Journal of Complex Networks (2015) **3**, 584–605
doi:10.1093/comnet/cnv006
Advance Access publication on 31 March 2015

Role of subgraphs in epidemics over finite-size networks under the scaled SIS process

June Zhang* and José M. F. Moura†
Electrical and Computer Engineering Department, Carnegie Mellon University, USA
Corresponding author. Email: junez@andrew.cmu.edu

Edited by: Matt Keeling

[Received on 9 October 2014; accepted on 27 January 2015]

In previous work, we developed the scaled SIS process, which models the dynamics of SIS epidemics over networks. We derived for the scaled SIS process a closed-form expression for the time-asymptotic probability distribution of the configurations of all the agents in the network, which explicitly exhibits the underlying network topology through its adjacency matrix. This is accomplished for networks that are of finite-size and of arbitrary topology. This paper determines which network configuration is the most probable. We prove that, for a range of epidemic parameters, this combinatorial inference problem leads to a submodular optimization problem, which can be solved in polynomial time. We relate the most-probable configuration to the network structure, and in particular to the existence of high-density subgraphs. Depending on the model parameters, subset of agents may be more likely to be infected than others; these more vulnerable agents form subgraphs that are denser than the overall network. We illustrate our results with a 193 node social network of drug users and with the 4941 node Western US power grid under different model parameters.

Keywords: SIS epidemics; network processes; dense subgraphs; submodular optimization; most-probable configuration; network topology; graph density.

1. Introduction

A network is a graph; it is a collection of nodes connected by edges. Networks have been used in science and engineering to represent systems of multiple interconnected, interdependent components. As a result, the network structure has a large impact on the behaviour of the system. Quantifying how network structure impacts network function, that is, the behaviour of dynamical processes over networks, is a difficult problem since the system components do not behave independently.

In this paper, we focus on analysing the behaviour of network processes that are like epidemics. Analytical results for epidemics over networks have been obtained under particular conditions: full-mixing models (i.e. the underlying network is a complete graph); infinite-size network models using mean-field approximation; or for scaled-free networks [1–6]. These approaches approximate the underlying network topology with mathematically simpler structures, because accounting for the exact graph topology is a combinatorial problem that is difficult to analyse and computationally expensive

†José M.F. Moura was a visiting professor with New York University and the Center for Urban Science and Policy (CUSP) in 2013–2014; E-mail: moura@andrew.cmu.edu

to compute. In previous work, we proposed a dynamic network epidemics model over any arbitrary, finite-size network with N agents [7,8]. We call this the *scaled SIS* (susceptible-infected-susceptible) process. For this process, it is possible to characterize its time-asymptotic behavior (i.e., equilibrium distribution) without having to approximate the network structure.

The scaled SIS process is Markov. It accounts for (1) exogenous (i.e. spontaneous) infection at rate λ; (2) endogenous (i.e. neighbour dependent) infection at rate β; and (3) healing at rate μ. The time-asymptotic behaviour of the process is described by its equilibrium distribution, which is a PMF (probability mass function) over 2^N possible network configurations. Our approach preserves the full microscopic states of all the agents in contrast to previous approaches that only provide results for aggregate or macroscopic states (e.g. fraction of infected agents) [9]. However, retaining the exact network configuration means that the computational complexity of solving for the equilibrium distribution, an eigenvalue–eigenvector problem, scales exponentially with the size of the network, N.

We have shown that, under specific assumptions on the form of the endogenous infection, the scaled SIS process is a *reversible* Markov process for which we can find its equilibrium distribution in closed form, avoiding solving a large eigenvalue–eigenvector problem. Further, the equilibrium distribution that we derived exhibits explicitly the underlying network structure through the network adjacency matrix. The equilibrium distribution is parametrized by two parameters: $(\lambda/\mu, \beta)$. Parameter λ/μ controls the exogenous, or the topology-independent behaviour of the scaled SIS process, whereas parameter β controls the endogenous or the topology-dependent behaviour of the process.

We used the equilibrium distribution to address the question of which 2^N possible configurations is the most likely to occur in the long run. We refer to this as the most-probable configuration, which is found by maximizing the equilibrium distribution. This inference problem (called the Most-Probable Configuration Problem) is difficult because: (1) it is combinatorial; (2) it depends on the healing/infection parameters of the scaled SIS process; and (3) it depends on the underlying network topology. Previously in Zhang & Moura [7], we partitioned the space of $(\lambda/\mu, \beta)$ values into four regimes and were able to find the most-probable configuration in Regime II *Endogenous Infection Dominant*, for which $0 < \lambda/\mu \leqslant 1, \beta > 1$, for only specific types of networks: k-regular, complete multipartite and complete multipartite with k-regular islands. We showed for these specific networks that the most-probable configuration solution space exhibits phase transition behaviour, depending on the network structure and epidemic parameters.

This paper considers the Most-Probable Configuration Problem in Regime II *Endogenous Infection Dominant* for arbitrary networks. We are able to prove that this leads to the minimization of a submodular function, which can be solved in polynomial time. Further, we show the connection between the most-probable configuration and subgraphs in the network that are more vulnerable to epidemics. These are relevant questions in different applications. For example, these are the clusters to focus on in marketing campaigns or when combating epidemics.

We review the scaled SIS process in Section 2 and set up the Most-Probable Configuration Problem in Section 3. In Section 4, we show that, in Regime II, the Most-Probable Configuration Problem can be transformed into an equivalent submodular problem, and that it is possible to solve for its *exact* solution in *polynomial time*. We apply this to solve the most-probable configuration for two example networks: the 193 node acquaintance network of drug users in Hartford, CT [10] and the 4941 node network of the Western US power grid [11]. Section 5 shows how the solution space of the Most-Probable Configuration Problem in Regime II relates to the density of subgraphs in the network. Section 6 concludes the paper.

2. Scaled SIS process

Consider a population of N agents whose interconnections are represented by a static, simple, unweighted, undirected, connected graph, $G(V,E)$, where $V(G)$ is the set of vertices and $E(G)$ is the set of edges. For background on graphs, see West *et al.* [12]. The topology of G is captured by the symmetric $N \times N$ adjacency matrix, A. The state of the ith agent is denoted by x_i. Agents can be in one of two states: susceptible ($x_i = 0$) or infected ($x_i = 1$). Let

$$\mathbf{x} = [x_1, x_2, \ldots, x_N]^\top.$$

We will refer to x_i as the *agent state* and $\mathbf{x}$ as either the *network state* or the *configuration*. The configuration state space is $\mathcal{X} = \{\mathbf{x}\}$, with cardinality $|\mathcal{X}| = 2^N$.

The scaled SIS process models the evolution of the network state, $\mathbf{x}$, over time according to the stochastic microscopic interaction rules of the SIS epidemics. The SIS framework assumes that infected agents can heal and become reinfected so it does not account for immunization [4,13]. Let $X(t) = \mathbf{x}$ be the state of the network at time t, $t \geqslant 0$. The scaled SIS process accounts for (1) exogenous infection (i.e. susceptibles spontaneously develop infection); (2) endogenous infection (i.e. susceptibles become infected due to infection from infective neighbours); and (3) healing events. Infection and healing processes are independent and can not occur simultaneously. We also assume that only one agent in the network can heal or become infected at any given time instant. By including both exogenous infection and healing, the scaled SIS process does *not* have an absorbing state at equilibrium.

The scaled SIS process is Markov; each configuration is a state of the Markov process. On the configuration, $\mathbf{x} = [x_1, x_2, \ldots, x_i, \ldots, x_j, \ldots, x_N]^\top$, we define two operators. We use the following notation [16]:

$$H_i\mathbf{x} = [x_1, x_2, \ldots, x_i = 1, \ldots, x_N]^\top,$$
$$H_{j\bullet}\mathbf{x} = [x_1, x_2, \ldots, x_j = 0, \ldots, x_N]^\top.$$

The operator H_i defines the operation that agent i becomes infected. If agent i is already infected, the operator does nothing. The operator $H_{j\bullet}$ defines the operation that agent j is healed. If agent j is already uninfected, the operator does nothing.

The time the process spends in a particular state is random and exponentially distributed, with the following transition rates corresponding to infection and healing events, respectively:

(1) $X(t)$ transitions to the configuration where the ith agent, which was healthy, becomes infected with transition rate

$$q(\mathbf{x}, H_i\mathbf{x}) = \lambda\beta^{d_i}, \quad \mathbf{x} \neq H_i\mathbf{x}, \tag{1}$$

where $d_i = \sum_{j=1}^{N} \mathbb{1}(x_j = 1)A_{ij}$ is the number of infected neighbours of node i. The symbol $\mathbb{1}(\cdot)$ is the indicator function, and $A = [A_{ij}]$ is the adjacency matrix of the arbitrary network G that captures the interactions among the agents. There are two components to the infection rate. If the ith agent has no infected neighbours, $d_i = 0$, then the transition rate reduces to $\lambda > 0$. We interpret λ as the exogenous infection rate, the rate a susceptible agent spontaneously becomes infected; it is the same for all the agents in the network. If the ith agent has d_i infected neighbours, the infective rate is $\lambda\beta^{d_i}$; it is the product of λ and the endogenous infection rate, $\beta > 0$, *scaled* by d_i,

the number of infected neighbours of agent i. Because of this factor, the infective rate depends on the network topology.

(2) $X(t)$ transitions to the configuration where the jth agent, which was infected, heals with transition rate:

$$q(\mathbf{x}, H_{j\bullet}\mathbf{x}) = \mu, \quad \mathbf{x} \neq H_{j\bullet}\mathbf{x}. \tag{2}$$

The healing rate, $\mu > 0$, is the same for all the agents in the system.

