dynamics provide a more accurate representation of the population dynamics. Fig. 6 illustrates the effect additional degree distribution information has on the network dynamics. By comparing the deterministic mean field trajectories for two SIS models, one with homophily and one without. We consider a scale free network, where individuals with d neighbours associate primarily with others having d neighbours. The network has a degree distribution $\rho(d)\propto d^{-\gamma}$ with $\gamma=3.0$, and maximum degree $D=20$. The degree distribution of each type

[28]The elements ς_{bc} are constrained such that the number of interactions from type b to type c are the same as those from type c to type b [41].

[29]The transition probabilities and the change in the fraction of infected/susceptible nodes in the population are evaluated, respectively, as

$$\mathbb{P}\left(\bar{x}_{n+1,b}(d)=\bar{x}_{n,b}(d)-\frac{1}{M^b(d)}\right)$$

$$=\sum_{a=0}^{d}\bar{P}_{12}^b(d,a)\binom{d}{a}\alpha_{n,b}^a(1-\alpha_{n,b})^{d-a}.$$

$$P_{21}^b(d,\bar{x}_{n,b}(d))\stackrel{\Delta}{=}(1-\bar{x}_{n,b}(d))\rho^{[b]}(d)\mathbb{P}$$

$$\times\left(\bar{x}_{n+1,b}(d)=\bar{x}_{n,b}(d)+\frac{1}{M^b(d)}\right)$$

[30]For example, solitary people may be more likely to also associate with other solitary people, resulting in high level of internal connections within nodes of low degree. Thus the degree distribution of low degree nodes differs from that of high degree nodes.

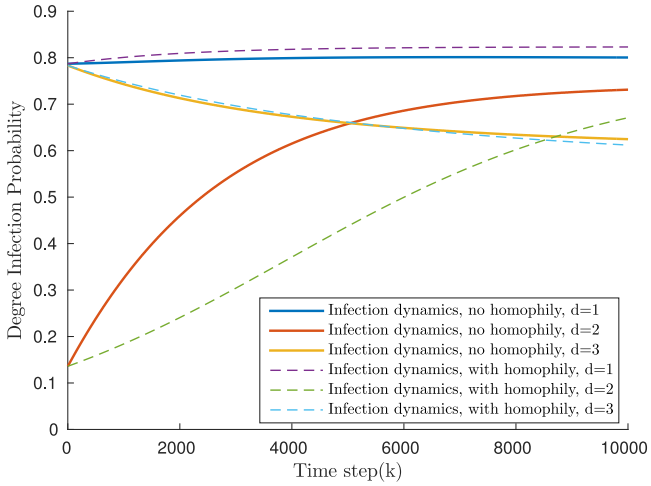


Fig. 6. Comparison of deterministic mean field trajectories of an SIS model with and without homophily for scale-free networks with $\gamma = 3.0$, and the degree specific degree distributions, $\rho^{[b]}(d)$, in the simulation with homophily is given by (40) with $\epsilon = 0.02$. Under the SIS model presented in Section II the systems modeled here are identical, however, the model with homophily incorporates additional degree distribution information so that the mean field dynamics provide a more accurate representation of the population dynamics.

$b \in \{1, \ldots, \mathcal{B}\}$ is given by:

$$\rho^{[b]}(d) = (1-\epsilon)\mathcal{I}(d = b-1) + \epsilon\rho(d). \tag{40}$$

where $\mathcal{I}$ is an indicator function and $\epsilon \in [0,1]$ is a parameter that captures the interconnectedness of nodes of different degrees. The key observation here is that we can utilize additional degree distribution information and the SIS model with homophily such that the mean field dynamics yield a more accurate representation of the population dynamics.

C. Analysis and Validation on Twitter Dataset

This section illustrates the tracking of infection diffusion on social networks using real diffusion data from the microblog platform Twitter. We analyze the diffusion of information through the Twitter Social network to demonstrate the effectiveness of the SIS model of Section II; see also [6], [42]. Twitter played a critical role in the Egyptian revolution of 2011 or January 25^{th} (#Jan25) uprising [43]. Twitter was used by protesters to organize the protest and recruit members and as a medium to discuss and share information about the protest. Below, we refer to the interest and engagement with the news of the uprising as *infection* and track the distribution of infection over time. Modeling this online process as an epidemic is supported by the virality of the #Jan25 hashtag.

1) Dataset: The dataset consists of tweets sampled between January 23^{rd} and February 8^{th}, 2011 and are available from Twitter (http://trec.nist.gov/data/tweets/). The tweet collection period encapsulates the time-frame of the first major developments relating to the January 25^{th} uprising event. In Twitter, a "hashtag" follows the discussion topics, i.e., a word or a phrase prefixed with the number sign #. We make use of the hashtags to track the spreading of a specific topic on Twitter. The most used hashtag related to this protest is "#Jan25". To obtain the information spreading among users participating in this protest, we filtered 26,313 tweets containing "#Jan25" published by 13,341 different users, from around 10 million tweets. These tweets contain the event of interest and the social network is (re-)constructed from them as follows: two users are connected if one user has mentioned another user ("@username") in the tweet containing "#Jan25" at least once over the duration of interest. We analyze information diffusion on this constructed social network.

