Search Engine Creates Cascading Effects in Social Networks

immediate

Abstract—The search engine, which is primarily designed for social network users to acquire information that interests them, is also a platform for spreading viruses. Recent years have witnessed a rapid development of viruses, and the wide variety of security threats caused by viruses heightens the need for studying virus propagation. Since the search engine plays a vital role in virus propagation, studying the relationship between search engines and the virus propagation process is vital. In this paper, we quantitatively discuss the characteristics of the virus propagation process in the presence of a search engine in social networks. First, we reveal that increasing the searching possibility leads to a change from a second to first percolation phase transition in the virus propagation process. We discuss the crucial searching possibility in both a Poisson distribution network and a power law distribution network. Second, we prove that the search engine accelerates virus propagation and expands the virus propagation area significantly by increasing the infection density and shrinking the network diameter. Third, we quantitatively characterize the effects of the search engine on the resilience of social networks by two metrics: the jump point and the giant components; we further find that the search engine makes social networks more vulnerable to virus infection. Moreover, we introduce the Google PageRank algorithm to analyze the positive feedback effect of the search engine.

Index Terms—Crucial searching possibility, giant components, infection density, jump point, network diameter, percolation phase transition, search engine, social network, virus propagation.

1 Introduction

The past few years have witnessed increasingly rapid development in the field of virus propagation. The wide variety of security threats caused by viruses makes studying virus propagation characteristics an urgent matter. The search engine, which is primarily designed for users to acquire information that interests them in social networks, has become a new threat that exacerbates the situation. Specifically, a search engine assists virus dissemination for social networks by inducing users to access content that includes the virus through keyword searches and transferring the virus from source users directly to potential interested users.

By accompanying such processes, the virus can break out on a large scale. One notable example, in which a search engine helped a virus to spread globally, was Bigfoot,[1] the plug-in of the World of Warcraft in September, 2010, which was widely used by many players. However, search engines had no strict control over advertisements and search results, which were utilized by the attacker. Search engines placed illegal search results above legal search results, which caused many identities to be exposed. In another example, Cisco Security discovered the virus Gumblar, which manipulated search results to increase advertising revenue, while allowing attackers to spread malicious codes. This malware would redirect search results to sites whose owners could profit from malicious activities. In the early years of virus propagation, a worm family, Mydoom,[2] used Web search engines to find new targets. These were mass mailer worms that searched email addresses by querying AltaVista, Google, Lycos, and Yahoo. These worms placed a significant burden on search engines. In one example from January, 2014, Yahoo's search engine spread viruses at a large scale: approximately 27,000 computers per hour were infected in countries such as Romania, Britain, and France.

As these examples illustrate, the search engine spreads the virus from one social network to another. Besides propagating the virus across networks, the search engine can also significantly promote the process with in a single network. Many social networks have a community structure. In the community structure, network nodes join together in tight, neat groups, and there are only a few connections between different groups. With the presence of a search engine, this limitation can be broken. Our paper focuses on this special case, although our analytical method can be extended to multi networks. Nowadays social network users are becoming increasingly dependent on one another, and are not separated from each other anymore. Their connections are built when users are interested in the same topic, thus coupling diverse social network users together. We show that owing to this coupling, networks are extremely sensitive to the virus, such that a random infection of a small fraction of nodes from fewer communities can produce an iterative cascade of virus propagation in many interdependent communities. Many large-scale virus outbreaks frequently result from a cascade of virus propagation among interdependent communities.

In tradition social networks, nodes in different communities can barely connect to each other, while in dependent communities, nodes from one community can connect with nodes from another community through the use of the search engine. A primary challenge for analyzing the virus propagation effect of the search engine lies in determining

1. https://www.secpulse.com/archives/6663.html

2. https://www.symantec.com/security response/

the processes whereby a search engine increases the propagation paths between two separated communities.

To address the problems noted above, several specific challenges must be addressed:

- Search engines enable users to access content of interest with keyword searching and by transferring content from the source users directly to potential interested users. ~~By~~ accompanying such processes, the social network evolves as new links emerge between users with common interests. In contrast to traditional virus propagation, the newly-emerging links increase the number of virus propagation paths, thus increasing the probability of getting infected. The virus can be spread among disconnected users or separated communities. Moreover, the mechanism of the search engine places malicious web pages at high priority, which further exacerbates virus propagation.
- Users in one community will search content that is published by users in another community, and thus two separated communities become "coupled." The likelihood that a user will use the search engine varies from user to user and usually changes over time.
- To quantitatively characterize the evolutionary process in the presence of a search engine, we need to properly model the virus propagation process driven by a search engine in social networks and adopt basic metrics to quantitatively evaluate the virus propagation process.
- The search engine brings new challenges to the virus defence, for it significantly increases the number of virus propagation paths. Whether the virus propagation process driven by a search engine in social networks can be controlled or not remains to be discussed.

To meet all these challenges, first we need to model the virus propagation process driven by the search engine. In this paper, we propose a ***virus propagation model***, which is inspired by the percolation theory in thermomechanics. Our model is an interdisciplinary model that combines computer science, thermomechanics, and complex networks. In our model, there is a social network that consists of several communities. If users from one community use the search engine to access content that is published by users from another community, or vice versa, we call these two networks "***coupling communities.***" In coupling communities, nodes from one community connect with nodes from another community with some likelihood. Consequently, when nodes from one community are infected, they cause nodes from the other community to also become infected. ~~This phenomenon hardly ever happens in separated communities.~~

In order to quantitatively characterize the virus propagation process, we adopt five metrics: the critical searching ratio, infection density, network diameter, jump point, and the giant components. The critical searching ratio is a critical value above which the virus propagation undergoes a first percolation transition. Infection density measures the coverage of the virus in the network, which is the essential metric for evaluating the virus propagation effect. Network diameter can be used to analyze the virus propagation speed. The jump point is a critical initial infection ratio: when the virus propagation undergoes a first-order transition, the stable infection density I_∞ has a single step discontinuity at the jump point. The giant components can be used to judge the network's robustness under the threat of virus.

