Chaos, Solitons and Fractals 107 (2018) 170–176

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Inferring infection rate based on observations in complex networks

Zhen Su[a], Fanzhen Liu[a], Chao Gao[a,b,c], Shupeng Gao[a], Xianghua Li[a,*]

[a] College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[b] Potsdam Institute for Climate Impact Research (PIK), Potsdam, 11473, Germany
[c] Institute of Physics, Humboldt University of Berlin, Berlin, 12489, Germany

ARTICLE INFO

Article history:
Received 10 November 2017
Revised 11 December 2017
Accepted 27 December 2017

Keywords:
Complex networks
Sensor nodes
Infection rate
SI model
Diffusion and inference

ABSTRACT

The infection rate of a propagation model is an important factor for characterizing a dynamic diffusion process accurately, which determines the scale and speed of a diffusion. Inferring an infection rate, based on an observed propagation phenomenon, can help us better estimate the threat of a diffusion in advance and deploy corresponding strategies to restrain such diffusion. Meanwhile, the infection rate is a vital and predefined parameter in the field of propagation network reconstruction and propagation source identification. Therefore, how to infer an infection rate effectively from observed diffusion data is of great significance. In this paper, a backpropagation-based maximum likelihood estimation (BP-ML) is used to infer such infection rate. More specifically, a set of sensors are first deployed into a network for collecting diffusion data (i.e., the infection time of a node). Then, a series of backpropagations are initiated by nodes resided by these sensors in order to deduce the more probable infection rate based on the maximum likelihood estimation. Some experiments in real-world networks show that by taking full advantage of observed diffusion data, our proposed method can infer the infection rate of a diffusion accurately.

1. Introduction

In the modern society, the phenomena of network diffusion become ubiquitous since many real-world systems take the form of networks [1]. For instance, rumors spread through online social networks (OSN) [2], computer viruses propagate over the Internet [3], and epidemics diffuse via human contact networks [4]. These diffusions, triggered by network risks (e.g., rumors, computer viruses and epidemics), often result in detrimental and uncertain social effects [5]. Therefore, it is crucial for us to understand the potential mechanism of a diffusion process in order to control and restrain such diffusion.

Over the past decades, extensive research, revolving around diffusion dynamics, has been conducted based on simulations. To simulate a dynamic process, a diffusion model with a specified infection rate, is predefined and employed as a test-bed in many scenarios [6–8]. For example, in the domain of vital nodes identification [9,10], a diffusion process, initiated by vital nodes, could be characterized by the susceptible-infected (SI) model or the susceptible-infected-susceptible (SIS) model [11]. Through comparing the speed and scale of such propagation, we can estimate the influence and importance of these vital nodes. In most cases, an infection rate is assigned according to predefined assumptions. Another typical example is the network immunization so as to restrain malicious virus propagation (e.g., email worms), unwanted information dissemination (e.g., rumors) or infectious diseases transmission (e.g., H1N1 influenza pandemic) [12,13]. The corresponding models (e.g., computer virus propagation models [14,15], information diffusion models [16,17] and classical epidemic models [11]) in terms of the three kinds of threats, are adopted to simulate a diffusion process and evaluate the efficiency of proposed immunization strategies [18]. Similarly, infection rate in these studies is predefined artificially. All these studies above consider infection rate as a priori knowledge. Hence, it raises an important question: is it possible to deduce and estimate the potential infection rate during a diffusion process? If so, we can give an assessment with respect to the possible impact of a diffusion and adjust corresponding strategies to control it.

In addition, inferring an infection rate can be further applied to propagation network reconstruction [19] and propagation source identification [5]. First, with an inferred infection rate, the underlying propagation network can be reconstructed from the macro level of the temporal feature of a propagation process [19]. On the other hand, identifying propagation sources based on a set of observations, is an effective method to control diffusions triggered by network risks [5]. As an instance of the temporal feature of a propagation process, the observations in terms of a diffusion phenomenon are relevant to an infection rate. Therefore, She et al.

* Corresponding author.
E-mail addresses: cgao@swu.edu.cn (C. Gao), li_xianghua@foxmail.com (X. Li).

https://doi.org/10.1016/j.chaos.2017.12.029
0960-0779/

have proposed a method in which they first deduce the infection rate for reproducing the temporal feature of a propagation. Furthermore, they have identified propagation sources based on such temporal feature [20]. However, they do not provide a detailed explanation and analyze the influence of deployment strategies of sensors on the accuracy of infection rate inference.

Practically, a diffusion phenomenon can be observed from infected nodes or individuals. For example, an infected computer will be detected by security software and reported to security companies [21]; a hospital will report the information of patients with infectious disease to centers for disease control (CDC) [4]. The security software and the hospital can be seen as sensors which are used to monitor the states of nodes in a network. Hence, in this paper, we aim to answer the following questions:

1) Can we infer an infection rate through deploying a set of sensors in a network?

2) Which strategy is better for us to deploy these sensors so that we can improve the efficiency of infection rate inference?