All users in the constructed social network are assumed to be susceptible initially. Users who initiate tweets on the event of interest are assumed to be infected and act as seeds for the spread of information. Once a user, say User#A, becomes infected, it has some constant probability of becoming susceptible in each time period. This modeling assumption is motivated by the frequently observed Poisson-like decay of an individual's interest in social media topics [44]. The decay probability was chosen to be 0.001, motivated heuristically by an average interest duration of 2 hours. Our dataset is thus a hybrid dataset, wherein the network and infection process are informed by true Tweets between users, and these infected individuals become susceptible according the synthetic Poisson-like process described above.

2) SIS Model for Twitter Data: Active users can be considered 'infected', and inactive users can become 'infected' by interacting with other 'infected' individuals, in particular, any of its active neighbors in a social network. In this way, engagement and knowledge of a topic spreads throughout the network. Users can also become disinterested in a subject they have already been exposed to, in this way they are not currently engaged, but may become engaged if contacted by an infected neighbor; thus inactive individuals are assumed to be susceptible.

Model Evaluation (Goodness of Fit for SIS model): We used the Kolmogorov Smirnov test on the empirical infected degree distribution at the final timepoint to evaluate the goodness of fit of the SIS model, where the infection was mapped to engagement in the January 25th uprising. The KS test statistic was 0.2286 with p-value 0.2813. The null hypothesis for this statistical test is that both the observed Twitter and SIS model infected degree distributions are samples of the same infected degree distribution. At a confidence level of 0.01, the null hypothesis cannot be rejected. This test therefore provides little evidence for or against the quality of SIS model for Twitter. To further explore the agreement between the model and the true data we computed the average and maximum square difference between the Twitter data and predicted SIS degree infection probabilities. These values can be seen in Table I and the trajectories are shown in Fig. 7.

The low magnitude of the model deviations in Table I and the failure to reject the hypothesis that the Twitter data and model infected degree distributions come from the same distribution, suggest that the SIS model is a satisfactory model with respect to the infection dynamics of interest in the January 25^{th} uprising.

3) Sampling for Tracking the Infected Population State: The mean field dynamics for the SIS model can be used to track and predict the evolution of the infection on Twitter. We must generate estimates of $\bar{P}_{12}$, $\bar{P}_{21}$, and determine the degree distribution from samples obtained from (16). We compute the

TABLE I
GOODNESS OF FIT OF THE SIS MODEL TO THE TWITTER DATASET: THE AVERAGE AND MAXIMUM DEVIATIONS BETWEEN THE TWITTER DATA AND SIS MODEL PREDICTIONS ARE PRESENTED. NETWORK IS SCALE FREE

Degree	1	2	3+
Average Square Difference	0.0011	0.0014	0.0235
Average Absolute Difference	0.0273	0.0294	0.1000
Maximum Absolute Difference	0.0644	0.0719	0.8403

Note that the degree distribution of the twitter

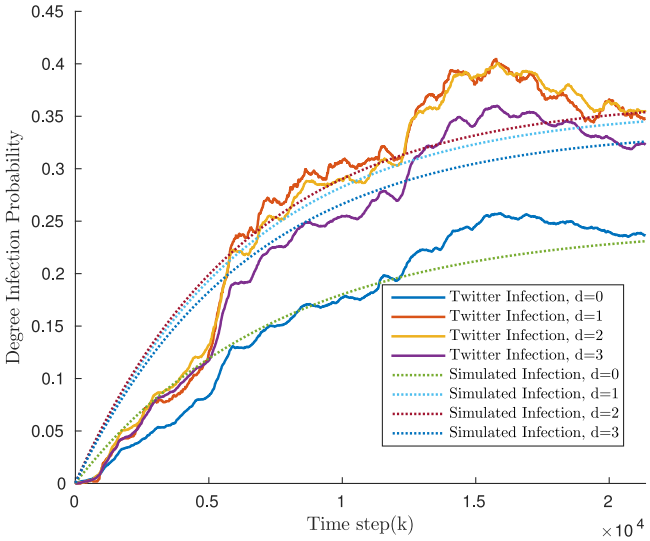


Fig. 7. The true Twitter infection and predicted Twitter infection using the deterministic mean field dynamics are compared for nodes of degree $d = 1, 2, 3$. The simulated trajectories use $\bar{P}_{12} = 0.001$ and the empirically generated ρ and $\hat{\bar{P}}_{21}$ in (41). Recall that without added state noise, the mean field dynamics are deterministic. The deterministic mean field dynamics satisfactorily capture the trajectory of the Twitter infection.

empirical transmission rates $\hat{\bar{P}}_{21}$ directly by observing the frequency with which an infected individual with d neighbors, a of which are infected, becomes infected over time horizon of length T time points. Here T is the time horizon of the data as stated in Section II-B.

$$\hat{\bar{P}}_{21}(d,a) = \times \frac{\sum_{n=0}^{T}\sum_{m=1}^{M}\mathcal{I}\left(s_{n+1}^{(m)}=1|s_n^{(m)}=2,\Xi^{(m)}=d,F_n^{(m)}=a\right)}{\sum_{n=0}^{T}\sum_{m=1}^{M}\mathcal{I}\left(s_n^{(m)}=2,\Xi^{(m)}=d,F_n^{(m)}=a\right)} \tag{41}$$

The true degree infection probabilities are computed directly from the entire network for each 1 minute time interval.