The main contributions are as follows:

- We reveal that virus propagation in a social network is driven by the search engine. The search engine creates short-cuts between separated communities. We propose a virus percolation model that takes into the search engine consideration. Our model is rooted in the percolation theory and also introduces the generating functions for degree distribution and the underlying branching processes to describe the complex virus propagation process.
- We introduce the Google PageRank algorithm to analyze the positive feedback effect of the search engine. Then we further discuss the relationship between the possibility that a user will use the search engine and the expected value of social networks. More infected nodes make malicious web pages get higher rankings, thus inducing more nodes to click the malicious web pages. As a result, more and more nodes are infected. The expected PageRank value can be used as an indicator for the explosion of the virus. When the virus propagation undergoes a first-order transition, the expected PageRank value of social networks significantly increases.
- We reveal that increasing the searching ratio leads to a change from a second to first percolation transition in the virus propagation process. For a strong searching ratio, the network undergoes a first-order transition, which means the infection density changes discontinuously as the initial infection ratio changes. For a weak searching ratio, it undergoes a second-order phase transition, which means the infection density changes continuously. The changing laws of the infection density with the initial infection ratio have instructional significance for virus defense. By controlling the number of immune nodes above a critical value, we can prevent the virus from a large-scale outbreak.
- We quantitatively characterize how a search engine influences the metrics of the virus propagation process, such as the infection density, network diameter, jump point, and giant components. We mathematically and theoretically prove that the search engine increases infection density, shrinks network diameter, and makes social networks more vulnerable.

The rest of this paper is organized as follows. In Section 2, we present a virus propagation model. The theoretical mathematical analyses of the related metrics are presented in Section 3, respectively. Evaluations are given in Section 4 and related work is presented in Section 5. Finally, we discuss conclusion about our work in Section 6.

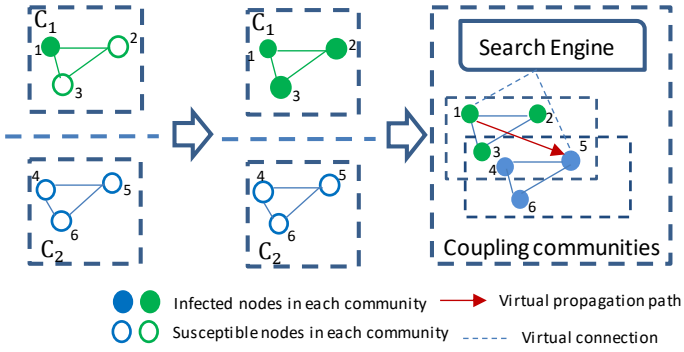


Fig. 1: Illustration of the effect of the search engine. (If a user uses the search engine to access interested content that is published by another user, or vice versa, we call these two nodes "coupling nodes," and these two communities "coupling communities." We suppose there are two separated communities C_1 and C_2, each of which has three nodes that are connected by internal links. At the beginning, node 1 is infected. When there is no search engine, the virus can only propagate to node 2 and node 3 through internal links. In the presence of a search engine, nodes can connect with each other without being limited by the community structure. "Coupling nodes" cause communities to become "coupling communities." Due to the virtual propagation path (red arrow), the virus is propagated from the virus outbreak community C_1 to another community C_2.)

2 Model

2.1 Social Network Model with a Search Engine

Our social network model comprises three entities: social actors (such as individuals or organizations), interactions, and a search engine. Our model consists of m nodes and can be described as an undirected graph whose nodes represent social network users and whose links represent connections between users. Links are built in two ways: the search engine (where links are built based on common interest), and the internal relationship (such as friendship, partnership, and kinship). Various connections built by the search engine and the internal relationship cause nodes to gather in multiple communities. ~~Our model highlights a property found in many social networks, community structure~~. In a community structure, network nodes join together in tight, neat groups, and there are only a few connections between different groups.

Beside the social actors and interactions that most current social network models include, our model emphasizes the role of the search engine. In recent years, search engines have been widely used by social network users. A user uses the search engine to search for content of interest that has been released by another user; a link is built between them. The search engine can be considered an intermediate node of this link that has connections with both users. Multiple similar links make the search engine become a super node in network A. We consider the super node as a complete graph consisting of m nodes, where the degree distribution of graph B can be described as $P_B(k) = 1, k = m$. In this way, all users in network A have their own virtual search engines, and they can choose to use them or not.

The search engine plays an important role in the information dissemination process in social networks. However, it is a double-edged sword. Consider a virus outbreak situation where due to the community structure, the virus propagation is limited to several communities. In the presence of a search engine, the virus can spread to the whole network without being limited by the community structure. If the users that use the search engine change over time, the initial small-scale infection may trigger a recursive process of virus spread that can completely cover the whole network.

Fig. 2 shows each stage of the virus propagation process in our model. Let I_{A_t} denote the infection density at stage t of the cascade process, and I_{B_t} denote the fraction of nodes that contains viruses ~~in graph B~~ at stage t of the recursive process of virus spread. The infected nodes infect their neighbor nodes in accordance with the percolation theory. We define g_A and g_B as the fraction of nodes belonging to the giant components of networks A and B. As discussed in [5]and [6], we will introduce the generating function of the degree distributions $G_{A_d}(\xi) = \sum_k P_A(k)\xi^k$, where $P_A(k)$ is the degree distribution of network A and ξ is an arbitrary variable. Analogously we will introduce the generating function of the underlying branching processes, $G_{A_u}(\xi) = \frac{G_{A_d}(\xi)'}{G_{A_d}(1)'}$. Random infection of fraction $1-p$ of nodes will change the degree distribution of the remaining nodes, so the generating function of the new distribution is equal to the generating function of the original distribution with the argument ~~equal~~ to $1-(1-\xi)$ [5]. The fraction of nodes that are not infected after the initial infection of $1-p$ nodes is [6]

$$g_A(p) = 1 - G_{A_d}(1-(1-f_A)p), \tag{1}$$

where $f_A = f_A(p)$ satisfies a transcendental equation

$$f_A = G_{A_u}(1-(1-f_A)p). \tag{2}$$

2.2 Basic Epidemic Propagation Models

There are three basic models of epidemic propagation: SI [12], SIS [13], and SIR [5]. The states of individuals are divided into three states: susceptible (S), infectious (I), and recovered (R). A user is either susceptible, meaning that he/she has not yet been infected, infectious, meaning that he/she is infected and is capable of spreading the virus to his/her contacts, or recovered, meaning that he/she is no longer spreading the virus. Let S_t, I_t, R_t represent the susceptible density, infection density, and recovered density, respectively. Nodes are infected with possibility λ, recover with possibility β, and become immune with possibility μ. We describe the three models as follows.

2.2.1 SI

The SI model is usually used to describe diseases that cannot be cured after the person is infected and can also describe the early behavior of diseases. The SI system can be expressed by the following set of ordinary differential equations [12]:

$$\begin{cases} \frac{dS_t}{dt} = -\lambda I_t S_t, \\ \frac{dI_t}{dt} = \lambda I_t S_t. \end{cases} \tag{3}$$

2.2.2 SIS

Some infections, for example those from the common cold and influenza, do not confer any long lasting immunity. Such infections do not give immunization upon recovery from infection, and individuals become susceptible again.