2.1 *Scaled SIS process vs. contact process*

We note the main differences between the scaled SIS process and the basic contact process [15], which has been used to study SIS epidemics on networks [17,18]. First, the dynamics of the basic contact process do not consider exogenous infection. Therefore, the configuration where all the agents are healthy is an absorbing state of the Markov process; in this case, the equilibrium distribution is trivial. Secondly, and more importantly, the form of the infection rate of the scaled SIS process differs from the basic contact process. Using our notation and including exogenous infection, the infection rate of the basic contact process is

$$q(\mathbf{x}, H_i\mathbf{x}) = \lambda + d_i\beta, \quad \mathbf{x} \neq H_i\mathbf{x}.$$

In contrast to the infection rate of the scaled SIS process (see Equation (1)), the endogenous infection rate of agent i is linearly dependent on the number of infected neighbours, d_i, in the basic contact process instead of multiplicative as in the scaled SIS process. As we pointed out in Zhang & Moura [7], one advantage of the scaled SIS process is its tractability. We show in a forthcoming paper the relationship between the infection and healing rates of the scaled SIS process and the basic contact process with exogenous infection.

2.2 *Equilibrium distribution*

The evolution of the scaled SIS process is captured by the rate (infinitesimal) matrix $\mathbf{Q}$ of the Markov process $X(t)$. The assumption that the underlying network G is connected assures that the Markov process is irreducible. Therefore, the equilibrium distribution, $\pi(\mathbf{x})$, exists and is given by the left eigenvector corresponding to the 0 eigenvalue of $\mathbf{Q}$ [14]. The problem in determining the equilibrium distribution $\pi(\mathbf{x})$ is that its computation is prohibitively expensive for meaningful size networks since $\mathbf{Q}$ is a $2^N \times 2^N$ matrix. This has limited the analysis of epidemics and spreading processes on networks to either: (1) full-mixing models (e.g. where every agent comes in contact with every other agent—the network is a complete graph); (2) small scale simulations, where N is small so that $O((2^N)^3)$ operations are feasible; or (3) to mean-field type approximations of special network configurations.

We proved in Zhang & Moura [8], see also Zhang & Moura [7], that the scaled SIS process is a *reversible* Markov process by showing that its equilibrium distribution satisfies not only the global balance equation but also the detailed balance equation. For reversible Markov processes, the equilibrium distribution is unique [16]. We derived the equilibrium distribution of the scaled SIS process to be

$$\pi(\mathbf{x}) = \frac{1}{Z}\left(\frac{\lambda}{\mu}\right)^{1^\top \mathbf{x}} \beta^{\mathbf{x}^\top A\mathbf{x}/2}, \quad \mathbf{x}, \in \mathcal{X}, \tag{3}$$

where Z is the partition function,

$$Z = \sum_{\mathbf{x}\in\mathcal{X}} \left(\frac{\lambda}{\mu}\right)^{1^\top \mathbf{x}} \beta^{\mathbf{x}^\top A\mathbf{x}/2}. \tag{4}$$

Previous epidemics over network models call the ratio λ/μ, the *effective infection rate* [19]; this is also known as the *reproductive ratio* [20]. The equilibrium distribution, $\pi(\mathbf{x})$, factors as the product of three terms: (1) the normalization by the partition function; (2) the term $(\lambda/\mu)^{1^\top \mathbf{x}}$ that is topology independent since the exogenous infection rate λ and the healing rate μ are identical for all the agents in the network, and the total number of infected agents, $1^\top \mathbf{x}$, does not depend on the topology; and (3) the term $\beta^{\mathbf{x}^\top A\mathbf{x}/2}$ that explicitly accounts for the exact network through its adjacency matrix A. This term is topology dependent because the number of edges where both end nodes are infected (we call them *infected edges*), $\mathbf{x}^\top A\mathbf{x}/2$, explicitly depends on the underlying network.

2.3 *Parameter regimes*

The scaled SIS Process can model different types of network diffusion processes depending on if the effective exogenous infection rate, λ/μ, and the endogenous infection rate, β, are between 0 and 1, or if they are greater than 1. In Zhang & Moura [7], we identified four regimes. In this paper, we focus our analysis on Regime II *Endogenous Infection Dominant:* $0 < \lambda/\mu \leqslant 1, \beta > 1$. Regime II best models epidemics and similar types of spreading processes.

The effective exogenous infection rate, λ/μ, indicates the preference of individual agents. With $0 < \lambda/\mu \leqslant 1$, the healing rate is larger than the exogenous infection rate; agents prefer the healthy state to the infected state. With $\beta > 1$, however, additional infected neighbours increase the rate at which a healthy agent becomes infected. The network helps to spread the infection. As a result, network topology is crucial to determine the behaviour of the scaled SIS process at equilibrium.

In the next section, we introduce the Most-Probable Configuration Problem, which solves for the network configuration with the maximum equilibrium probability. Because there is *competition* between a topology-independent term and a topology-dependent term, the most-probable configuration exhibits complex phase transition behaviour depending on the effective exogenous infection rate λ/μ, the endogenous infection rate β, and the underlying network topology.

3. Most-probable configuration problem

In the previous section, we showed that, for the scaled SIS process, we are able to derive analytically its equilibrium distribution, $\pi(\mathbf{x})$ (see Equation (3)). The equilibrium distribution describes the long-run behaviour of the network epidemics. While the partition function (4) renders the exact calculation of the equilibrium distribution infeasible for meaningful size networks, knowing the equilibrium distribution expression allows us to quickly compare between network configurations, addressing, for example, questions like which of the two is more probable. Of all the possible 2^N network configurations, one is of particular interest, namely, the configuration of infected and healthy agents that has the highest chance of occurring in the long run. This is the configuration that maximizes $\pi(\mathbf{x})$. Formally, $\mathbf{x}^*$ maximizes the equilibrium probability:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x}) = \arg\max_{\mathbf{x}\in\mathcal{X}} \left(\frac{\lambda}{\mu}\right)^{1^\top \mathbf{x}} \beta^{\mathbf{x}^\top A\mathbf{x}/2}. \tag{5}$$

We call this the Most-Probable Configuration Problem and $\mathbf{x}^*$ the *most-probable configuration*. The Most-Probable Configuration Problem is a combinatorial optimization problem as agents can only be in one of two states: its solution is dependent on the effective exogenous infection rate, λ/μ, the endogenous infection rate β, and the underlying network topology, captured by the adjacency matrix, A.

Previously in Zhang & Moura [7], we provided analytical results for the Most-Probable Configuration Problem in Regime II *Endogenous Infection Dominant:* $0 < \lambda/\mu \leqslant 1, \beta > 1$ for particular networks, namely, structured network topologies such as k-regular, complete multipartite, complete multipartite with k-regular islands. We observed a phase transition behaviour. Below a threshold condition that depends on the parameters $(\lambda/\mu, \beta)$ and on the network topology, the most-probable configuration is $\mathbf{x}^0 = [0, 0, \ldots, 0]$, the configuration where all agents are susceptibles. Above the threshold condition, the most-probable configuration is $\mathbf{x}^N = [1, 1, \ldots, 1]$, the configuration where all agents are infected.

This paper extends the analysis of the Most-Probable Configuration Problem in Regime II to *arbitrary* network topology. We will show that, for arbitrary networks, the most-probable configuration may be configurations other than $\mathbf{x}^0$ and $\mathbf{x}^N$. We call these solutions to the Most-Probable Configuration Problem *non-degenerate configurations*. These solutions are useful for identifying agents and communities that are more vulnerable to the epidemics. We will relate these communities to the structure of the networks in detail later. Figures 1 and 2 show the most-probable configurations obtained by the method of Section 4 for two example networks: a 193 node acquaintance network [10] and the 4941 node power grid [11]. These are non-degenerate configurations where only a subset of agents are infected.

In Section 4, we prove that we can solve for the most-probable configuration in Regime II in polynomial time using submodular optimization. Then, in Section 5, we discuss the relationship between the most-probable configuration and the network topology, in particular, the relation between non-degenerate configurations and subgraphs in the network.

4. Submodularity and the most-probable configuration

In this section, we show that the Most-Probable Configuration Problem in Regime II can be transformed into a submodular function. First, we review the definition of submodular functions.

4.1 *Submodular function*

The Most-Probable Configuration Problem is the maximization of a pseudo-Boolean function. Pseudo-Boolean functions are functions that map N binary variables to a real number [21]. Minimization of general pseudo-Boolean functions is NP-hard [22]. Grötschel *et al.* [23] proved that the minimization of a pseudo-Boolean function that is submodular can be solved in polynomial time. If the function is supermodular, its maximization can be found in polynomial time.

A pseudo-Boolean function, $f : \{0, 1\}^N \to \mathcal{R}$, is also a set function $g : \mathcal{P}(V) \to \mathcal{R}$, where $\mathcal{P}(V)$ is the power set of $V = \{1, 2, \ldots, N\}$. There are many equivalent definitions of submodularity [24]. The one we use in this paper is the following:

Definition 4.1 ([21]) A set function, $g : \mathcal{P}(V) \to \mathcal{R}$, is submodular if and only if for any $\alpha_1 \subseteq V, \alpha_2 \subseteq \alpha_1, i \in V \setminus \alpha_1$

$$g(\alpha_1 \cup \{i\}) - g(\alpha_1) \leqslant g(\alpha_2 \cup \{i\}) - g(\alpha_2).$$

For a submodular function, the incremental gain of adding an element to the set α_1 is less than or equal to the gain of adding the element to a smaller subset of α_1. A supermodular function has the inequality in the opposite direction.