Next, we sample the data using the RDS sampling scheme described in Section II-C, every 1 minute and track the infected population states over time using the non-linear Bayesian Filtering technique described in Section III-A. The parameters used in the Bayesian filter are the empirically generated ρ, $\hat{\bar{P}}_{21}$ and all entries of $\bar{P}_{12} = 0.001$. The filter estimates are shown in

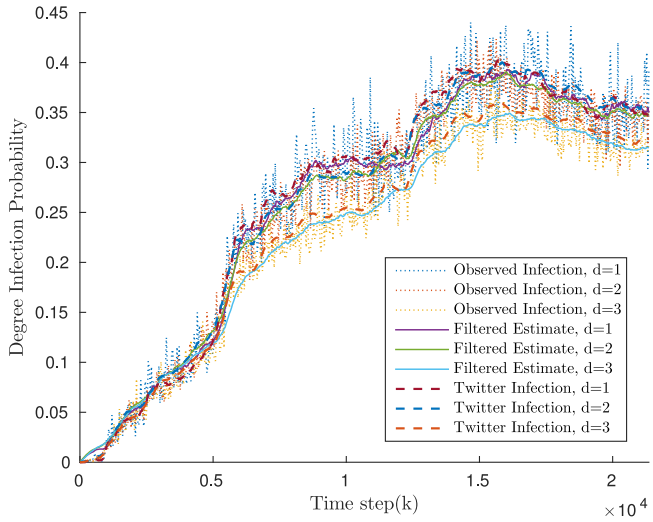


Fig. 8. The true Twitter infection and the filter estimates of the Twitter infected population state are compared. Samples are generated by RDS sampling on the #Jan25 social network. At each timestep a 10,000 node walk is performed and from this walk an observation of the infected population state is generated.

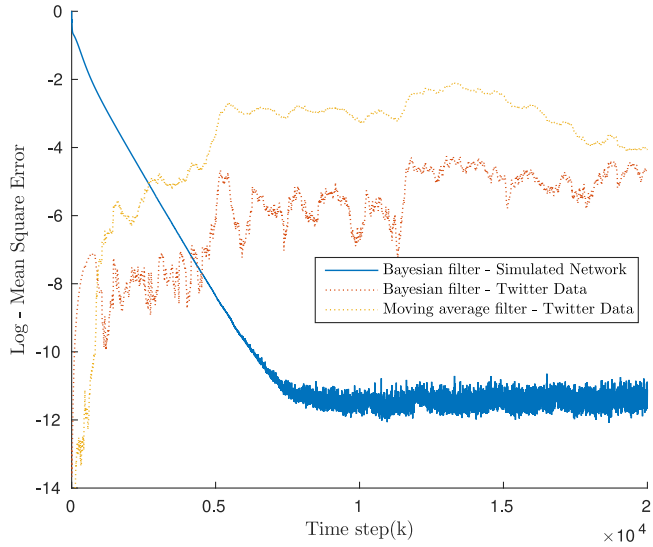


Fig. 9. The MSE of the filtered estimate of the true twitter infection probability state and the MSE of the estimate of the infected population state of a simulated network are shown. The MSE of the Twitter estimate starts extremely low. During early timesteps in the Twitter data, there are almost no infected nodes. The sampled observations similarly estimate nearly 0 infection, which is why the filtered infection is more accurate at these early timesteps. As the system evolves, the state observations become less accurate.

Fig. 8. It is seen that the filtered estimates satisfactorily track the true infected population states over time. Fig. 9 shows the mean square errors of the filter estimate of Twitter infection and the mean square errors of the filter estimate of the numerical example of Section V-D. This figure illustrates the superiority of the filter on the numerical example. The filter performs adequately for both the Twitter data and the simulated network, however the estimate for the simulated network, the infection of which follows the SIS model, performs dramatically better at a filtered estimate MSE of 10^{-11} versus the filtered estimate of the Twitter data around 10^{-6}.

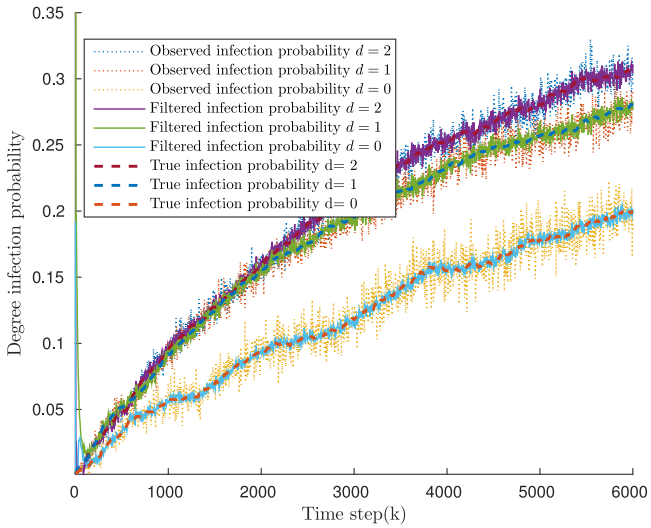


Fig. 10. Diffusion of infected population states and their corresponding filtered estimates in a scale-free network. The network was simulated, sampled and filtered according to 'Scheme B' of Section III-B. It can be seen that the estimates converge towards the true state for all degrees.