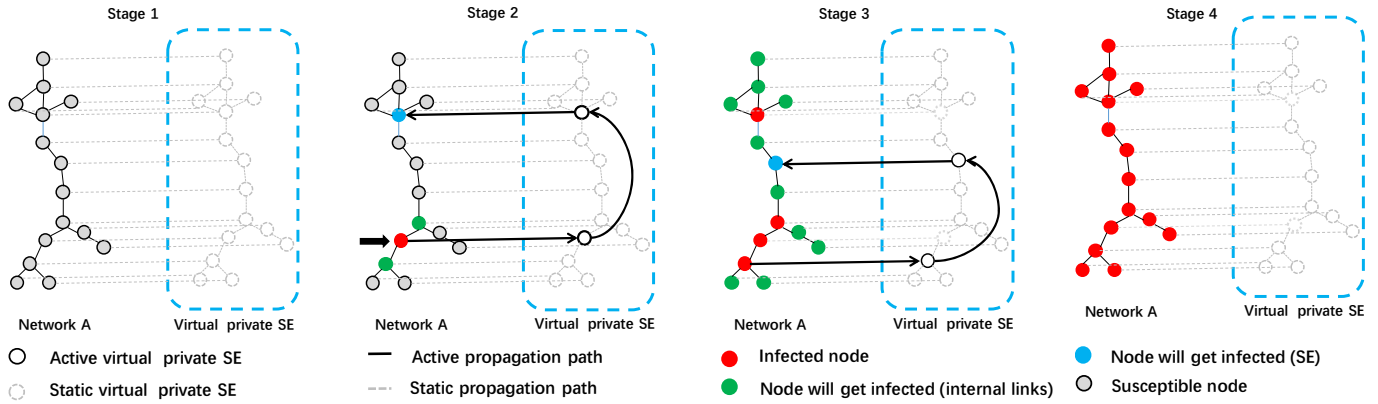


Fig. 2: Modeling an iterative process of a cascade of virus propagation. We suppose the search engine is a complete graph consisting of m virtual private search engines. All users in network A have their own virtual search engines, and they can choose to use them or not. **Stage 1**: The initial stage in which each node in network A has its virtual private search engine and all nodes are susceptible. **Stage 2**: A node in network A is attacked (red node). Then the virus will be propagated through internal links (green node) and virtual links added by the search engine (blue node). On one side, virus propagates from an infected node to its neighbors through internal links. On the other side, if a susceptible node uses a search engine and locate to an infected node, it will be infected by a virtual link. **Stage 3**: The newly-infected nodes make some of their branches depart from the giant components, and these nodes will get infected at next stage. Also, a susceptible node can use a search engine and locates to an infected node: it will be infected by a virtual link. **Stage 4**: Following the iterative process above, all nodes become infected.

The SIS system can be expressed by the following set of ordinary differential equations [13]:

$$\begin{cases} \frac{dS_t}{dt} = -\lambda I_t S_t + \beta I_t, \\ \frac{dI_t}{dt} = \lambda I_t S_t - \beta I_t. \end{cases} \tag{4}$$

2.2.3 SIR

The SIR model is reasonably predictive for infectious diseases that are transmitted from human to human, and where recovery confers lasting resistance, such as measles, mumps, and rubella. The SIR system can be expressed by the following set of ordinary differential equations [5]:

$$\begin{cases} \frac{dS_t}{dt} = -\lambda I_t S_t, \\ \frac{dI_t}{dt} = \lambda I_t S_t - \mu I_t, \\ \frac{dR_t}{dt} = \mu I_t. \end{cases} \tag{5}$$

The three basic models can describe the dynamic change relation between I_t and S_t well; however, they have limitations when we hope to analyze more parameters in the virus propagation process, such as the initial infection density and the distribution of infection nodes. Moreover, they ignore the greater number propagation paths added by a search engine, which makes the virus propagation process more complicated. Actually, epidemic propagation is a very complicated process and varies in different situations. Different situations need different models. In this paper, we propose a virus percolation model based on the classical percolation theory. Theory and practice have proved that our model is a good explanation of the various virus propagation characteristics in the presence of a search engine.

2.3 Problem Statement

To the best of our knowledge, our work is the first to adopt the percolation theory [3] in the analysis of virus propagation in the presence of a search engine. In our study, we divide the network nodes into three states: susceptible (S), infectious (I), and recovered (R) [4]. S_t, I_t, and R_t denote the susceptible density, the infection density, and the recovered density respectively. We divided the phase of network nodes during the virus propagation process into three phases:

- phase 1 (initial phase): all susceptible ($S_t = 1, I_t = 0, R_t = 0$).
- phase 2 (middle phase): partially susceptible ($S_t < 1, I_t < 1, R_t < 1$).
- phase 3 (stable phase): global outbreak ($S_t = 0, I_t = 1$ in SI, $S_t = c_1, I_t = c_2$ in SIS, $S_t = c_3, I_t = c_4, R_t = c_5$ in SIR, where $c_1 \sim c_5$ are all constants).

By changing the searching ratio q, as the initial infection ratio $1 - p$ grows, the final infection density I_{A_∞} undergoes a first-order transition (I_{A_∞} changes discontinuously from $phase1$ to $phase3$) or a second-order transition (I_{A_∞} changes continuously from $phase1$ to $phase2$ to $phase3$). On the one hand, a search engine can influence phase transition in the virus propagation process. On the other hand, many virus propagation characteristics change due to the presence of a search engine. In order to analyze the exact effects of a search engine on the virus propagation process, we adopt the following metrics:

- ***Critical searching ratio.*** ~~Virus propagation in a general social network undergoes a second-order phase transition.~~ We find that increasing the searching ratio of the search engine leads to a change from a second to first-order phase transition in the virus propagation process. When the searching ratio increases to a critical value, the virus propagation undergoes a first-order phase transition. We call this critical value the critical searching ratio.
- ***Infection density.*** Infection density is defined as $I_t = M_t/N_t$, where M_t is the number of infected nodes at time t, and N_t is the total number of nodes at time t. Infection density is usually used to measure the coverage of viruses in a social network. It is the essential metric for evaluating the virus propagation effect.
- ***Network diameter.*** The network diameter is the maximum of the shortest paths. The mathematical expectation value of the shortest path in a social

network is defined as: $\mathbb{E}_{route} = \sum N_k p_k$, where N_k denotes the number of a certain kind of route, and p_k denotes the probability of this route. This metric is mainly used to measure the virus propagation speed in a social network.

- ***Jump point.*** When the virus propagation undergoes a first-order phase transition, the stable infection density I_∞ changes discontinuously with the initial infection ratio and has a single step discontinuity at the threshold p_c. It is quite different from a second-order phase transition, where I_∞ changes continuously at $p = p_c$. We call the crucial value p_c the jump point. The jump point can be used to measure a network's level of resistance in the case of a virus outbreak.
- ***The giant components.*** A random non-connected graph can be divided into several connected branches. The connected branch with the largest number of nodes is "the giant components" of the network. ~~The ratio of the giant components' order to the whole network's order is generally understood as the network's robustness~~. The giant components are always used to analyze network stability.

