Physica A 508 (2018) 10–17

Contents lists available at ScienceDirect

Physica A

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

Sequential seeding for spreading in complex networks: Influence of the network topology

Qipeng Liu *, Tao Hong

Institute of Complexity Science, Qingdao University, 308 Ningxia Road, Qingdao, Shandong, PR China

HIGHLIGHTS

- Sequential seeding strategies are constructed based on network centralities.
- Sequential seeding strategies are examined in both real and synthetic networks.
- Sequential seeding strategies perform better in networks with heterogeneous degrees.
- Average degree and assortativity coefficient also have influence on the performance.

ARTICLE INFO

Article history:
Received 16 January 2018
Received in revised form 25 March 2018

MSC:
05C82
91D30

Keywords:
Complex networks
Spreading models
Sequential seeding
Centrality measures

ABSTRACT

In this paper we investigate the problem of sequential seeding for spreading in complex networks. We focus on the influence of network topology on the performance of seeding strategies. The classic independent cascade model (ICM) is adopted to represent the spreading process. We examine the centrality measures—degree, K-shell, and H-index in several real networks and confirm that degree is a good indicator for spreading efficiency. Scale-free networks with tunable parameters such as power-law exponent, density, and assortativity coefficient are constructed as the testbed of the study. By simulations, we find that the advantage of sequential seeding strategy is large in a degree-heterogeneous network with relatively small average degree and large assortativity coefficient.

1. Introduction

Spreading in complex networks is a very common phenomenon in the real world, ranging from disease spreading among human beings through contact, information and rumor spreading in online social networking websites, to advertising though viral marketing in the society [1,2].

A hot research topic related to spreading in complex networks is how to choose a set of nodes in the network as initial seeds such that the spreading process starting from these seeds can reach the maximum fraction of the population in the network. This is the so-called influence maximization problem [3]. Another similar topic is influential nodes identification problem, i.e., find out a set of nodes which have the highest rank according to some specific measure [4–6]. In fact, if we set the measure as the size of the population reached in the spreading process, the problem of identifying influential nodes is specialized to influence maximization problem. Conversely, choosing influential nodes in the network as seeds to start

* Corresponding author.
E-mail address: qipengliu@qdu.edu.cn (Q. Liu).

https://doi.org/10.1016/j.physa.2018.05.057

the spreading process is in general a better strategy than randomly choosing seeds among all nodes. Therefore, these two research areas can provide valuable insight and inspiration to each other.

Searching for the optimal seeding strategy is in general an NP-hard problem. Thus, greedy algorithms and other heuristic strategies are used to search for approximate solutions in many applications. Especially, when dealing with a large-scale network, heuristic strategies seem to be the only choice [3,7]. Some classic centrality measures are used in constructing heuristic algorithms, such as degree, betweeness, closeness, PageRank, and K-shell. In a heuristic strategy, a set of nodes with highest centrality measure are selected as the seeds to start the spreading process.

A popular assumption in the existing work is that all of the seeds are activated at the initial time step and then they spread the active state to other nodes through the connections among them. Here we call it single-stage seeding strategy. Most recently, Jankowski et al. propose several sequential seeding strategies, in which only a part of seeds are activated at the initial time step and other seeds are activated in the subsequent steps according to some rules [8,9]. They show that for a fixed number of seeds, sequential seeding strategies can active more nodes than the single-stage seeding strategy. They provide extensive simulations to investigate the influence of many factors, such as centrality measures, seed percentage, and propagation probability, on the performance of sequential seeding strategies. However, the influence of network topology is rarely mentioned in their work. In this paper, we construct scale-free networks and provide a systematic study of the influence of network topological properties (such as network density, degree distribution, and assortativity coefficient) on the performance of sequential seeding strategies.

The paper is organized as follows: In Section 2, we introduce a widely used spreading model—the independent cascade model which captures the spreading dynamic in our work. In Section 3, resorting to real networks, we compare the performance of three sequential seeding strategies, which are constructed based on degree, K-shell, and H-index, respectively. Due to its best performance, we choose degree as the centrality measure in our sequential seeding strategy, and study the influence of network topology in artificial scale-free networks in Section 4. Concluding remarks are given in Section 5.

2. Spreading models

To study the spreading behavior in networks theoretically, we first need to set up the spreading model. In the literature, a variety of spreading models are proposed in different research areas. A review of the models and relations among them can be found in [10]. In our paper, we adopt a commonly used model—the independent cascade model (ICM).

There are many variants of ICM in the literature. Here we adopt the conceptually simplest version that is studied in [3,11]. The spreading process unfolds in discrete steps as follows:

(1) At the initial time step t_0, a set of nodes are activated, which are regarded as the seeds to trigger the spreading process.
(2) At any time step $t > t_0$, each of the active nodes, say v, has a change to activate each of the inactive nodes, say w, which is directly connected with it. It succeeds with a probability p called spreading rate, and whether or not it succeeds, it could not make further attempts to activate w in the subsequent time steps.
(3) The process runs until no more nodes can be possibly activated. The size of the population activated represents the coverage of the spreading process.

Based on this ICM spreading model, next we will study how the seeding strategies (sequential and single-stage) affect the final population coverage of the spreading process.

3. Sequential seeding in real networks

3.1. Single-stage seeding and sequential seeding

We compare two kinds of seeding strategies in this paper: single-stage seeding and sequential seeding. After ranking the nodes descendingly according to a given centrality measure, the two strategies are as follows:

Single-Stage Seeding: Select the top k nodes as the seeds and activate them all at the initial time step.

Sequential Seeding: Only select the top $k/2$ nodes as the first half number of seeds and activate them at the initial time step. Once the spreading process ends, select top $k/2$ nodes which are not activated yet as the other half number of the seeds and activate them.

The comparison of the two strategies is first conducted on several real networks from SNAP Datasets [12]. Among all the networks in SNAP datasets, we mainly choose social networks, where the spreading phenomenon is usually observed. And due to the restrict of our computational capacity, we have to choose medium size networks with less than 100 thousand nodes. We choose seven networks from SNAP and provide the summary of them in Table 1, where the first column shows the names of the networks, the second column contains the types of the networks (directed or undirected), the subsequent two columns show the numbers of nodes and edges, respectively, and the last column has brief descriptions of the networks and their corresponding references.