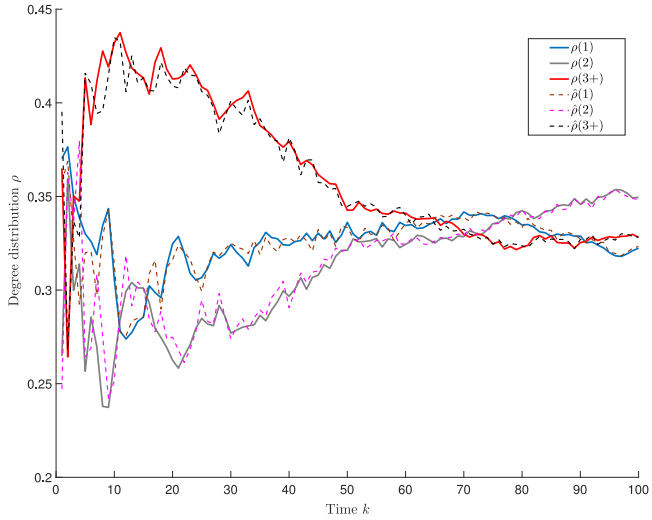


Fig. 11. Optimal filter performance on a network generated using the preferential attachment model (33), with $N^{(+)} = 3$ in (34). At each time k the true degree distribution evolves according to (34) and an infection is propagated according to 'Scheme B' of Section III-B on networks having 13000 nodes. The curves in the figure correspond to the true (ρ) and estimated ($\hat{\rho}$) degree distribution. The degree distribution is estimated using the procedure in Section IV-B.

D. Two-Time Scale Simulation

This section presents a numerical example that encompasses all the models used in this paper. It involves an evolving network (on a slow time scale) and diffusion of infection with the SIS model (on a fast time scale). For this simulation, the true state of the infection and underlying network are exactly known (in contrast with the Twitter data of Section V-C). This network is sampled to generate observations of the infected population state according to Section II-C, and then these observations are filtered using the Bayesian filter of Section III-A. The parameters of the network simulation are chosen to emulate the Twitter network explored in Section V-C.

1) Tracking the Infected Population State: Following *Scheme B* of Section III-B, we generated scale free networks with $M = 13000$ (number of nodes) to emulate the Twitter Social network. The infections were initialized by infecting nodes at time 0 with probability 0.01. This infection was propagated for 2×10^4 timesteps[31] with a healing probability $\bar{P}_{12} = 0.001$ and $\bar{P}_{21}$ generated empirically from Twitter data of Section V-C. At each timestep, 10^4 samples were obtained according to the uniform sampling described in Section II-C. The state and observation covariance matrices $\mathbf{Q}_n$ and $\mathbf{R}_n$ used in filtering were computed empirically from the true underlying state and observation data. The resulting observations and filtered estimates are shown in Fig. 10. The mean square error of the filter estimates are shown in Fig. 9 and compared to the filtered estimate for the Twitter data of Section V-C. The displayed mean square errors are the average of 50 independent simulations.

2) Tracking the Degree Distribution: In Section V-D1, we estimated the infected population state, where the infection dynamics evolved on the fast time scale. The slow time scale simulation is performed as follows: At each time k on the slow time scale, the infection dynamics is tracked using the filter discussed in Section III-A, for a duration of $n = 2 \times 10^4$ points. The length of the time window for tracking the infection dynamics is on the same order as the number of nodes in the network, as discussed in Theorem 1. The probability u in the degree evolution matrix for generating the true degree distribution is a function of the mode of the infected degree distribution as discussed in Section IV-B. We chose $u_k = \frac{1}{z_k+1}$ to capture this link. This was motivated by the rationale that higher the mode of infected degree distribution, smaller the u and hence higher is the probability of forming edges at the next time $k+1$. The mode $\hat{z}_k$ is observed in noise via the observation matrix B_k estimated according to the procedure given in Footnote 21 (Section IV-B). Fig. 11 displays the performance of the HMM filter for tracking the degree distribution of the co-evolving system.

[31] Recall from Theorem 1 that the duration is of the order of number of nodes.

VI. Conclusion

We considered the problem of tracking infection diffusion over large social networks by modeling the diffusion process using a SIS model. Using mean field dynamics, the evolution of infection has a generative model with polynomial dynamics. This was exploited to track the infection using a finite dimensional non-linear filter. Posterior Cramér-Rao lower bounds were computed for the mean field dynamics and it was shown that these bounds are relatively insensitive to the type of underlying social network (Erdős-Rényi vs Scale Free network). In large co-evolving networks modeled using a preferential attachment scheme, we provided a monotone comparative static result on the relation between the transition probabilities and the diffusion thresholds. A Bayesian filter to track the evolving degree distribution was also provided. The SIS model was then

extended to include homophily, and filtering on these networks was illustrated. Finally, a Twitter dataset was used to illustrate how infection diffusion can be modeled by a mean field dynamical SIS model, and we can filter and satisfactorily track the infection diffusion over time.

APPENDIX A
PROOF OF THEOREM 1 (MEAN FIELD DYNAMICS)

Part 1 of the theorem is a standard martingale representation of a Markov chain, see for example [22, page 20].