2.4 Notations

TABLE 1: Notations.

Variable	Definition
q	Nodes in network A search for their topics of interest and navigate to another node with ratio q.
a	The average degree of network A.
m	Number of nodes in network A.
$1-p$	The initial infection fraction.
μ	The probability that an infected node will recover at each stage of the infection cascade process.
r	The fraction of infected nodes that are immune at each stage of the infection cascade process.
t	The stage of the infection cascade process.
I_{A_i}	The infection density at stage i of the cascade process of network A.
I_{B_i}	The infection density at stage i of the cascade process of graph B.
g_A	The fraction of nodes belonging to the giant components of networks A.
g_B	The fraction of nodes belonging to the giant components of graph B.
f_A	The possibility that a randomly chosen outgoing link will not lead to an infinitely large giant component in network A.
f_B	The possibility that a randomly chosen outgoing link will not lead to an infinitely large giant component in graph B.
$P_A(k)$	The degree distribution of network A.
$P_B(k)$	The degree distribution of graph B.
k	The degree of a node.
E_{A_0}	Number of edges in network A at the initial time.
D_{max}	The distance between two unconnected nodes is regarded as infinity, and is denoted as a big constant D_{max}.

3 Analysis

3.1 Critical Searching Ratio

Theorem 1. *When the network follows the Poisson Distribution, increasing the searching ratio q leads to a change from a first to a second-order phase transition in the virus propagation process. There exists a crucial searching ratio q_c, where any q that satisfies $q > q_c$ will cause virus propagation to undergo a first-order phase transition, and when $q < q_c$, the virus propagation will undergo a second-order phase transition.*

Proof. In the following subsections, we will discuss the crucial searching ratio q_c in two aspects SIS (SI) and SIR.

3.1.1 SIS (SI)

The iterative process of virus propagation is initiated by randomly infecting a fraction $1-p$ of network A nodes. All the neighbor nodes that are connected to them with A edges may get infected. Moreover, because of the presence of the search engine, users use it to search for information of interest, thus building connections with more nodes which can never connect without the search engine. By accompanying this process, the virus can spread on a larger scale. As more and more nodes are infected, the network A breaks up into connected components (clusters). Our key idea, based on the percolation theory, is that when the network is fragmented, the nodes belonging to the largest component (giant components) connecting a finite fraction of the network are still susceptible, while nodes that are parts of the remaining small clusters become infected. Since each node can potentially use the search engine, any node may get infected. This leads to the infection of more dependent nodes.

Since search engines are widely used by social network users, when users search for information that interests them, there exist links between users and the search engine. The search engine can be considered to be a virtual super node B in network A. The virtual super node B is connected with nodes in network A with ratio q, which means the users in network A use the search engine to access information with a searching ratio q. To simplify the analysis and prevent loss of generality, we suppose the search engine is a complete graph consisting of m virtual search engines which we call graph B. All users in network A have their own virtual private search engines, and they can choose to use them or not. Then our model can be mapped into two networks A and B, which communicate with each others by search engine with ratio q.

Next we present the virus propagation process step by step. The initial infection is $1-p$, namely $I_{A_0} \equiv 1-p$. A ratio q of nodes from graph B connects with nodes from network A, and consequently $q(1-pg_A(p))$ nodes get infected through the search engine. Consider that at each stage of the infection cascade process, the infected nodes recover with possibility μ, $I_{A_0} \sim I_{A_n}$ times coefficient $1-\mu$. Similarly, we can construct the sequence I_{A_i} and I_{B_i} at each stage of the cascade of virus propagation. The general form is given by

$$
\begin{cases}
I_{A_0} = & 1-p \\
I_{B_0} = & q(1-g_A(1-I_{A_0})p), \\
I_{A_1} = & (1-\mu)(1-(1-q(1-g_B(1-I_{B_0})))p), \\
I_{B_1} = & q(1-g_A(1-I_{A_1})p), \\
\dots & \\
I_{A_n} = & (1-\mu)(1-(1-q(1-g_B(1-I_{B_{n-1}})))p), \\
I_{B_n} = & q(1-g_A(1-I_{A_n})p), \\
\dots &
\end{cases}
\tag{6}
$$

$$
g_A(p) = 1 - G_{A_d}(1-(1-f_A)p),
$$

Consider I_{A_n} and I_{B_n}: at the end of the cascade process, the infection density does not change anymore, so we have $I_{A_n} = I_{A_{n-1}}$, $I_{B_n} = I_{B_{n-1}}$. Let $I_{A_n} \to 1-x$, $I_{B_n} \to 1-y$; then we have

$$\begin{cases} x = 1-(1-\mu)(1-(1-q(1-g_B(y)))p), \\ y = 1-q(1-g_A(x)p). \end{cases} \tag{7}$$

For the search engine, which is considered to be a complete graph consisting of m nodes, the degree distribution of graph B can be described as $P_B(k) = 1, k = m$.

When network A follows the Poisson distribution, the problem can be solved explicitly. From (1), we have

$$\begin{cases} G_{A_u}(\xi) = G_{A_d}(\xi) = exp[a(\xi-1)], \\ G_{B_d}(\xi) = \xi^m, \\ G_{B_u}(\xi) = \xi^{m-1}, \\ g_A(x) = 1-f_A, \\ g_B(y) = 1-(1-y(1-f_B))f_B. \end{cases}$$

$$g_A(p) = 1-G_{A_d}(1-(1-f_A)p),$$

Substitute (8) into (7) to get

$$f_A = G_{A_u}(1-(1-f_A)p).$$

$$\begin{cases} x = 1-(1-\mu)(1-(1-qf_B)p), \\ y = 1-q(1-(1-f_A)p), \end{cases} \tag{9}$$

where f_A and f_B satisfy the transcendental equations (2), and then we have

$$\begin{cases} f_A = exp[ax(f_A-1)], \\ f_B = (1-(1-f_B)y)^{m-1}. \end{cases} \tag{10}$$

Substitute x and y into the above formula:

$$\begin{cases} f_A = exp[a(1-(1-\mu)(1-(1-qf_B)p))(f_A-1)], \\ f_B = (f_B+q(1-f_B)(1-(1-f_A)p))^{m-1}. \end{cases} \tag{11}$$

The solutions of system (11) can be graphically presented on a f_A, f_B, as shown in Fig. 3. Let $f(f_A) = \frac{log f_A + \frac{1}{f_A} + 1}{(f_A-1)^2} = C(constant)$. where $C \in (\frac{1}{2}, +\infty)$. The critical searching ratio satisfies

$$q_c = \frac{1}{\sqrt{2a(1-\mu)p}} \sqrt{\frac{(m-2)C}{(m-1)^2}}. \tag{12}$$

So we get $\frac{1}{2p}\sqrt{\frac{m-2}{a(1-\mu)(m-1)^2}} \le q_c \le 1$. When the network follows Poisson distribution, there exists a crucial searching ratio q_c, where any q that satisfies $q > q_c$ will cause the virus propagation to undergo a first-order phase transition, and when $q < q_c$, the virus propagation will undergo a second-order transition. We take the min value and propose the conclusion that when the searching ratio q satisfies $q < \frac{1}{2p}\sqrt{\frac{m-2}{a(1-\mu)(m-1)^2}}$, the virus propagation undergoes a second-order phase transition.