4.2 *Most-probable configuration: a submodular problem*

The Most-Probable Configuration Problem (5) seeks the maximum of a pseudo-Boolean function that maps a 0-1 vector, the network configuration $\mathbf{x}$, to a scalar. The network configuration $\mathbf{x} \in \{0,1\}^N$ is the characteristic vector or characteristic function of the set of infected agents: $\alpha_{\mathbf{x}} = \{i \mid i \in V, x_i = 1\}$. Let $h(\alpha_{\mathbf{x}})$ be the set of infected edges (i.e. edges where both end nodes are infected) in configuration $\mathbf{x}$: $h(\alpha_{\mathbf{x}}) = \{\{i,j\} \mid i,j \in V, A_{ij} = 1, x_i = 1, x_j = 1\}$.

The number of infected agents in configuration $\mathbf{x}$ is $|\alpha_{\mathbf{x}}| = 1^\top \mathbf{x}$. The number of infected edges is $|h(\alpha_{\mathbf{x}})| = \mathbf{x}^\top A\mathbf{x}/2$. The Most-Probable Configuration Problem is then to solve for the maximum argument of

$$g(\alpha_{\mathbf{x}}) = \left(\frac{\lambda}{\mu}\right)^{|\alpha_{\mathbf{x}}|} \beta^{|h(\alpha_{\mathbf{x}})|}. \tag{6}$$

We will prove in Theorem 4.3 that $-\log(g(\alpha_{\mathbf{x}}))$ is a submodular function. Therefore, we can solve for its minimum argument in polynomial time. Lemma 4.2 sets up some basic conditions that makes proving Theorem 4.3 easier.

LEMMA 4.2 Consider two sets of infected agents, $\alpha_1, \alpha_2 \subseteq V$ and $i \in V \setminus \alpha_1$. The cardinalities of α_1 and α_2 are $|\alpha_1| = n_1$ and $|\alpha_2| = n_2$, respectively; then $|\alpha_1 \cup \{i\}| = n_1 + 1$, and $|\alpha_2 \cup \{i\}| = n_2 + 1$. The numbers of infected edges induced by α_1 and α_2 are $|h(\alpha_1)| = e_1$ and $|h(\alpha_2)| = e_2$, respectively. Let $|h(\alpha_1 \cup \{i\})| = e_1 + m_1$ and $|h(\alpha_2 \cup \{i\})| = e_2 + m_2$; therefore, m_1 is the number of additional infected edges created with the inclusion of agent i in α_1 and m_2 is the number of additional infected edges created with the inclusion of agent i in α_2. Let $\alpha_2 \subseteq \alpha_1$. Then

(1) $n_1 \geqslant n_2$.

(2) $e_1 \geqslant e_2$.

(3) $m_1 \geqslant m_2$.

Proof. (1) When $\alpha_2 \subset \alpha_1$, α_2 must have strictly fewer infected agents than α_1. When $\alpha_2 = \alpha_1$, then they contain the same number of infected agents. Hence, $n_1 \geqslant n_2$.

(2) When $\alpha_2 \subset \alpha_1$, infected agents in α_2 can not induce more infected edges than the number of infected edges induced by the infected agents in α_1. When $\alpha_2 = \alpha_1$, then the infected agents in α_1 and α_2 will induce the same number of infected edges. Hence, $e_1 \geqslant e_2$.

(3) Every infected agent in α_2 is an infected agent in α_1. Every new infected edge that is induced when adding infected agent i to α_2 is also a new infected edge when adding infected agent i to α_1. Therefore, $m_1 \geqslant m_2$. □

THEOREM 4.3 Let $g(\alpha_{\mathbf{x}})$ be the set function given in (6). If $\lambda > 0, \mu > 0$ and $\beta \geqslant 1$, then $-\log(g(\alpha_{\mathbf{x}}))$ is a submodular function, where

$$-\log(g(\alpha_{\mathbf{x}})) = -|\alpha_{\mathbf{x}}| \log\left(\frac{\lambda}{\mu}\right) - |h(\alpha_{\mathbf{x}})| \log(\beta).$$

Proof. To prove submodularity of $-\log(g(\alpha_{\mathbf{x}}))$, we need to show that

$$-\log(g(\alpha_1 \cup \{i\})) + \log(g(\alpha_1)) \leqslant -\log(g(\alpha_2 \cup \{i\})) + \log(g(\alpha_2)), \tag{7}$$

for any $\alpha_1 \subseteq V, \alpha_2 \subseteq \alpha_1, i \in V \setminus \alpha_1$.

The left-hand side (LHS) of (7) is

$$-(n_1+1)\log\left(\frac{\lambda}{\mu}\right) - (e_1+m_1)\log(\beta) + n_1\log\left(\frac{\lambda}{\mu}\right) + e_1\log(\beta), \tag{8}$$

which reduces to

$$-\log\left(\frac{\lambda}{\mu}\right) - m_1\log(\beta). \tag{9}$$

The right-hand side (RHS) of (7) is

$$-(n_2+1)\log\left(\frac{\lambda}{\mu}\right) - (e_2+m_2)\log(\beta) + n_2\log\left(\frac{\lambda}{\mu}\right) + e_2\log(\beta), \tag{10}$$

which reduces to

$$-\log\left(\frac{\lambda}{\mu}\right) - m_2\log(\beta). \tag{11}$$

Expression (7) reduces to

$$-\log\left(\frac{\lambda}{\mu}\right) - m_1\log(\beta) \leqslant -\log\left(\frac{\lambda}{\mu}\right) - m_2\log(\beta).$$

Since $\beta \geqslant 1$, we know that $\log(\beta) \geqslant 0$ and that $m_1 \geqslant m_2$ by Lemma 4.2. Therefore, the LHS of (7) is less than or equal to the RHS of (7) for any $\alpha_1 \subseteq V, \alpha_2 \subseteq \alpha_1, i \in V \setminus \alpha_1$. By definition, $-\log(g(\alpha_{\mathbf{x}}))$ is a submodular function. □

Theorem 4.3 proves that $-\log(g(\alpha_{\mathbf{x}}))$ is submodular if $\lambda > 0, \mu > 0$ and $\beta \geqslant 1$; this means that $\log(g(\alpha_{\mathbf{x}}))$ is supermodular under the same condition. Since the logarithm function is a monotonic function, the maximum argument of $\log(g(\alpha_{\mathbf{x}}))$ is also the maximum argument of $g(\alpha_{\mathbf{x}})$, which is the solution to the Most-Probable Configuration Problem. As Regime II *Endogenous Infection Dominant*: $0 < \lambda/\mu \leqslant 1, \beta > 1$ satisfies the condition that $\beta \geqslant 1$, using submodular minimization, we can find the *exact* most-probable configuration of the scaled SIS process in Regime II for *arbitrary* network topology in polynomial time.

4.3 *Social networks and the power grid*

The most-probable configuration allows us to identify a set of agents that are more vulnerable to the network epidemics; agents who are infected in the most-probable configuration are more vulnerable to the epidemics than agents who remain healthy. Because the most-probable configuration is derived from a dynamical process, the set of vulnerable agents depends on the infection and healing rates, λ, β, μ.

As we showed in Zhang & Moura [7], the most-probable configuration changes depending on these parameters. When the healing rate dominates over the infection rates, $\mathbf{x}^* = \mathbf{x}^0$; this means that