The proof of the mean field dynamics approximation in [23] is not readily accessible to an engineering reader. We show below that the proof is a simple consequence of Azuma-Hoeffding inequality and Gronwall's inequality; see [22].

Theorem 3 (Azuma-Hoeffding Inequality [45]): Suppose $S_T = \sum_{\tau=1}^{T} \zeta_\tau + S_0$ where $\{\zeta_\tau\}$ is a martingale difference process with bounded differences satisfying $|\zeta_\tau| \leq \Delta_\tau$ almost surely where Δ_τ are finite constants. Then for any $\epsilon > 0$,

$$\mathbb{P}(|S_T - S_0| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{\tau=1}^{T}\Delta_\tau^2}\right) \quad \blacksquare$$

A bound on the deviation between the mean field dynamics $\mathbf{x}_n$ in (12) and infected population state $\bar{\mathbf{x}}_n$ in (11) is evaluated in the form of two lemmas, namely, Lemma 2 and Lemma 3 below. Recall that ζ_τ is a $2D$-dimensional finite-state martingale increment process defined in (11) with $\|\zeta_\tau\|_2 \leq \frac{\Gamma}{M}$ for some positive constant Γ.

Lemma 2: Let $\varphi_n = x_n - \bar{x}_n$. Then $\|\varphi_n\|$ satisfies

$$\|\varphi_{n+1}\|_\infty \leq \|\varphi_0\|_\infty + \frac{\beta}{M}\sum_{\tau=1}^{n}\|\varphi_\tau\|_\infty + S_T.$$

where β is a positive constant (explicitly specified in (42) below).

Lemma 3: Let $S_T = \max_{1\leq n\leq T}\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty$. Then

$$\mathbb{P}\big(S_T \geq \epsilon\big) \leq 2\exp\left(-\frac{\epsilon^2 M^2}{2\Gamma T}\right)$$

Proof of Theorem 1: With Lemmas 2 and 3, the proof of Theorem 1 is as follows. Applying Gronwall's inequality[32] to Lemma 2 yields $\|\varphi_n\|_\infty \leq S_T \exp\left[\frac{\beta n}{M}\right]$, which in turn implies

$$\max_{1\leq n\leq T}\|\varphi_n\|_\infty \leq S_T \exp\left[\frac{\beta T}{M}\right].$$

As a result

$$\begin{aligned}\mathbb{P}(\max_{1\leq n\leq T}\|\varphi_n\|_\infty > \epsilon) &\leq \mathbb{P}\left(S_T \exp\left[\frac{\beta T}{M}\right] > \epsilon\right)\\ &= \mathbb{P}\left(S_T > \exp\left[-\frac{\beta T}{M}\right]\ \epsilon\right)\end{aligned}$$

[32]Gronwall's inequality: if $\{x_n\}$ and $\{b_n\}$ are non-negative sequences and $a \geq 0$, then

$$\bar{x}_n \leq a + \sum_{j=1}^{n-1} x_j b_j \Rightarrow x_n \leq a\exp\left(\sum_{j=1}^{n-1} b_j\right)$$

Next applying Lemma 3 to the right hand side yields

$$\mathbb{P}(\max_{1\leq n\leq T}\|\varphi_n\|_\infty > \epsilon) \leq 2\exp\left(-\exp\left(\frac{-2\beta T}{M}\right)\epsilon^2\frac{M^2}{2\Gamma T}\right)$$

Finally choosing $T = c_1 M$, for some positive constant c_1 yields

$$\mathbb{P}(\max_{1\leq n\leq T}\|\varphi_n\|_\infty > \epsilon) \leq 2\exp(-C_2\epsilon^2 M)$$

where $C_2 = \exp(-2\beta c_1)\frac{1}{2\Gamma c_1}$. $\blacksquare$

Proof of Lemma 2: Define the $2D-$ dimensional vectors:

$$\begin{aligned}\mathcal{H}(\mathbf{x}_n) &= \mathcal{P}_{21}(\mathbf{x}_n) - \mathcal{P}_{12}(\mathbf{x}_n),\\ \mathcal{H}(\bar{\mathbf{x}}_n) &= \mathcal{P}_{21}(\bar{\mathbf{x}}_n) - \mathcal{P}_{12}(\bar{\mathbf{x}}_n).\end{aligned}$$

Recall from (11) and (12),

$$\begin{aligned}\varphi_{n+1} &= \varphi_n + \frac{1}{M}[\mathcal{H}(\mathbf{x}_n) - \mathcal{H}(\bar{\mathbf{x}}_n)] + \zeta_n\\ &= \varphi_0 + \frac{1}{M}\sum_{\tau=1}^{n}[\mathcal{H}(\mathbf{x}_\tau) - \mathcal{H}(\bar{\mathbf{x}}_\tau)] + \sum_{\tau=1}^{n}\zeta_\tau\end{aligned}$$

$$\begin{aligned}\|\varphi_{n+1}\|_\infty &\leq \|\varphi_0\|_\infty + \frac{1}{M}\sum_{\tau=1}^{n}\|\mathcal{H}(\mathbf{x}_\tau)\\ &\quad - \mathcal{H}(\bar{\mathbf{x}}_\tau)\|_\infty + \|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty\\ &\leq \|\varphi_0\|_\infty + \frac{\beta}{M}\sum_{\tau=1}^{n}\|\varphi_\tau\|_\infty + S_T\end{aligned}$$