3.1.2 SIR

We consider the situation in which, at each stage of the infection cascade process, a fraction r of the infected nodes become immune. Being limited to the technical level, resources, and recovery ability, r cannot be infinity, we denote r_{max} as the maximum value of r.

Following a similar approach we can construct the sequence I_{A_i} and I_{B_i} at each stage of the cascade of virus propagation. The general form is given by

$$\begin{cases} I_{A_0} = & 1-p, \\ I_{B_0} = & q(1-g_A(1-I_{A_0})p), \\ I_{A_1} = & 1-(p-r)(1-q(1-g_{B_0}(1-I_{B_0}))), \\ I_{B_1} = & q(1-(p-r)g_{A_1}(1-I_{A_1})), \\ \dots & \\ I_{A_n} = & (1-(p-nr)(1-q(1-g_{B_{n-1}}(1-I_{B_{n-1}})))), \\ I_{B_n} = & q(1-(p-nr)g_{A_n}(1-I_{A_n})), \\ \dots & \end{cases} \tag{13}$$

At the end of the cascade process, the infection density does not change anymore, so we have $I_{A_n} = I_{A_{n-1}}$, $I_{B_n} = I_{B_{n-1}}$. Let $M_{A_n} = x$, $M_{B_n} = y$, $g_{A_\infty} = g_A(x)$, $g_{B_\infty} = g_B(y)$, then we have

$$\begin{cases} x = (p-r_{max})(1-q(1-g_B(y))), \\ y = 1-q(1-(p-r_{max})g_A(x)). \end{cases} \tag{14}$$

Random infection of fraction $1-p$ of nodes will change the degree distribution of the remaining nodes, so the generating function of the new distribution is equal to the generating function of the original distribution with the argument equal to $1-(1-\xi)$ [5]. At each stage of the infection cascade process, a fraction r of the infected nodes become immune. So we have

$$g_{A_i}(p-ir) = 1-G_{A_d}(1-(p-ir)(1-f_{A_i})), \tag{15}$$

where $f_{A_i} = f_{A_i}(p-ir)$ satisfies a transcendental equation

$$f_{A_i} = G_{A_u}(1-(p-ir)(1-f_{A_i})). \tag{16}$$

For the search engine that is considered to be a complete graph consisting of m nodes, the degree distribution of graph B can be described as $P_B(k) = 1, k = m$.

When network A follows the Poisson distribution, the problem can be solved explicitly. Therefore, we have

$$\begin{cases} G_{A_u}(\xi) = G_{A_d}(\xi) = exp[a(\xi-1)], \\ G_{B_d}(\xi) = \xi^m, \\ G_{B_u}(\xi) = \xi^{m-1}, \\ g_A(x) = 1-f_A, \\ g_B(y) = 1-(1-(1-f_B)y)f_B. \end{cases} \tag{17}$$

At the stable stage, we have

$$\begin{cases} x = (p-r_{max})(1-qf_B), \\ y = 1-q(1-(p-r_{max})(1-f_A)), \end{cases} \tag{18}$$

where f_A and f_B satisfy the transcendental equations

$$\begin{cases} f_A = exp[ax(f_A-1)], \\ f_B = (1-(1-f_B)y)^{m-1}. \end{cases} \tag{19}$$

Substitute x and y into the above formula:

$$\begin{cases} f_A = exp[a(p-r_{max})(1-qf_B)(f_A-1)], \\ f_B = (f_B+q(1-f_B)(1-(p-r_{max})(1-f_A)))^{m-1}. \end{cases} \tag{20}$$

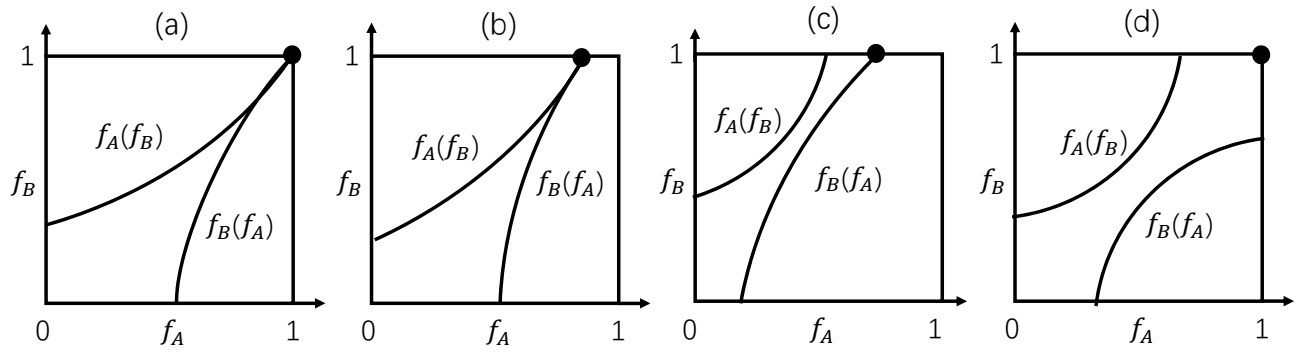


Fig. 3: Different graphical solutions of system (11). (System (11) can be considered to be an equation system consisting of two functions: $f_A(f_B)$ (independent variable f_B and dependent variable f_A) and $f_B(f_A)$ (independent variable f_A and dependent variable f_B). The solutions are presented as a crossing of either $f_B(f_A)$ or $f_A = 1$ with $f_A(f_B)$ or $f_B = 1$ and are restricted to the square $0 \le f_A \le 1; 0 \le f_B \le 1$. Since the search engine is considered to be a complete graph, we have $f_B = 1$, the possible solutions can be describe as Fig. 3. As shown in Fig.3 (a)∼(b), f_A and f_B change continuously which corresponds to the second-order phase transition. At the touching point of the two curves, the system satisfies $\frac{df_B(f_A)}{df_A} \cdot \frac{df_A(f_B)}{df_B} = 1$, in which an infinitesimal change $\triangle z$ in the vector of the system parameters $z = (a, m, p, q, \mu)$ may lead to a first-order phase transition in which the size of one or both of the giant components changes discontinuously from a finite value to zero: Fig.3 (c)∼(d). Then $\frac{df_B(f_A)}{df_A} \cdot \frac{df_A(f_B)}{df_B} < 1$.