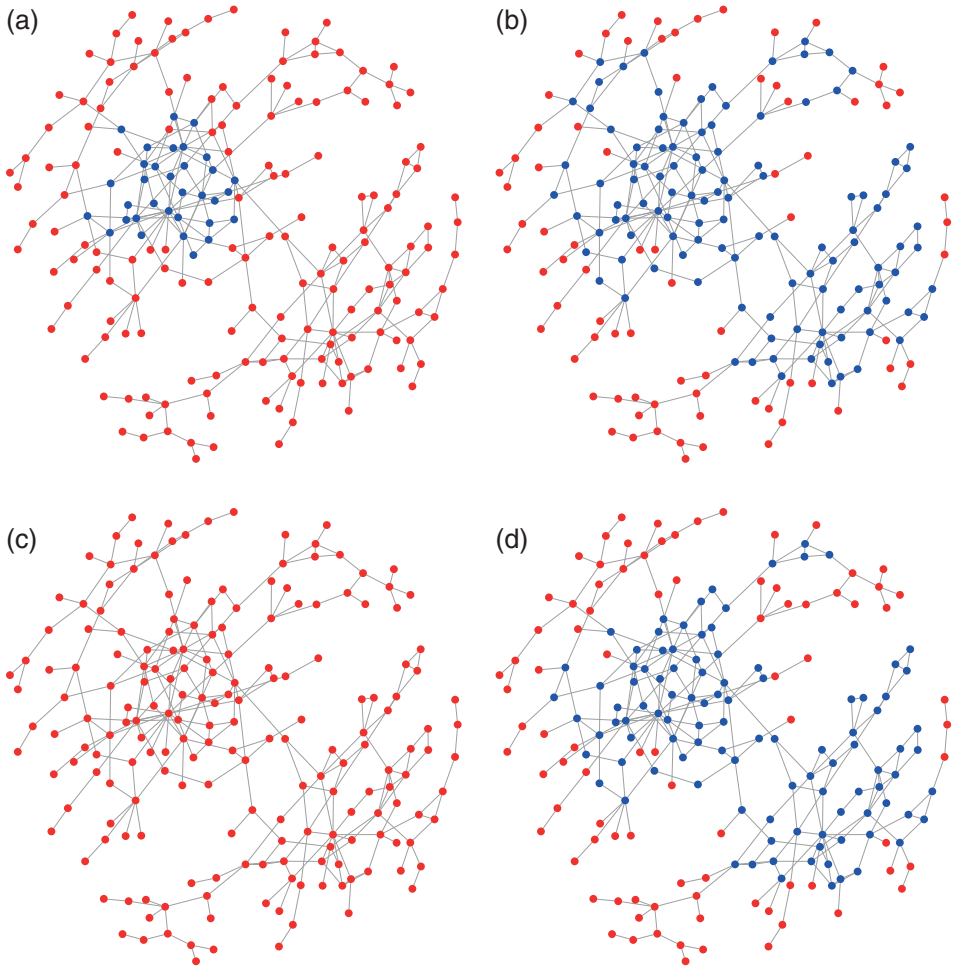

FIG. 1. Most-probable configuration $\mathbf{x}^*$ under different $(\lambda/\mu, \beta)$ parameters (Blue = Infected, Red = Healthy). (a) $\lambda/\mu = 0.2, \beta = 2.4$. (b) $\lambda/\mu = 0.267, \beta = 3$. (c) $\lambda/\mu = 0.4, \beta = 1.2$. (d) $\lambda/\mu = 0.5, \beta = 1.6$.

the epidemics is not severe. When the infection rates dominate over the healing rate, $\mathbf{x}^* = \mathbf{x}^N$; this means that the epidemics is severe. When $\mathbf{x}^*$ is a non-degenerate configuration (i.e. $\mathbf{x}^* \neq \mathbf{x}^0, \mathbf{x}^N$), this indicates that sets of agents in the network are more vulnerable than others to the epidemics. We illustrate this by solving for the most-probable configuration using [25] under different $(\lambda/\mu, \beta)$ parameters for two real-world networks: a social network [10] and the Western US power grid [11], obtained from [26]

The network shown in Fig. 1 is a 193 node, 273 edge social network of drug users in Hartford, CT. The network was determined through interviews. Reference [27] looked for influential agents in the network by considering it as a graph connectivity problem. However, they did not consider a dynamical model of influence. Assuming that we can model drug habits as an epidemics (i.e. there is a social contagion aspect to the behaviour), we applied the scaled SIS process to this

network and solved for the most-probable configuration under different parameters to find influential network structures.

We show the resultant most-probable configurations in Fig. 1(a–d) as we change $(\lambda/\mu, \beta)$. We can see from these results that there is a small community of users who are infected when others are healthy. The size of this community increases or decreases depending on the parameters. If there is a social contagion component to drug usage, then these agents may be more vulnerable to the social contagion component of drug usage and therefore more likely to persist in their habit.

The network shown in Fig. 2 is the 4941 node, 6595 edge power grid network of the Western United States used by Watts and Strogatz [11]. They showed through simulation of the SIR epidemic model on the western power grid that small-world networks like the western power grid are more conducive to spreading infection/failures than lattice networks. This is useful for explaining why failures propagate so quickly in a blackout. However, they can not identify *which* components in the power grid are more vulnerable to the epidemics with their approach. Here, we model the blackout as an SIS epidemics by assuming that failures and recoveries of grid components (e.g. power stations, substations, generators, switches, lines) are intermittent; a failed component may return to power, possibly failing again, as often happens in practice. Using the scaled SIS process, we can identify the most vulnerable substructures in the network.

Figure 2(a) and (b) show the most-probable configuration for the western US power grid when for the scaled SIS process parametrized at $(\lambda/\mu = 0.33, \beta = 2)$ and $(\lambda/\mu = 0.33, \beta = 2.6)$, respectively. We can see that for the same λ/μ, as β increases, thereby increasing the infectiousness of cascading failures (i.e. epidemics), the number of vulnerable components increases. This is intuitive since, for large β, the epidemics is severe, and the most-probable configuration is driven towards $\mathbf{x}^N$, the configuration where all the components are infected. Moreover, the most-probable configurations are both non-degenerate configurations. The components that are infected at equilibrium are more vulnerable to the cascading failures than components that remain healthy. By using submodular optimization, we can identify these more vulnerable components, by solving for the most-probable configuration out of 2^{4941} total possible configurations in polynomial time.

Using Matlab [25], on a desktop with 3.7 GHz Quad Core Xeon processor and 16 GB of RAM, the computation for the most-probable configuration for the 193 node network and parameters shown in Fig. 1(a–d) took 1.54, 0.96, 0.14 and 1.76 s, respectively, and for the 4941 node network and parameters shown in Fig. 2(a) and (b) took 1.29×10^4 s and 3.71×10^3 s, respectively. Although large, the increase in computation time from the 193 node network to the 4941 node network is much smaller than the corresponding increase (2^{4748} times) in the number of configurations. Additionally, computation time varies depending on the parameter values, most likely due to the complexity of the solution space of the optimization function.

An important question is to relate the most-probable configuration to network structure. We will show in the next section that the most-probable configuration is related to subgraph density by rewriting the equilibrium distribution (3) in term of induced subgraphs instead of network configurations.

5. Most-probable configuration: network structure

In the previous section, we showed that we can exactly solve for the most-probable configuration with a polynomial time algorithm. The exact solution, however, does not give insight on how the most-probable configuration depends on the network topology. In this section, we draw the connection between the most-probable configuration and subgraphs in the network. As per our intuition for

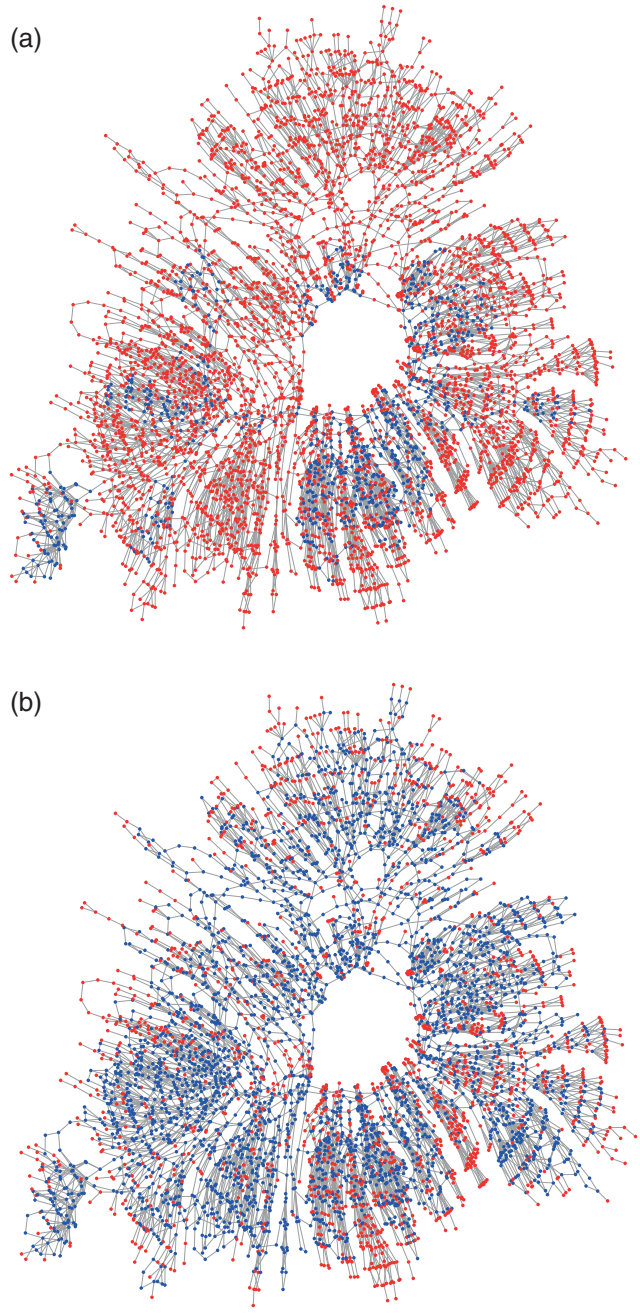

FIG. 2. Most-probable configuration $\mathbf{x}^*$ under different $(\lambda/\mu, \beta)$ parameters (Blue = Infected, Red = Healthy). (a) $\lambda/\mu = 0.33, \beta = 2$. (b) $\lambda/\mu = 0.33, \beta = 2.6$.

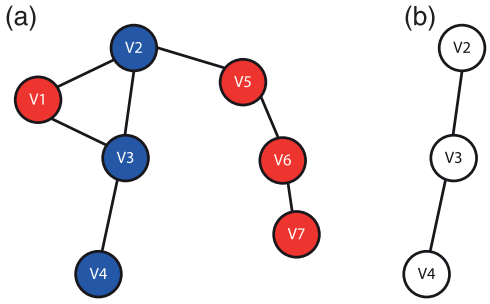

FIG. 3. (a) Configuration $\mathbf{x}_1 = [0, 1, 1, 1, 0, 0, 0]^\top$. (b) Induced Subgraph $H(\mathbf{x}_1) = H_1$.