The last inequality is justified as follows: From (9), $P_{21}(d, \bar{x}_n(d)) = \rho(d)(1-\bar{x}_n(d))\bar{P}_{21}(d,a)$ and $P_{12}(d,\bar{x}_n(d)) = \rho(d)\bar{x}_n(d)\bar{P}_{12}(d,a)$ for some a, where a is the number of infected neighbors. Hence

$$\mathcal{H}(\mathbf{x}_n, i) - \mathcal{H}(\bar{\mathbf{x}}_n, i) \leq \beta(\bar{x}_n(i) - x_n(i)) \tag{42}$$

where $\beta = \max_d\big[\rho(d)(\bar{P}_{21}(d,a) + \bar{P}_{12}(d,a))\big]$ is bounded.

Proof of Lemma 3: $\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty = \max_i|\sum_{\tau=1}^{n}e_i'\zeta_\tau| = |\sum_{\tau=1}^{n}e_{i^*}'\zeta_\tau|$ for some i^*, where e_i denotes the vector having 1 at the i^{th} position and zero elsewhere. Since $e_{i^*}'\zeta_\tau$ is a martingale difference process with $|e_{i^*}'\zeta_\tau| \leq \sqrt{\Gamma}/M$ applying the Azuma-Hoeffding inequality (Theorem 3) yields

$$\begin{aligned}\mathbb{P}\left(\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty \geq \epsilon\right) &= \mathbb{P}\left(|\sum_{\tau=1}^{n}e_{i^*}'\zeta_\tau| \geq \epsilon\right)\\ &\leq 2\ \exp\left[-\frac{\epsilon^2 M^2}{2\Gamma n}\right]\end{aligned}$$

The right hand side is increasing with n. So clearly,

$$\mathbb{P}\left(\max_{1\leq n\leq T}\|\sum_{\tau=1}^{n}\zeta_\tau\|_\infty \geq \epsilon\right) \leq 2\ \exp\left[-\frac{\epsilon^2 M^2}{2\Gamma T}\right] \quad \blacksquare$$

APPENDIX B
PROOF OF THEOREM 2

A. Definitions: With $\mathbf{1}$ denoting the D-dimensional vector of ones, let $\Pi(D) \triangleq \{\rho \in \mathbb{R}^D : \mathbf{1}'\rho = 1, 0 \leq \rho(i) \leq 1,\ i \in$

$\{1, 2, \ldots, D\}\}$ denote the $D-1$ dimensional unit simplex comprised of probability vectors of dimension D.

Definition 2: First-Order Stochastic Dominance ($\geq_{sd}$): Let $\rho^1, \rho^2 \in \Pi(D)$ be any two probability vectors. Then $\rho^2 \geq_{\rm sd} \rho^1$ if

$$\sum_{i=j}^{D} \rho^2(i) \geq \sum_{i=j}^{D} \rho^1(i) \text{ for } j \in \{1, \ldots, D\}.$$

Let $\mathcal{V}$ denotes the space of D dimensional vectors ϑ, with non-decreasing components, i.e., $\vartheta(1) \leq \vartheta(2) \leq \cdots \vartheta(D)$. Then $\rho^2 \geq_{\rm sd} \rho^1$ iff for all $\vartheta \in \mathcal{V}$,

$$\vartheta' \rho^2 \geq \vartheta' \rho^1. \tag{43}$$

Definition 3: Second-Order Stochastic Dominance ($\geq_{\rm ssd}$): Let $\rho^1, \rho^2 \in \Pi(D)$ be two probability vectors with cumulative distribution functions F_1 and F_2. Then $\rho^1 \geq_{\rm ssd} \rho^2$ if

$$\sum_{j=1}^{i} F_1(j) \leq \sum_{j=1}^{i} F_2(j) \text{ for } i \in \{1, \ldots, D\}.$$

B. Proofs:

Theorem 4 ([1]): Consider two networks with degree distributions ρ^1 and ρ^2 respectively, where $\rho^1 \leq_{\rm ssd} \rho^2$. Then the diffusion thresholds satisfy $\theta^1_* > \theta^2_*$. ■

An important consequence of Theorem 4 is that as the number of nodes with higher degree increase, the probability of a large fraction of agents becoming infected increases.

Lemma 4: For any $u \in (0, 1)$, the transition matrix $L_k(u)$ in (33) satisfies

$$L^i_k(u) \leq_{sd} L^{i+1}_k(u) \;\; i = 1, 2, \ldots$$

where $L^i_k(u)$ denotes the i^{th} row of $L_k(u)$.

Lemma 5: Let $L_k(u_1)$ and $L_k(u_2)$ be two transition matrices with $u_i > 0$. If $u_1 > u_2$, then

$$L^i_k(u_2) \geq_{\rm sd} L^i_k(u_1)$$

where $L^i_k(u)$ denotes the i^{th} row of $L_k(u)$.

The proofs of Lemma 4 and 5 follow immediately since the matrix $L_k(u)$ is upper bidiagonal (34).