The first equation can be solved with respect to f_B, and the second equation can be solved with respect to f_A:

$$\begin{cases} f_A = 1 - \frac{1}{p - r_{max}} + \frac{f_B^{\frac{1}{m-1}} - f_B}{(p - r_{max})q(1 - f_B)}, f_B \neq 1, \forall f_A, f_B = 1 \\ f_B = \frac{1}{q}(1 - \frac{log f_A}{a(p - r_{max})(f_A - 1)}), f_A \neq 1, \forall f_A, f_B = 1. \end{cases} \tag{21}$$

We get the critical searching ratio

$$q_c = \frac{1}{\sqrt{2a}(p - r_{max})} \sqrt{\frac{(m-2)C}{(m-1)^2}}. \tag{22}$$

So we get $\frac{1}{2(p-r_{max})}\sqrt{\frac{m-2}{a(m-1)^2}} \le q_c \le 1$. When the network follows a Poisson distribution, there exists a crucial searching ratio q_c, in which any q that satisfies $q > q_c$ will cause the virus propagation to undergo a first-order phase transition, and when $q < q_c$, the virus propagation undergoes a second-order transition. We take the min value and propose the conclusion that when the searching ratio q satisfies $q < \frac{1}{2(p-r_{max})}\sqrt{\frac{m-2}{a(m-1)^2}}$, the virus propagation undergoes a second-order phase transition.

3.2 Virus Propagation Speed in the Presence of a Search Engine

Theorem 2. *The search engine increases infection density and shrinks network diameter, thus accelerating virus propagation and expanding the virus propagation area significantly in social networks.*

Proof. In this section, we choose infection density and the network diameter to measure the characteristics of virus propagation in the presence of a search engine.

3.2.1 Infection Density

Basically, infection density measures the coverage of the virus in the network. It is the essential metric for evaluating the virus propagation effect. We compare infection densities with and without a search engine in a social network to observe the effect of the search engine.

First, we consider the infection density in a social network with the presence of search engine (we take SIS as an example). At time t, the infection density in network A reaches

$$I_{A_t} = (1-\mu)(1-(1-q(1-g_B(1-I_{B_{t-1}}))p). \tag{23}$$

Without the presence of the search engine, namely $q = 0$, the infection density in network A is

$$I'_{A_t} = (1-\mu)(1-p). \tag{24}$$

We can know $I_{A_t} - I'_{A_t} = pq(1-\mu)(1-g_B(1-I_{B_{t-1}})) \ge 0$, which means the search engine increases the infection density and expands the virus propagation area significantly.

3.2.2 Network Diameter

Network diameter is closely related to the shortest path. We analyze the changing trend of the network diameter to observe the virus propagation effect of the search engine. The value of the network diameter with a search engine is expected to be lower than it is without a search engine. In an extreme condition, where a user connects all topics through use of the search engine, the network diameter will be close to 1, and any two users can share information. So we come to a preliminary conclusion that the search engine shrinks the network diameter; a rigorous proof follows.

When there is a search engine, the mathematical expectation value of the shortest path is as follows:

$$\begin{aligned} D =& E_{A_0} \cdot 1 + (\frac{m(m-1)}{2} - E_{A_0})q + \\ & D_{\max}(\frac{m(m-1)}{2} - E_{A_0} - (\frac{m(m-1)}{2} - E_{A_0})q) \\ =& E_{A_0} + (\frac{m(m-1)}{2} - E_{A_0})q(1 - D_{\max}) \\ & + D_{\max}(\frac{m(m-1)}{2} - E_{A_0}). \end{aligned} \tag{25}$$

When there is no search engine, the mathematical expectation value of the shortest path is as follows:

$$\begin{aligned} D' =& E_{A_0} \cdot 1 + D_{\max}(\frac{m(m-1)}{2} - E_{A_0}) \\ =& E_{A_0} + D_{\max}(\frac{m(m-1)}{2} - E_{A_0}). \end{aligned} \tag{26}$$

Suppose that when q increases $\triangle q$, the variation of the diameter can be measured as

$$\triangle D = (\frac{m(m-1)}{2} - E_{A_0})(1 - D_{\max}) \triangle q. \quad (27)$$

We can know that $D_t < D'_t$. The result shows that the search engine shrinks the network diameter in social networks, and the higher q is, the shorter the diameter is. In the presence of a search engine, nodes in different communities can also be connected by the search engine with a certain degree of probability. Thus the probability that a route will be created between two nodes will be increased. This results in the decrease in the number of separated nodes, so the value of the maximum shortest path will be decreased. Therefore, in social networks, the mathematical expectation value of the network diameter will be shorten in the presence of a search engine.

3.3 Effects of the Search Engine on the Resilience of Social Networks

Theorem 3. *The search engine decreases the resilience of coupling networks and makes the coupling networks more vulnerable to virus infection.*

In this section, we choose the jump point p_c and the giant components to measure the effects of the search engine on the resilience of social networks.

3.3.1 Jump Point

It is known that reducing the coupling strength leads to a first to second-order transition in coupling networks [3]. When virus propagation undergoes a first-order phase transition, the stable infection density I_∞ changes discontinuously with the initial infection ratio and has a single step discontinuity at the threshold p_c, where it changes from zero for $p < p_c$ to $I(p_c) > 0$ for $p = p_c$. This behavior is characteristic of a first-order phase transition, quite different from a second-order phase transition such as that characterizing percolation of a single network, where $I(p)$ is a continuous function at $p = p_c$. We call the crucial value p_c the *jump point*. In this section, we analyze the jump point with and without the use of a search engine. The larger the initial infection ratio $1 - p_c$, the more resilient the coupling networks. We prove that the jump point is always larger in the presence of a search engine than without a search engine, which means that a search engine decreases the resilience of coupling networks.