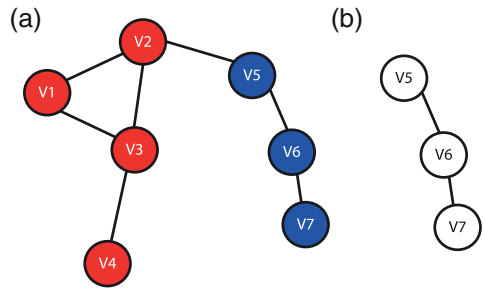

FIG. 4. (a) Configuration $\mathbf{x}_2 = [0, 0, 0, 0, 1, 1, 1]^\top$. (b) Induced Subgraph $H(\mathbf{x}_2) = H_2$.

epidemics, densely connected network structures are more vulnerable to network epidemics; the scaled SIS process quantifies this intuition. First, we will define the graph theoretic terms used in this section.

5.1 *Induced subgraphs and graph density*

DEFINITION 5.1 (From [28]) The graph H is an induced subgraph of G if two vertices in H are connected if and only if they are connected in G and the vertex set and edge set of H are subsets of the vertex set and edge set of G.

$$V(H) \subseteq V(G), E(H) \subseteq E(G).$$

DEFINITION 5.2 The graph $H(\mathbf{x})$ is an induced subgraph of configuration $\mathbf{x} = [x_1, x_2, \ldots, x_N]^\top$ if the nodes/edges in the subgraph are the infected agents/edges in $\mathbf{x}$.

$$V(H(\mathbf{x})) = \{v_i \in V(G) \mid x_i = 1\},$$
$$E(H(\mathbf{x})) = \{(i,j) \in E(G) \mid x_i = 1, x_j = 1\}.$$

By definition, $|V(H(\mathbf{x}))| = 1^\top \mathbf{x}$ and $|E(H(\mathbf{x}))| = \mathbf{x}^\top A\mathbf{x}/2$. Figures 3 and 4 show two network configurations and their corresponding induced subgraphs. We proved in Zhang & Moura [29] that configurations whose induced subgraphs are isomorphic are equally probable. Unless we need to refer explicitly to the underlying network configuration $\mathbf{x}$, for notational simplicity, we will write H to denote an induced subgraph instead of writing $H(\mathbf{x})$.

DEFINITION 5.3 The set of all possible induced subgraphs of G is $\mathcal{H} = \{H(\mathbf{x})\}$, $\forall \mathbf{x} \in \mathcal{X}$.

The set $\mathcal{H}$ includes the empty graph, which is induced by the configuration $\mathbf{x}^0 = [0, 0, \ldots, 0]^\top$, and G, which is the subgraph induced by the configuration $\mathbf{x}^N = [1, 1, \ldots, 1]^\top$.

DEFINITION 5.4 (From [30]) The density of graph G is

$$d(G) = \frac{|E(G)|}{|V(G)|}.$$

There is an alternative definition for graph density that is the number of edges divided by the total number of possible edges [31]. Unfortunately, these two definitions of density are not equivalent.

We will refer to the density of the entire network, $d(G) = d(H(\mathbf{x}^N))$, as the *network density*, and the density of an induced subgraph of G as the *subgraph density*. The density of the empty graph,

$d(H(\mathbf{x}^0))$, is 0 by definition. The subgraphs in $\mathcal{H}$ can be partially ordered by their density. There may be many subgraphs with the same density. A special induced subgraph in $\mathcal{H}$ is the densest subgraph.

DEFINITION 5.5 Let $\overline{H}$ be the densest subgraph in G. Then

$$d(\overline{H}) \geqslant d(H) \quad \forall H \in \mathcal{H}.$$

Finding $\overline{H}$ is known as the *Densest Subgraph Problem*. This problem can be solved in polynomial time exactly and in linear time in approximation for undirected graphs [30].

5.2 *Equilibrium distribution of the scaled SIS process*

Since there is a one-to-one relationship between the network configuration $\mathbf{x}$ and its induced subgraph $H(\mathbf{x})$, we can rewrite the equilibrium distribution (3) of the scaled SIS process in terms of the induced subgraph density and the size of the induced subgraph:

$$\pi(H) = \frac{1}{Z}\left(\left(\frac{\lambda}{\mu}\right)\beta^{d(H)}\right)^{|V(H)|}, \quad H \in \mathcal{H}, \tag{12}$$

where $d(H)$ is the density of the subgraph and Z is the partition function.

The Most-Probable Configuration Problem (5) is then also an optimization problem over all the possible induced subgraphs in G:

$$H(\mathbf{x}^*) = \arg\max_{H\in\mathcal{H}}\left(\left(\frac{\lambda}{\mu}\right)\beta^{d(H)}\right)^{|V(H)|}. \tag{13}$$

The subgraph induced by the most-probable configuration, $H(\mathbf{x}^*)$, is the *most-probable subgraph*, but this is *not* necessarily the same subgraph as the densest subgraph, $\overline{H}$.

Stating the equilibrium distribution in terms of the induced subgraph will allow us to derive several theorems regarding the most-probable configuration. For the theorems that follow, we make the following assumptions:

ASSUMPTION 1 The scaled SIS process operates in Regime II *Endogenous Infection Dominant*. This limits the effective exogenous infection and the endogenous infection rates to the range, $0 < \lambda/\mu \leqslant 1$ and $\beta > 1$.

ASSUMPTION 2 The underlying network G is a simple, undirected, unweighted and connected graph.

5.3 *Most-probable configuration and subgraphs*

THEOREM 5.6 (Proof in Appendix A) The most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^0$ if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\beta^{d(H)} > \mu$.

THEOREM 5.7 (Proof in Appendix B) The most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^N$ if and only if there exists at least one induced subgraph $H \in \mathcal{H} \setminus G$ with density $d(H) = E'/N'$ for which

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} < \frac{N'}{N}.$$

Combining Theorems 5.6 and 5.7, we can obtain the following corollary regarding the non-degenerate most-probable configurations.

COROLLARY 5.8 (Proof in Appendix C) Let the density of the network be $d(G) = E/N$. Then, the most-probable configuration is a non-degenerate configuration, $\mathbf{x}^* \in \mathcal{X} \setminus \{\mathbf{x}^0, \mathbf{x}^N\}$, if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H) = E'/N'$ for which $\lambda\beta^{d(H)} > \mu$, and

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} < \frac{N'}{N}.$$

In Regime II, individual agents have a preference for being healthy, but the epidemics might spread to other agents through neighbour-to-neighbour contagion. Under the scaled SIS process, the subgraph density $d(H)$ scales the exogenous infection rate β, thereby affecting the overall infection rate. Theorem 5.6 states that, if the network contains *dense-enough* subgraphs, then even when the effective exogenous infection rate, λ/μ, is small (i.e. $0 < \lambda/\mu \ll 1$), the exogenous infection rate, β, can leverage dense subgraphs to spread the infection throughout the network.

On the other hand, if the endogenous infection rate, β, is large (i.e. $\beta \gg 1$), then most certainly the epidemics will spread throughout the entire network and $\mathbf{x}^* = \mathbf{x}^N$. Theorem 5.7 states the condition when $\mathbf{x}^* \neq \mathbf{x}^N$. It also shows that it is important to consider if there are subgraphs in the network denser than the overall structure. Corollary 5.8 proves that the existence of the non-degenerate configurations is related to the existence of subgraphs with density larger than the network density. The existence of these *denser than G* subgraphs is crucial to the existence of non-degenerate configurations (i.e. different from $\mathbf{x}^0$ and $\mathbf{x}^N$) as solutions to the Most-Probable Configuration Problem; when the most-probable configuration is a non-degenerate configuration, agents belonging to denser subgraphs are more vulnerable to the epidemics.

In network science, dense clusters of agents have often been identified as either network *core* or *community* [32–34]. Solving for the non-degenerate configuration is an alternative method for determining these network structures. Previous works in core/community detection are algorithmic and do not consider dynamical processes on the network. The scaled SIS process, however, is a model for dynamical processes on networks and, therefore, what is considered a *community* changes depending on the parameters of the dynamical process: the most-probable configuration changes depending on the rates λ/μ and β.

We illustrate Theorems 5.6 and 5.7 with two small 16 node examples previously used in [7]: Network A in Fig. 5 and Network B in Fig. 6. For each network, we fix the effective exogenous infection rate, $\lambda/\mu = 0.5$. We then solve for the most-probable configuration for different β, ranging from 1.2 to 3. As the endogenous infection rate, β, changes, the most-probable configuration also changes. In Figs. 5(a) and 6(a), neither network supports dense-enough subgraphs for the epidemics to be severe. But as β increases, the infection starts to spread. In Network A, there is at least one subgraph denser than the network. The subgraph induced by $V1, V2, V3, V4, V5, V7, V8, V9, V10$ has a density of 1.33 whereas the density of the entire network is 1.19. In Fig. 5(b), the most-probable configuration has these nine agents infected while the other seven agents remain healthy. The nine agents in the dense subgraph are more vulnerable to the epidemics when $\lambda/\mu = 0.5$ and $\beta = 1.7$.

In Network B, there are at least two subgraphs denser than the network and they are induced by the set of infected agents of the most-probable configuration as shown in Fig. 6(b) and (c). We can see by solving for the most-probable configuration for different parameter values that, as the endogenous infection increases, the most-probable configuration goes towards $\mathbf{x}^N$ as all agents become vulnerable to the epidemics.