Lemma 6: (i) Let $L_k(u)$ be such that $L^i_k(u) \leq_{\rm sd} L^{i+1}_k(u)$ for $i = 1, 2, \ldots$, where $L^i_k(u)$ denotes the i^{th} row of $L_k(u)$. Then for probability vectors ρ^1 and ρ^2 with $\rho^1 \leq_{\rm sd} \rho^2$,

$$L'_k(u)\rho^1 \leq_{\rm sd} L'_k(u)\rho^2 \tag{44}$$

(ii) Let $u_1 > u_2$ and $L^i_k(u_2) \geq_{\rm sd} L^i_k(u_1)$ for $i = 1, 2, \cdots, D$. Then for any probability vector ρ,

$$L'_k(u_1)\rho \leq_{\rm sd} L'_k(u_2)\rho \tag{45}$$

Proof: From Definition 2, (44) is equivalent to

$$\sum_{i=1}^{D} \sum_{j \geq m} L^{\rm ij}_k(u)\rho^1(i) \leq \sum_{i=1}^{D} \sum_{j \geq m} L^{\rm ij}_k(u)\rho^2(i) \tag{46}$$

for $m = 1, \ldots, D$. Since $L^i_k(u) \leq_{\rm sd} L^{i+1}_k(u)$, it follows that $\sum_{j \geq m} L^{\rm ij}_k(u)$ is increasing in i. Then since $\rho^1 \leq_{\rm sd} \rho^2$, (43) yields (46).

Similarly, (45) is equivalent to

$$\sum_{i=1}^{D} \sum_{j \geq m} \left(L^{\rm ij}_k(u_1) - L^{\rm ij}_k(u_2)\right)\rho(i) \leq 0 \tag{47}$$

Since $L^i_k(u_2) \geq_{\rm sd} L^i_k(u_1)$, each term $L^{\rm ij}_k(u_1) - L^{\rm ij}_k(u_2) \leq 0$, thereby yielding (47).

Proof of Theorem 2:

1) Let $u_1 > u_2 > 0$. Assume by induction that at $k-1$, $\rho^{u_1}_{k-1} \leq_{\rm sd} \rho^{u_2}_{k-1}$. We have,

$$\begin{aligned} L'_k(u_1)\rho^{u_1}_{k-1} &\leq_{\rm sd} L'_k(u_1)\rho^{u_2}_{k-1} \;\; \text{from (44)} \\ L'_k(u_1)\rho^{u_2}_{k-1} &\leq_{\rm sd} L'_k(u_2)\rho^{u_2}_{k-1} \;\; \text{from (45)} \\ \Rightarrow L'_k(u_1)\rho^{u_1}_{k-1} &\leq_{\rm sd} L'_k(u_2)\rho^{u_2}_{k-1} \\ \Rightarrow \rho^{u_1}_k &\leq_{\rm sd} \rho^{u_2}_k \end{aligned}$$

2) Let $u_1 > u_2 > 0$. From the first part of Theorem 2 we have, $\rho^{u_1}_k \leq_{sd} \rho^{u_2}_k$. But $\rho^{u_1}_k \leq_{\rm sd} \rho^{u_2}_k$ implies $\rho^{u_1}_k \leq_{\rm ssd} \rho^{u_2}_k$. Then Theorem 4 implies $\theta^1_*(k) > \theta^2_*(k)$.

REFERENCES

[1] D. López-Pintado, "Diffusion in complex social networks," *Games Econ. Behav.*, vol. 62, no. 2, pp. 573–590, 2008.

[2] C. Chamley, *Rational Herds: Economic models of social learning*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[3] M. Granovetter, "Threshold models of collective behavior," *Amer. J. Sociol.*, vol. 83, pp. 1420–1443, 1978.

[4] N. Chen, "On the approximability of influence in social networks," *SIAM J. Discrete Math.*, vol. 23, no. 3, pp. 1400–1415, 2009.

[5] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1386–1396, May 1998.

[6] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 591–600.

[7] L. A. Adamic and B. A. Huberman, "Power-law distribution of the world wide web," *Science*, vol. 287, no. 5461, pp. 2115–2115, 2000.

[8] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Rev.*, vol. 42, no. 4, pp. 599–653, 2000.

[9] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[10] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton Univer. Press, 2008.

[11] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, 2001, Art. no. 3200.

[12] F. Vega-Redondo, *Complex Social Networks*, vol. 44. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[13] V. Krishnamurthy, O. N. Gharehshiran, and M. Hamdi, "Interactive sensing and decision making in social networks," *Found. Trends Signal Process.*, vol. 7, no. 1/2, pp. 1–196, 2014.

[14] M. A. Porter and J. P. Gleeson, *Dynamical Systems on Networks: A Tutorial*, vol. 4. New York, NY, USA: Springer-Verlag, 2016.

[15] M. Hamdi, V. Krishnamurthy, and G. Yin, "Tracking a Markov-modulated stationary degree distribution of a dynamic random graph," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6609–6625, Oct. 2014.

[16] L. Shi, "Kalman filtering over graphs: Theory and applications," *IEEE Trans. Autom. Control*, vol. 54, no. 9, pp. 2230–2234, Sep. 2009.

[17] M. O. Jackson and B. W. Rogers, "Relating network structure to diffusion properties through stochastic dominance," *BE J. Theor. Econ.*, vol. 7, no. 1, pp. 1–16, 2007.