Proof. Here, we consider a social network with a power law distribution. We denote the degree distribution of network A as $P_A(k) = ak^{-\alpha}$. The generating function of the degree distributions is $G_{A_d}(\xi) = \sum_k P_A(k)\xi^k$, and the generating function of the underlying branching processes is $G_{A_u}(\xi) = \frac{G_{A_d}(\xi)'}{G_{A_d}(1)'}$. So we get

$$\begin{cases} G_{A_d}(\xi) = \sum_k ak^{-\alpha}\xi^k, \\ G_{A_u}(\xi) = \frac{\sum_k ak^{1-\alpha}\xi^{k-1}}{\sum_k ak^{1-\alpha}}. \end{cases} \quad (28)$$

As p decreases, the nontrivial solution $f_A < 1$ of Eq. (4) gradually approaches the trivial solution $f_A = 1$. Accordingly, I_{A_∞} (selected as the order parameter of the transition) gradually approaches 1, as in a second-order phase transition, and becomes zero when two solutions of Eq. (4) coincide at $p = p_c$. At this point the straight line corresponding to the right-hand side of Eq. (4) becomes tangent to the curve corresponding to its left-hand side, yielding $p_c = 1/G_{A_1}(1)'$. So we have

$$p_c = (\frac{\sum_k k^{2-\alpha}}{\sum_k k^{1-\alpha}} - 1)^{-1}. \quad (29)$$

In the presence of a search engine, the distribution of network A follows a summation of a series of power law [7], and we denote the new distribution as $P_{A_{SE}}(k) = ak^{-\alpha} + b\sum_{i=1}^{k} i^{-\alpha-1}$. Similarly, we have

$$\begin{cases} G_{A_{dSE}}(\xi) = \sum_k (ak^{-\alpha} + b\sum_{i=1}^{k} i^{-\alpha-1})\xi^k, \\ G_{A_{uSE}}(\xi) = \frac{\sum_k (ak^{1-\alpha} + bk\sum_{i=1}^{k} i^{-\alpha-1})\xi^{k-1}}{\sum_k (ak^{-\alpha} + b\sum_{i=1}^{k} i^{-\alpha-1})}, \\ p_{c_{SE}} = \left(\frac{\sum_k (ak^{2-\alpha} + bk^2\sum_{i=1}^{k} i^{-\alpha-1})}{\sum_k (ak^{1-\alpha} + bk\sum_{i=1}^{k} i^{-\alpha-1})} - 1\right)^{-1}. \end{cases} \quad (30)$$

It is obvious that $p_{c_{SE}} > p_c$. The result shows that in the presence of a search engine, fewer initial infection nodes can nevertheless lead to the collapse of the whole network. So we conclude that a search engine decreases the resilience of coupling networks.

Studying the jump point has great significance for virus defense. By increasing the immune nodes above p_c, we can control the virus propagation process.

3.3.2 The Functional Node Density

A random non-connected graph can be divided into several connected branches. The connected branch with the largest number of nodes is called "the largest connected components" of the network. Our insight based on the percolation theory is that when the network ids fragmented, the nodes belonging to the giant component connecting a finite fraction of network nodes are still functional, whereas the nodes that are part of the remaining small clusters become non-functional. The functional node density can be used to measure the resilience of a network after a virus outbreak.

When there is no search engine, different users in different communities can hardly connect, and thus the virus coverage is limited. The situation is completely different in the presence of a search engine, where users can use the search engine to search for web pages of interest to them in any community: thus virus coverage can increase significantly. As a result, at the end of the virus propagation process, the largest connected components level is greater with a search engine than without a search engine, which means that social networks become more vulnerable after a virus outbreak.

Proof. For network A, when the cascade process of the virus propagation reaches a stable state, the fraction of nodes in the giant components can be calculated as

$$g_A(p) = 1 - G_{A_d}(1 - (1 - f_A)p), \qquad f_A = G_{A_u}(1 - (1 - f_A)p).$$

$$G_{A_u}(\xi) = G_{A_d}(\xi) = exp[a(\xi - 1)],$$

$G_{A_\infty} = M_{A_\infty} g_A(M_{A_\infty})$, and in a Poisson distribution SI model we have

$$G_{A_\infty} = (1 - f_A)(1 - q f_B)p. \tag{31}$$

When there is no search engine, the fraction of the giant components can be calculated as

$$G'_{A_\infty} = (1 - f_A)p. \tag{32}$$

Obviously, $G_{A_\infty} < G'_{A_\infty}$. Actually, at each time slot the results still held up. According to Theorem 2, we know that $I_{A_t} > I'_{A_t}$, so $M_{A_t} < M'_{A_t}$. Since g_A is a monotonic increasing function, it is clear that $G_{A_t} < G'_{A_t}$. The presence of the search engine makes social networks more vulnerable to virus infection.

3.4 Relationship between the Searching Ratio and the Expected PageRank Value

PageRank, the webpage scoring approach adopted by Google developers Brin and Page, produces rankings independent of a user's query. The importance of a webpage in this case is determined (in large part) by the number of other important webpages that are pointing to that page and the number of outlinks from those other webpages. Suppose that there are M normal web pages $S_1, S_2, ..., S_M$ and N infected webpages $I_1, I_2, ..., I_N$. Then the PageRank value of S_1 can be calculated as

$$PR(S_1) = \sum_{i=1}^{M} \beta_i \frac{PR(S_i)}{L(S_i)} + \sum_{j=M+1}^{M+N} \beta_j \frac{PR(I_j)}{L(I_j)}, \tag{33}$$

where $\beta_i \sim \beta_M$ represent the possibility of webpages S_i pointing to S_1, $\beta_{M+1} \sim \beta_{M+N}$ represent the possibility of webpages I_j pointing to S_1, and L is the total number of outgoing links. For a normal web page, $\beta_{M+1} = ... = \beta_{M+N} = 0$. In the presence of a search engine, $\beta_{M+1} \sim \beta_{M+N}$ will increase with some possibility, thus causing S_1 to become infected. The PageRank value of S_1 can be calculated as

$$\begin{aligned} PR(S_1)' = &\sum_{i=1}^{M} (\beta_i + \triangle\beta_i) \frac{PR(S_i)}{L(S_i) + q} \\ &+ \sum_{j=M+1}^{M+N} (\beta_j + \triangle\beta_j) \frac{PR(I_j)}{L(I_j) + q}, \end{aligned} \tag{34}$$

where $\triangle\beta_{i(j)}$ are possibilities increased by the search engine. $\triangle\beta_{i(j)}$ has positive correlation of q and k_{S_1} (the degree of S_1). Let $\triangle\beta_i = c_i q k_{S_1}$, where c_i is a positive constant. So we get

$$\begin{aligned} PR(S_1)' = &\sum_{i=1}^{M} (\beta_i + c_i q k_{S_1}) \frac{PR(S_i)}{L(S_i) + q} \\ &+ \sum_{j=M+1}^{M+N} (\beta_j + c_j q k_{S_1}) \frac{PR(I_j)}{L(I_j) + q}. \end{aligned} \tag{35}$$

Comparing every item that votes for S_1 in (35) and (36), we find that when $c_i k_{S_1} > \beta_i L(S_i)$, $PR(S_1)' > PR(S_1)$. $\beta_i L(S_i)$ can be understood as the average number of edges that from S_i to S_1, there must be $\beta_i L(S_i) \leq 1 < c_i k_{S_1}$. So we get $PR(S_1)' > PR(S_1)$, and as q increases, $PR(S_1)'$ increases. This conclusion can be extended to all the web pages S_i. In the presence of a search engine, more webpages are infected and the PageRank value of each webpage increases. On the other hand, more infected webpages and the increase of PageRank make more outlinks become virus propagation paths. Essentially, increasing q can be considered to increase the number of virus propagation paths, so increasing infected outlinks is equal to increasing q. So, we can know that $PR(S_i)' \propto q$.