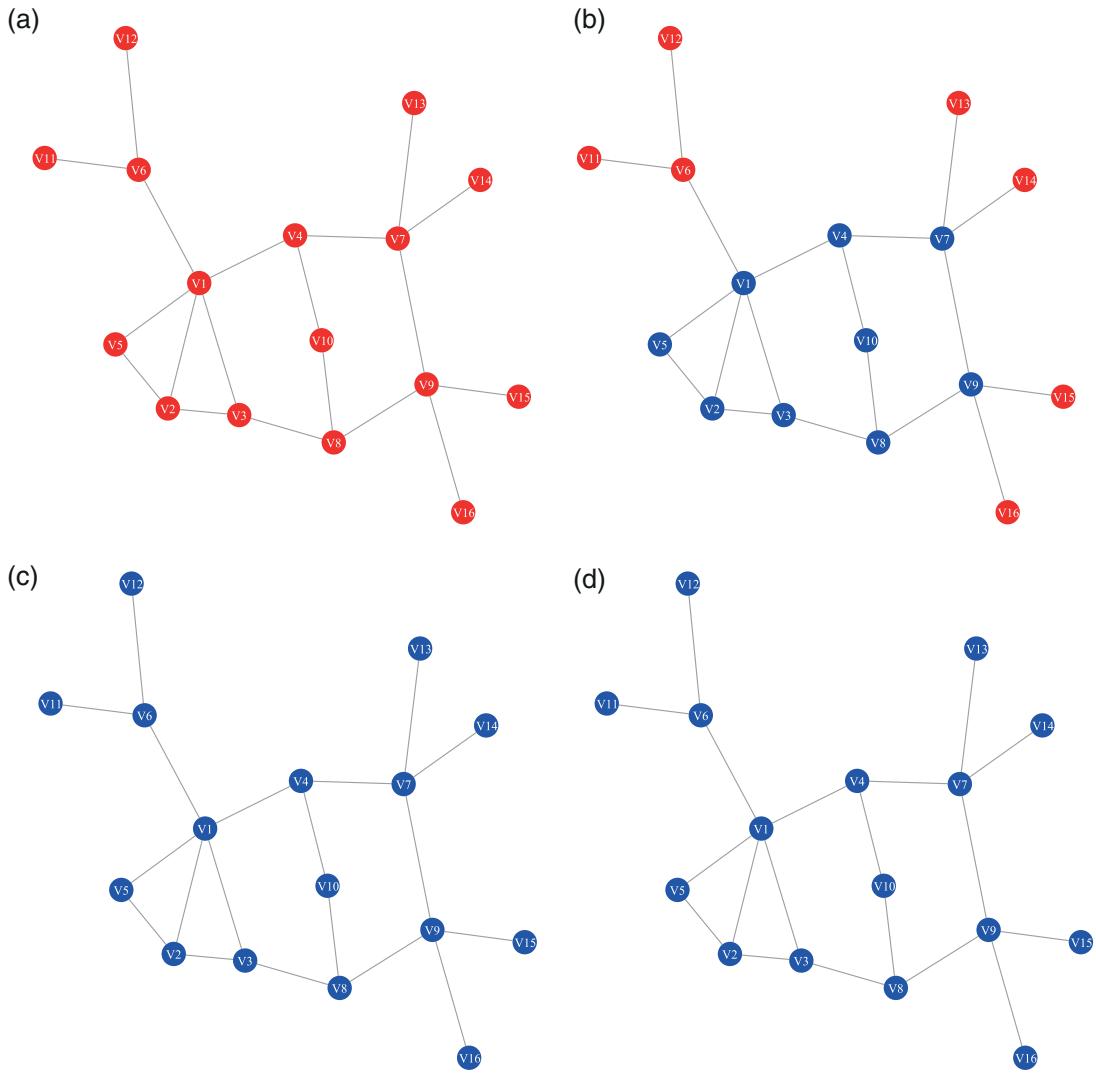

FIG. 5. Most-probable configuration $\mathbf{x}^*$ under different $(\lambda/\mu, \beta)$ parameters (Blue = Infected, Red = Healthy). (a) $\lambda/\mu = 0.5, \beta = 1.2, d(H(\mathbf{x}^*)) = 0$. (b) $\lambda/\mu = 0.5, \beta = 1.7, d(H(\mathbf{x}^*)) = 1.33$. (c) $\lambda/\mu = 0.5, \beta = 2. d(H(\mathbf{x}^*)) = 1.19$. (d) $\lambda/\mu = 0.5, \beta = 3, d(H(\mathbf{x}^*)) = 1.19$.

It is easier for the infection to spread in Network B than in Network A, since, for the same rate parameters, $\mathbf{x}^* = \mathbf{x}^N$ for Network B while $\mathbf{x}^* \neq \mathbf{x}^N$ for Network A. This is because Network B is a denser graph ($d(G) = 2.4375$) than Network A ($d(G) = 1.19$).

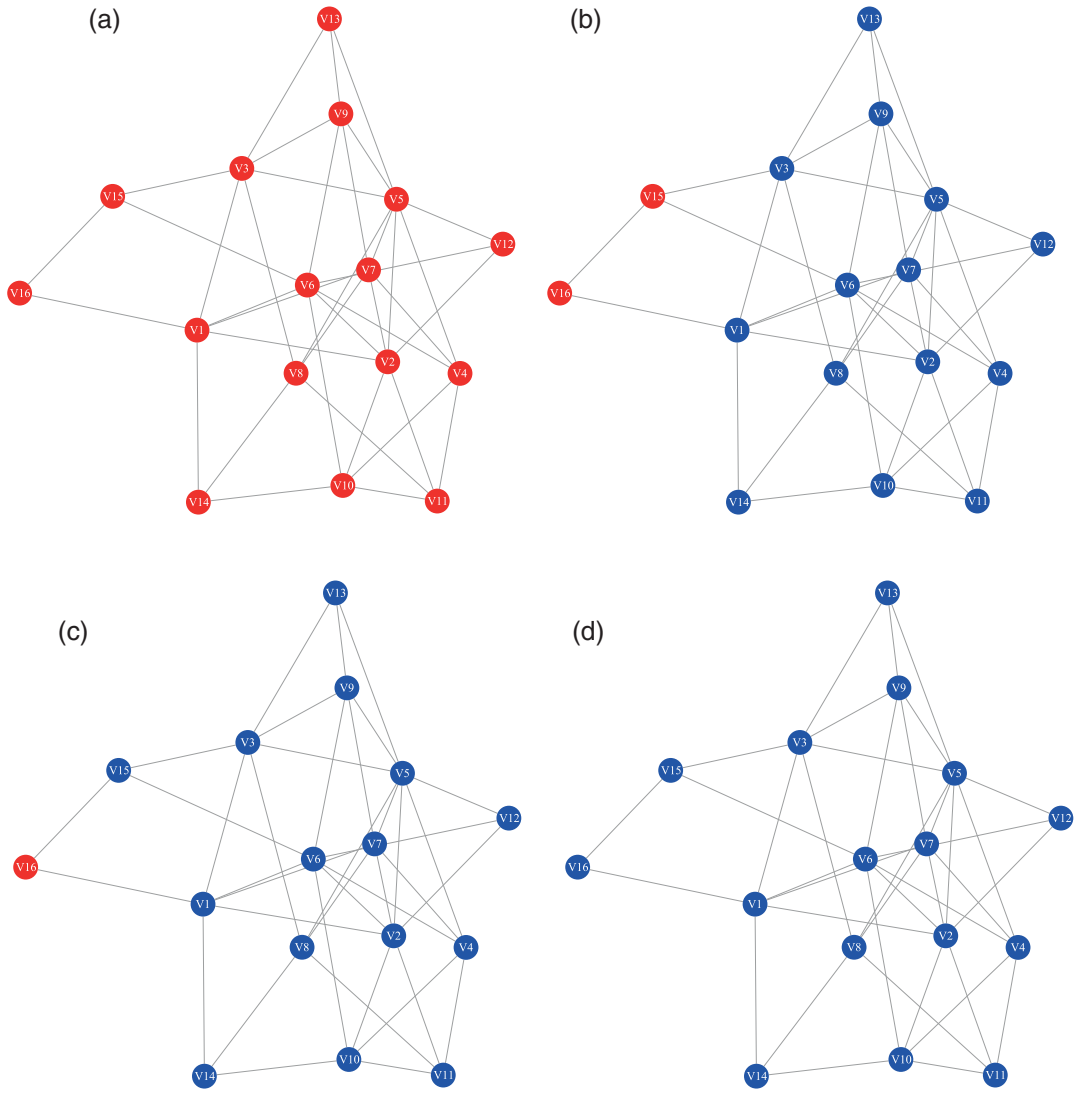

FIG. 6. Most-probable configuration $\mathbf{x}^*$ under different $(\lambda/\mu, \beta)$ parameters (Blue = Infected, Red = Healthy). (a) $\lambda/\mu = 0.5, \beta = 1.2$, $d(H(\mathbf{x}^*)) = 0$. (b) $\lambda/\mu = 0.5, \beta = 1.38$, $d(H(\mathbf{x}^*)) = 2.5$. (c) $\lambda/\mu = 0.5, \beta = 1.41$, $d(H(\mathbf{x}^*)) = 2.467$. (d) $\lambda/\mu = 0.5, \beta = 1.7$, $d(H(\mathbf{x}^*)) = 2.4375$.

5.4 *Most-probable configuration and the densest subgraph*

We showed that the most-probable configuration is related to the density of induced subgraphs in the network. The densest subgraph, $\overline{H}$, is a special induced subgraph. In this section, we focus specifically on the relationship between the most-probable configuration and the densest subgraph.

COROLLARY 5.9 (Proof in Appendix D) The most-probable configuration $\mathbf{x}^* = \mathbf{x}^0$ if and only if $\lambda\beta^{d(\overline{H})} \leqslant \mu$.