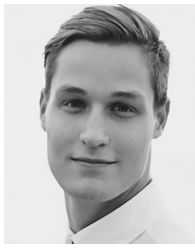
[18] G. Ghoshal, L. Chi, and A.-L. Barabási, "Uncovering the role of elementary processes in network evolution," *Sci. Rep.*, vol. 3, 2013, Art. no. 2920.

[19] L. Allen, "Some discrete-time SI, SIR, and SIS epidemic models," *Math. Biosci.*, vol. 124, no. 1, pp. 83–105, 1994.

[20] C. Castillo-Chavez and A.-A. Yakubu, "Discrete-time SIS models with complex dynamics," *Nonlinear Anal.: Theory, Methods Appl.*, vol. 47, no. 7, pp. 4753–4762, 2001.
[21] M. Taylor, "Exact and approximate epidemic models on networks: theory and applications," Ph.D. dissertation, University of Sussex, Brighton, U.K., 2013.
[22] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[23] M. Benaim and J. Weibull, "Deterministic approximation of stochastic evolution in games," *Econometrica*, vol. 71, no. 3, pp. 873–903, 2003.
[24] M. Granovetter, "Network sampling: Some first steps," *Amer. J. Sociol.*, vol. 81, pp. 1287–1303, 1976.
[25] P. J. Carrington, J. Scott, and S. Wasserman, *Models and Methods in Social Network Analysis*, vol. 28. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[26] N. K. Ahmed, J. Neville, and R. Kompella, "Network sampling: From static to streaming graphs," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 2, 2014, Art. no. 7.
[27] D. D. Heckathorn, "Respondent-driven sampling: A new approach to the study of hidden populations," *Soc. Probl.*, vol. 44, no. 2, pp. 174–199, 1997.
[28] D. D. Heckathorn, "Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations," *Soc. Probl.*, vol. 49, no. 1, pp. 11–34, 2002.
[29] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 631–636.
[30] S. Goel and M. J. Salganik, "Respondent-driven sampling as Markov chain Monte carlo," *Stat. Med.*, vol. 28, no. 17, pp. 2202–2229, 2009.
[31] M. Hern ández-González and M. V. Basin, "Discrete-time filtering for nonlinear polynomial systems over linear observations," *Int. J. Syst. Sci.*, vol. 45, no. 7, pp. 1461–1472, 2014.
[32] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*. Hoboken, NJ, USA: Wiley, 2004.
[33] F. Chung and L. Lu, *Complex Graphs and Networks*, vol. 107. Providence, RI, USA: American Mathematical Society, 2006.
[34] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos, "Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication," in *Proc. Eur. Conf. Principles Data Mining Knowl. Discovery*, 2005, pp. 133–145.
[35] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
[36] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the world-wide web," *Physica A: Statist. Mech. Appl.*, vol. 281, no. 1, pp. 69–77, 2000.
[37] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu.Rev. Sociol.*, vol. 27, pp. 415–444, 2001.
[38] H. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*. New York, NY, USA: Springer-Verlag, 2012.
[39] D. Topkis, *Supermodularity and Complementarity*. Princeton, NJ, USA: Princeton Univ. Press, 1998.
[40] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociol.*, vol. 27, no. 1, pp. 415–444, 2001.
[41] D. López-Pintado, *An Overview of Diffusion in Complex Networks*. New York, NY, USA: Springer-Verlag, 2016.
[42] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on Twitter," in *Proc. 7th Workshop Soc. Net. Mining. Anal.*, 2013, Paper 8.
[43] Z. Papacharissi and M. de F. Oliveira, "Affective news and networked publics: The rhythms of news storytelling on #Egypt," *J. Commun.*, vol. 62, no. 2, pp. 266–282, 2012.
[44] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proc. Nat. Acad. Sci.*, vol. 105, no. 41, pp. 15 649–15 653, 2008.
[45] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. London, U.K.: Oxford Univ. Press, 2013.

Vikram Krishnamurthy (F'05) received the Ph.D. degree from the Australian National University, Canberra, ACT, Austrailia, in 1992. He is currently a Professor in the School of Electrical and Computer Engineering, Cornell Tech, Cornell University, Ithaca, NY, USA. From 2002 to 2016, he was the Canada Research Chair Professor in statistical signal processing at the University of British Columbia, Vancouver, BC, Canada. His current research interests include statistical signal processing and stochastic control with applications in social networks. He served as a Distinguished Lecturer for the IEEE signal processing society and Editor-in-Chief of IEEE Journal Selected Topics in Signal Processing. He received a honorary doctorate from KTH (Royal Institute of Technology), Stockholm, Sweden in 2013.

Sujay Bhatt received the M.Tech. degree from the Indian Institute of Technology Bombay, Mumbai, India. He is currently working toward the Ph.D. degree at Cornell University, Ithaca, NY, USA. His current research interests include social learning, Bayesian learning over networks, statistical signal processing, stochastic control, and game theory.



Tavis Pedersen received the B.A.Sc. in engineering physics (with distinction) from the University of British Columbia, Vancouver, BC, Canada, in 2015, where he is currently working toward the M.A.Sc. degree in electrical engineering. His current research interests include statistical signal processing, mutations detection, mathematical modeling of contagion, and evolutionary game theory.