Considering the stochastic effect of a real-world diffusion, we employ a backpropagation-based maximum likelihood estimation (BP-ML) to infer the potential infection rate. Specifically, the diffusion data in terms of a propagation process, is reported by the deployed sensors. Based on the BP-ML, a desirable result of infection rate inference is that the temporal feature of a propagation process, reproduced by a candidate infection rate, optimally matches the observed diffusion data. Furthermore, we estimate the influence of different deployment strategies of sensors on the inference accuracy. Experimental results show that our method can effectively deduce the potential infection rate of a diffusion.

The rest of this paper is organized as follows. Section 2 elucidates some important research questions. Section 3 introduces the BP-ML approach used for the infection rate inference. Section 4 presents some experiments to evaluate the performance of the BP-ML in real-world networks. Finally, we conclude this paper in Section 5.

2. Problem statement

In this section, we formulate the problem of infection rate inference. Then, the metric, i.e., the accuracy of infection rate inference, is used for estimating the performance of proposed method.

Mathematically speaking, the problem of inferring an infection rate is equivalent to a maximum likelihood problem. Specifically, our goal is to deduce an infection rate which triggers a diffusion process similar to a given observed phenomenon. In this paper, before deducing the potential infection rate, a set of sensors are deployed into a network in order to obtain observed diffusion data. A network with the deployed sensors can be formulated as a graph and defined in Definition 1.

Definition 1. A **graph** $G = (V, E, O)$ is a finite and undirected network, in which $V = \{v_i\}_{i=1}^{N}$ refers to a set of nodes; $E = \{e_{ij} | 1 \le i, j \le N, i \ne j\}$ is a set of links; $O = \{o_k\}_{k=1}^{K}$ denotes a set of sensors which are deployed on nodes based on a predefined strategy. The total number of nodes in the network is denoted as $N = |V|$; $|E|$ stands for the total number of edges and e_{ij} represents the link between nodes v_i and v_j; the total number of sensors is $K = |O|$ and $K \le N$.

Given a structure of G, a diffusion is initiated by an infection source with a predefined infection rate p^*. In most cases, p^* is fixed and determined by the type of infection. Meanwhile, such infection rate is unknown to users. It is a challenging problem for us to deduce p^* from the observed diffusion phenomena in the real world. Currently, it is feasible and easy-to-implement to deploy a set of sensors O into a network to monitor the states of nodes in order to collect necessary diffusion information. Here, the limited diffusion data, reported by the deployed sensors O, is called the observed infection time. Such time is denoted as $T^o = \{t_k^o\}_{k=1}^{K}$ where K is the total number of sensors. More specifically, t_k^o is the observed infection time reported by a sensor o_k.

Based on the observed infection time T^o, the main task of infection rate inference is to deduce and estimate an infection rate p which should be as close as possible to the real infection rate p^*. The optimal estimator of the infection rate inference is defined in Definition 2.

Definition 2. The **infection rate inference** refers to deducing an infection rate based on the observed infection time T^o, i.e.,

$$p = x : \max \mathbf{S}(T^o, T^r(x)), x \in [0.01, 0.09]$$

where p is the inferred infection rate; $T^r(x)$ is the simulated infection time with a given infection rate x; the $\max \mathbf{S}(\cdot)$ returns an infection rate x which yields the highest probability of generating the observed infection time T^o.

Intuitively, the performance of infection rate inference can be measured by the success rate which is defined as follows:

Definition 3. The **success rate** of the infection rate inference, is defined as the degree to which the inferred infection rate p is close to the real p^*, i.e.,

$$success\_rate = \begin{cases} 1, & p == p^* \\ 0, & p \ne p^* \end{cases}$$

Technically speaking, our method is completely data-driven because it is based on the observed diffusion data. As long as a diffusion process is detected by the deployed sensors, the potential infection rate can be deduced based on the BP-ML. Details on our proposed method are elaborated in the next section.

3. Infection rate inference

In this section, we first introduce the deployment strategy with respect to sensors in Section 3.1. Section 3.2 gives a brief description of the selected propagation model. Section 3.3 describes the BP-ML approach used for infection rate inference.

3.1. The deployment strategy

As defined in Section 2, this paper aims to deduce the potential infection rate by using the diffusion data reported by a set of sensors. Given a network G, we deploy sensors into the network based on a predefined strategy.

Many centrality measures have been proposed to estimate the influence of nodes [9,22], such as degree [23], betweenness [24], and k-core [25]. In general, nodes with a high centrality can facilitate a diffusion process prominently. Therefore, to ensure that a diffusion phenomenon can be detected by sensors as soon as possible, we deploy sensors based on different centrality measures. In detail, the efficiency of different deployment strategies is compared in Section 4.3.

3.2. SI model-based diffusions

In this paper, the diffusion process is modeled based on the SI model. In detail, each node in G has two corresponding states: (i) Infected, if it has been informed (infected) by one of its neighbors; (ii) Susceptible, if it has not been informed (infected). The initial infection source is selected randomly from G. At time t, a diffusion starts from infected nodes to susceptible ones through linking