Corollary 5.9 follows the result of Theorem 5.6. If the densest subgraph in the network is not *dense enough* to overcome individual preferences for being healthy, then the endogenous infection rate β will not be able to drive the most-probable configuration away from $\mathbf{x}^0$.

Lastly, because of the connection between the most-probable configuration of the scaled SIS process and the densest subgraph, we can prove a *general* statement regarding network structure using results from dynamical processes on networks.

COROLLARY 5.10 (Proof in Appendix E) If G is a k-regular, complete multipartite or complete multipartite with k-regular islands network, then $\overline{H} = G$. That is, for these structured networks, the densest subgraph is the overall graph.

6. Conclusion

We introduced in the previous works the scaled SIS process, which is a mathematically analysable model for modelling diffusion processes on a static network [7]. The scaled SIS process is a reversible Markov process and has a closed-form equilibrium distribution that explicitly accounts for the underlying network topology via the adjacency matrix. It is controlled by two parameters: $(\lambda/\mu, \beta)$. The effective exogenous infection rate λ/μ controls the exogenous, or the topology-independent behaviour of the scaled SIS process whereas the exogenous infection rate β controls the endogenous or the topology-dependent behaviour of the process.

Depending on if the parameter values are in $(0, 1]$ or $(1, \infty)$, the scaled SIS process models qualitatively different network diffusion processes. In Regime II *Endogenous Infection Dominant:* $0 < \lambda/\mu \leqslant 1, \beta > 1$, the scaled SIS process best models a network epidemic process; individuals prefer to be healthy, while the network helps to spread the epidemics throughout the population.

This paper analyses the Most-Probable Configuration Problem, which solves for the network state with the maximum equilibrium probability, in Regime II for arbitrary networks. First, we proved that the Most-Probable Configuration Problem in Regime II is submodular. This means that we can compute the *exact* most-probable configuration in *polynomial time*. We use the most-probable configuration of the scaled SIS process to identify sets of vulnerable agents/components for a social network of drug users and the Western US power grid under different infection/healing rates.

We then showed that the most-probable configuration is dependent on certain classes of subgraphs in the networks. If there exist *dense-enough* subgraphs, conditioned on the right set of parameters, the most-probable configuration will shift away from $\mathbf{x}^0$, the network state where all the agents are healthy. However, if there exist subgraphs that are *denser than* the entire network, conditioned on the right range of infection and healing rates, the most-probable configuration may not reach $\mathbf{x}^N$, the network state with all agents infected. We call the solution of the Most-Probable Configuration Problem that is neither $\mathbf{x}^0$ nor $\mathbf{x}^N$, the non-degenerate configuration. Non-degenerate configurations identify subsets of agents that are more vulnerable to the network epidemics than others.

We also proved in this paper using results from Zhang & Moura [7] that structured networks such as k-regular, complete multipartite, complete multipartite with k-regular islands do not contain subgraphs that are denser than the overall network. Therefore, if we want to avoid subsets of agents being more vulnerable than others, we should use these types of structured networks. Our analysis of the scaled SIS process in Regime II informs us that network subgraph structures are important for understanding network diffusion processes. For future work, we are interested in statistically characterizing the subgraphs in network classes such as small-world networks and scaled-free networks.

Acknowledgement

We wish to thank Prof. João P. Costeira and Prof. João M.F. Xavier of the Department of Electrical and Computer Engineering at Instituto Superior Técnico, Lisbon, Portugal, for discussions regarding submodular optimization.

Funding

This work was partially supported by the Air Force Office of Scientific Research (FA95501010291) and by the National Science Foundation (CCF1011903, CCF1018509).

References

1. Pastor-Satorras, R. & Vespignani, A. (2002) Epidemic dynamics in finite size scale-free networks. *Phys. Rev. E*, **65**, 035108.
2. Keeling, M. J. & Eames, K. T. (2005) Networks and epidemic models. *J. R. Soc. Interface*, **2**, 295–307.
3. De Souza, D. R. & Tomé, T. (2010) Stochastic lattice gas model describing the dynamics of the SIRS epidemic process. *Phys. A: Stat. Mech. Appl.*, **389**, 1142–1150.
4. Newman, M. (2010) *Networks: An Introduction*. Oxford: Oxford University Press.
5. Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O. & Vernon, M. C. (2011) Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on Infectious Diseases.* doi:10.1155/2011/284909.
6. House, T. & Keeling, M. J. (2011) Insights from unifying modern approximations to infections on networks. *J. R. Soc. Interface*, **8**, 67–73.
7. Zhang, J. & Moura, J. M. F. (2014) Diffusion in social networks as SIS epidemics: beyond full mixing and complete graphs. *IEEE J. Sel. Top. Signal Process.*, **8**, 537–551. doi: 10.1109/JSTSP.2014.2314858.
8. Zhang, J. & Moura, J. M. F. (2013) Threshold behavior of epidemics in regular networks. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5411–5414. doi: 10.1109/ICASSP.2013.6638697.
9. Wang, Y., Chakrabarti, D., Wang, C. & Faloutsos, C. (2003) Epidemic spreading in real networks: an eigenvalue viewpoint. *Proceedings of the International Symposium on Reliable Distributed Systems*, Florence, Italy, pp. 25–34. doi: 10.1109/RELDIS.2003.1238052.
10. Weeks, M. R., Clair, S., Borgatti, S. P., Radda, K. & Schensul, J. J. (2002) Social networks of drug users in high-risk sites: finding the connections. *AIDS Behav.*, **6**, 193–206.
11. Watts, D. J. & Strogatz, S. H. (1998) Collective dynamics of small-world networks. *Nature*, **393**, 440–442.
12. West, D. B. *et al.* (2001) *Introduction to Graph Theory*, vol. 2. Upper Saddle River: Prentice Hall.
13. Ross, R. (1915) Some a priori pathometric equations. *Br. Med. J.*, **1**, 546.
14. Norris, J. R. (1998) *Markov Chains*. Cambridge: Cambridge University Press.
15. Liggett, T. M. (1999) *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, vol. 324. Berlin: Springer.
16. Kelly, F. P. (2011) *Reversibility and Stochastic Networks*. Cambridge: Cambridge University Press.
17. Draief, M., Ganesh, A. & Massoulié, L. (2006) Thresholds for virus spread on networks. *Proceedings of the International Conference on Performance Evaluation Methodologies and Tools*, Pisa, Italy. New York, NY, USA: ACM, p. 51.
18. Ganesh, A., Massoulie, L. & Towsley, D. (2005) The effect of network topology on the spread of epidemics. *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, Miami, USA, pp. 1455–1466.
19. Bonaccorsi, S., Ottaviano, S., De Pellegrini, F., Socievole, A. & Van Mieghem, P. (2014) Epidemic outbreaks in two-scale community networks. *Phys. Rev. E*, **90**, 012810.

20. Ball, F. (1999) Stochastic and deterministic models for SIS epidemics among a population partitioned into households. *Math. Biosci.*, **156**, 41–67.
21. Billionnet, A. & Minoux, M. (1985) Maximizing a supermodular pseudo-Boolean function: a polynomial algorithm for supermodular cubic functions. *Discrete Appl. Math.*, **12**, 1–11.
22. Boros, E. & Hammer, P. L. (2002) Pseudo-Boolean optimization. *Discrete Appl. Math.*, **123**, 155–225.
23. Grötschel, M., Lovász, L. & Schrijver, A. (1981) The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, **1**, 169–197.
24. Lovász, L. (1983) Submodular functions and convexity. *Mathematical Programming The State of the Art*. Berlin: Springer, pp. 235–257.
25. Krause, A. (2010) SFO: A toolbox for submodular function optimization. *J. Mach. Learn. Res.*, **11**, 1141–1144.
26. Leskovec, J. & Krevl, A. (2003) SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
27. Borgatti, S. P. (2003) The key player problem. Dynamic social network modeling and analysis: Workshop summary and papers, p. 241.
28. Godsil, C. G. R. (2001) *Algebraic Graph Theory*. Berlin: Springer.
29. Zhang, J. & Moura, J. M. F. (2014) Subgraph density and epidemics over networks. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1125–1129. doi: 10.1109/ICASSP.2014.6853772.
30. Khuller, S. & Saha, B. (2009) On finding dense subgraphs. *Automata, Languages and Programming*. Berlin: Springer, pp. 597–608.
31. Wasserman, S. (1994) *Social Network Analysis: Methods and Applications*, vol. 8. Cambridge: Cambridge University Press.
32. Csermely, P., London, A., Wu, L. Y. & Uzzi, B. (2013) Structure and dynamics of core/periphery networks. *J. Complex Netw.*, **1**, 93–123.
33. Borgatti, S. P. & Everett, M. G. (2000) Models of core/periphery structures. *Soc. Netw.*, **21**, 375–395.
34. Brandes, U., Pfeffer, J. & Mergel, I. (2012) *Studying Social Networks: A Guide to Empirical Research*. Campus.

Appendix A. Proof for Theorem 5.6

Theorem The most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^0$ if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\beta^{d(H)} > \mu$.

Proof. *Sufficiency*: If there exists at least one subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\beta^{d(H)} > \mu$, then $\mathbf{x}^* \neq \mathbf{x}^0$.

Using the equilibrium distribution (3), $\pi(\mathbf{x}^0) = 1/Z$. Let the subgraph $H \in \mathcal{H}$ be the subgraph induced by configuration $\mathbf{x}' \in \mathcal{X} \setminus \mathbf{x}^0$. The number of infected agents in configuration $\mathbf{x}'$ is $1^\top \mathbf{x}' = |V(H)| > 0$. Using (12), its equilibrium probability is

$$\pi(\mathbf{x}') = \pi(H) = \frac{1}{Z}\left(\left(\frac{\lambda}{\mu}\right)\beta^{d(H)}\right)^{|V(H)|}.$$

If $(\lambda/\mu)\beta^{d(H)} > 1$, we know that $\pi(\mathbf{x}') > \pi(\mathbf{x}^0)$. Therefore, $\mathbf{x}^0$ can not be the most-probable configuration.

Necessity: If $\mathbf{x}^* \neq \mathbf{x}^0$, then there exists at least one subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\beta^{d(H)} > \mu$.

If $\mathbf{x}^* \neq \mathbf{x}^0$, this means that there is some configuration $\mathbf{x}'$ for which $\pi(\mathbf{x}') > \pi(\mathbf{x}^0)$. We know that $\pi(\mathbf{x}^0) = 1/Z$. Using the equilibrium distribution in (12) and the fact that $1^\top \mathbf{x} = |V(H)| > 0, \forall \mathbf{x} \in \mathcal{X} \setminus \mathbf{x}^0$,

we can conclude that there must exist some induced subgraph whose density satisfies this condition $(\lambda/\mu)\beta^{d(H(\mathbf{x}'))} > 1$. □

Appendix B. Proof for Theorem 5.7

THEOREM The most-probable configuration $\mathbf{x}^* \neq \mathbf{x}^N$ if and only if there exists at least one induced subgraph $H \in \mathcal{H} \setminus G$ with density $d(H) = E'/N'$ for which

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} < \frac{N'}{N}.$$

Proof. *Sufficiency:* If there exists at least one induced subgraph $H \in \mathcal{H} \setminus G$ with density $d(H) = E'/N'$ such that $\log((\lambda/\mu)\beta^{d(G)})/\log((\lambda/\mu)\beta^{d(H)}) < N'/N$, then $\mathbf{x}^* \neq \mathbf{x}^N$.

The subgraph H is induced by the configuration $\mathbf{x}' \in \mathcal{X}$. The log equilibrium probabilities according to (12) for $\mathbf{x}'$ and $\mathbf{x}^N$, respectively, are

$$\log(\pi(\mathbf{x}')) = \log\left(\frac{1}{Z}\right) + N' \log\left(\frac{\lambda}{\mu}\beta^{d(H)}\right)$$

and

$$\log(\pi(\mathbf{x}^N)) = \log\left(\frac{1}{Z}\right) + N \log\left(\frac{\lambda}{\mu}\beta^{d(G)}\right).$$

Condition $\log((\lambda/\mu)\beta^{d(G)})/\log((\lambda/\mu)\beta^{d(H)}) < N'/N$ implies that $N\log((\lambda/\mu)\beta^{d(G)}) < N'\log((\lambda/\mu)\beta^{d(H)})$. Therefore, $\log(\pi(\mathbf{x}')) > \log(\pi(\mathbf{x}^N))$. Since the logarithm is a monotonic function, we can conclude that $\mathbf{x}^* \neq \mathbf{x}^N$.

Necessity: If $\mathbf{x}^* \neq \mathbf{x}^N$, then there exists at least one induced subgraph $H \in \mathcal{H}$ such that $\log((\lambda/\mu)\beta^{d(G)})/\log((\lambda/\mu)\beta^{d(H)}) < N'/N$.

Let $\mathbf{x}^* = \mathbf{x}'$, which induces a subgraph $H \in \mathcal{H}$ with density $d(H)$. Using (12),

$$\pi(\mathbf{x}') = \log\left(\frac{1}{Z}\right) + N' \log\left(\frac{\lambda}{\mu}\beta^{d(H)}\right),$$
$$\pi(\mathbf{x}^N) = \log\left(\frac{1}{Z}\right) + N \log\left(\frac{\lambda}{\mu}\beta^{d(G)}\right).$$

Since $\mathbf{x}'$ is the most-probable configuration, this means $\pi(\mathbf{x}') - \pi(\mathbf{x}^N) > 0$, which implies

$$N' \log\left(\frac{\lambda}{\mu}\beta^{d(H)}\right) - N \log\left(\frac{\lambda}{\mu}\beta^{d(G)}\right) > 0.$$

This reduces to the condition that

$$\frac{\log\left((\lambda/\mu)\beta^{d(G)}\right)}{\log\left((\lambda/\mu)\beta^{d(H)}\right)} < \frac{N'}{N}.$$

□

Appendix C. Proof for Corollary 5.8

COROLLARY Let the density of the network be $d(G) = E/N$. Then, the most-probable configuration is a non-degenerate configuration, $\mathbf{x}^* \in \mathcal{X} \setminus \{\mathbf{x}^0, \mathbf{x}^N\}$, if and only if there exists at least one induced subgraph $H \in \mathcal{H}$ with density $d(H) = E'/N'$ for which $\lambda\beta^{d(H)} > \mu$, and

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} < \frac{N'}{N}.$$

Proof. We want to determine the necessary and sufficient conditions such that $\mathbf{x}^* = \mathbf{x}'$, which induces subgraph H, such that we have both $\mathbf{x}' \neq \mathbf{x}^0$ and $\mathbf{x}' \neq \mathbf{x}^N$. This is equivalent to showing

$$\pi(\mathbf{x}') > \pi(\mathbf{x}^0) \tag{14}$$

and

$$\pi(\mathbf{x}') > \pi(\mathbf{x}^N). \tag{15}$$

Condition (14) holds if and only if $\lambda\beta^{d(H)} > \mu$ by Theorem 5.6. Condition (15) holds if and only if

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} < \frac{N'}{N},$$

by Theorem 5.7. This proves the corollary. □

Appendix D. Proof for Corollary 5.9

COROLLARY The most-probable configuration $\mathbf{x}^* = \mathbf{x}^0$ if and only if $\lambda\beta^{d(\overline{H})} \leqslant \mu$.

Proof. *Sufficiency:* If $\lambda\beta^{d(\overline{H})} \leqslant \mu$, then $\mathbf{x}^* = \mathbf{x}^0$.

Recall the definition of the densest subgraph Definition 5.5. With $\beta > 1$, $\lambda\beta^{d(H(\mathbf{x}))} \leqslant \lambda\beta^{d(\overline{H}(\mathbf{x}))} \leqslant \mu$ for all possible induced subgraphs in G. This means that there is no subgraph, $H \in \mathcal{H}$, for which $\lambda\beta^{d(H)} > \mu$. We can conclude that $\mathbf{x}^* = \mathbf{x}^0$ using the contrapositive of Theorem 5.6: If there is no subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\beta^{d(H)} > \mu$, then $\mathbf{x}^* = \mathbf{x}^0$.

Necessity: If $\mathbf{x}^* = \mathbf{x}^0$, then $\lambda\beta^{d(\overline{H})} \leqslant \mu$.

The result follows from the contrapositive of Theorem 5.6: If $\mathbf{x}^* = \mathbf{x}^0$, then there is no subgraph $H \in \mathcal{H}$ with density $d(H)$ for which $\lambda\beta^{d(H)} > \mu$. Therefore, all induced subgraphs, including the densest subgraph have density for which $\lambda\beta^{d(H)} \leqslant \mu$. □

Appendix E. Proof for Corollary 5.10

LEMMA If G is a k-regular, complete multipartite or complete multipartite with k-regular islands network, then $\overline{H} = G$. That is, for these structured networks, the densest subgraph is the overall graph.

Proof. We proved previously in Zhang & Moura [7] that the solution of the Most-Probable Configuration Problem, in Regime II, can not be a non-degenerate configuration when the underlaying network is k-regular, complete multipartite, or complete multipartite with k-regular islands. We will use this and Corollary 5.8 to prove this corollary.

Consider the contrapositive of Corollary 5.8: Let the density of the network be $d(G) = E/N$. Then, the most-probable configuration is not a non-degenerate configuration, $\mathbf{x}^* \in \{\mathbf{x}^0, \mathbf{x}^N\}$, if and only if there

does not exist any subgraph $H \in \mathcal{H}$ with density $d(H) = E'/N'$ for which $\lambda\beta^{d(H)} > \mu$, or

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} < \frac{N'}{N}.$$

The most-probable configuration is not a non-degenerate configuration for k-regular, complete multipartite, complete multipartite with k-regular islands networks. This implies that all the induced subgraphs, $H \in \mathcal{H}$, in these types of networks, satisfy the condition that $\lambda\beta^{d(H)} \leqslant \mu$ or

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} \geqslant \frac{N'}{N},$$

for all $0 < \lambda/\mu \leqslant 1, \beta > 1$ parameter values.

Depending on the effective infection rate and the endogenous infection rate, $(\lambda/\mu, \beta)$, the first condition $\lambda\beta^{d(H)} \leqslant \mu$ may not be satisfied. However, since N'/N can not be larger than 1 regardless of parameters and the underlying network, the second condition is satisfied if

$$\frac{\log((\lambda/\mu)\beta^{d(G)})}{\log((\lambda/\mu)\beta^{d(H)})} \geqslant 1 \quad \forall H \in \mathcal{H}.$$

With $\beta > 1$, this means that $d(H) \leqslant d(G)$ for all possible induced subgraph. As this only depends on the structure of the underlying network, we can conclude that $d(H) \leqslant d(G)$ for networks whose most-probable configuration can only be $\mathbf{x}^0$ and/or $\mathbf{x}^N$. □