Let $PR(A)$ denote the expected PageRank value of network A, β_{ji} denote the possibility that one of web page j's outgoing links points to S_i, and $\triangle\beta_{ji}$ denote the possibility increased by the search engine. $\triangle\beta_{ji} \propto q k_{S_i}$, $\beta_{ii} = \triangle\beta_{ii} = 0$, and so we have

$$\begin{aligned} PR(A) = &\frac{1}{M+N} \sum_{i=1}^{M+N} \sum_{j=1}^{M+N} (\beta_{ji} + \triangle\beta_{ji}) \frac{PR(S_j)}{L(S_j) + q} \\ = &\frac{1}{M+N} \sum_{j=1}^{M+N} PR(S_j)'. \end{aligned} \tag{36}$$

Since $PR(S_i)' \propto q$, we know that $PR(A) \propto q$, that is, as the searching ratio q increases, the expected PageRank value of network A increases. So the expected PageRank value can be used as another indicator for the explosion of virus. When the virus propagation undergoes a first-order phase transition, the expected PageRank value of social networks significantly increases.

4 Experiments

Our experiments are based on the real-world data sets from the Slashdot, the DBLP, the P2P, and a simulated data set. We continuously trace the change in network properties, such as the crucial searching ratio, the network stability, the infection density, and the network diameter during the virus propagation process. Moreover, we compare these metrics with and without a search engine to verify the conclusions discussed above: that virus propagation is promoted and the network becomes more vulnerable to virus infection in the presence of a search engine. Finally, we simulate the virus propagation process in different epidemic models to show the superiority of our virus propagation model.

Facebook is a typical social network that includes node features (profiles), circles, and ego networks [8]. Facebook users use the search engine to search for content of interest, thus building social relationships. By observing this data set, it is easy to analyze the effect of a search engine. Facebook consists of $4,039$ nodes and $88,234$ edges.

The P2P data set is a sequence of snapshots of the Gnutella peer-to-peer file [9]. P2P is the earliest social network designed for sharing files. P2P users upload files and make them easily found by search engines. There are nine snapshots of the Gnutella network collected in this data set. Nodes represent hosts and edges represent connections. P2P consists of $62,586$ nodes and $147,892$ edges.

Slashdot is a technology-related news website known for its specific user community [10]. In the presence of a search engine, users can search for news without being limited by community structure. The website features user-submitted and editor-evaluated current, primarily technology-oriented, news. In 2002 Slashdot introduced

the Slashdot Zoo feature, which allows users to tag each other as friends or foes. The network contains friend/foe links between the users of Slashdot. The network was obtained in November 2008. Slashdot consists of $77,360$ nodes and $905,468$ edges.

The DBLP data set provides a comprehensive list of research papers in computer science [11]. It constructs a co-authorship network where two authors are connected if they publish at least one paper together. There is abundant communication among co-authors, and this can be done through a search engine. DBLP consists of $317,080$ nodes and $1,049,866$ edges.

4.1 Critical Searching Ratio

We change the searching ratio of the network users, and observe the infection density when the infection reaches a stable state I_∞ as the initial infection ratio $1-p$ increases. The searching ratio p_{SE} is set from small to large. The numbers 0∼4 represent different levels of the searching ratio, where the level is 0 if there is no search engine, and 4 if the search engine is widely used by the users in social networks.

Fig. 4 shows that, as the initial infection ratio $1-p$ increases, the stable infection density I_∞ increases and finally reaches a stable state. Specifically, a higher searching ratio leads to a higher stable infection density. And the search engine makes the virus propagation process reach the stable state faster. Moreover, when the searching ratio is small, I_∞ increases continuously with $1-p$. When the searching ratio is large, for example, $p_{SE}=4$, with the increase of $1-p$, a jump phenomenon appears at p_c in which I_∞ changes discontinuously from 0 to a large value.

The experiment results show that wide use of the search engine causes a greater number of nodes to become infected. Specifically, the jump phenomenon shows that the search engine significantly promotes the virus propagation process and can lead to a large-scale infection. The effect of the search engine can be particularly evident in social networks, for a small $1-p$ may lead to large-scale infection. When there is no search engine, different users in separated communities are much less likely to connect, and the virus is limited to the communities where the virus initially appears. The situation is completely different in the presence of a search engine, where users can employ the search engine to search for web pages in any community, and thus the virus can infect the whole network.

4.2 Network Stability Experiment

The functional node density can be used to measure network resilience. We mainly focus on the functional node density of normal users at the end of the virus cascade process. We fix the searching ratio, and observe the functional node density when the infection reaches a stable state as the initial infection ratio $1-p$ increases.

As shown in Fig. 5, as $1-p$ increases, the functional node density of both networks decrease. In the presence of a search engine, the functional node density is always lower than that in networks without search engine.

This phenomenon proves that the presence of a search engine makes social networks more vulnerable to a virus infection. As more and more nodes are infected, the functional node density of normal users decreases. In the presence of a search engine, the functional node density decreases faster, and the structure of the network becomes unstable. As a result, the search engine cuts down the network's stability.

4.3 Infection Density Experiment

We change the searching ratio of the network users, and observe the infection density as time goes. The searching ratio p_{SE} is set from small to large. The numbers 0∼4 represent different levels of the searching ratio, where the level is 0 if there is no search engine, and 4 if the search engine is widely used by the users in social networks. We fix the initial infection ratio $1-p=0.01$.

As shown in Fig. 6, the infection densities with and without the search engine increase over time and finally reach stable states. In the presence of a search engine, the infection density is always higher than in networks without a search engine.

The experiment results show that the search engine increases infection density and expands the virus propagation area significantly. When there is no search engine, virus propagation is limited to the communities where the virus first appears.

4.4 Network Diameter Experiment

We change the searching ratio of the network users, and observe the average network diameter. The searching ratio p_{SE} is set from small to large. In order to measure the average network diameter more easily, we regard the shortest path between two unconnected nodes as a large constant.

As shown in Fig. 7, network diameter decreases as the searching ratio increases. By analyzing the results, we conclude that a search engine cuts down the network diameter in social networks and that the higher the searching ratio is, the lower the diameter is. As an intuition of the decrease of the network diameter, social network users contact more tightly with each other. The virus propagation consequently becomes faster and wider.

Network diameter is used to analyze information propagation, which is the key function of social networks. Nodes gather together based on the same interest. This process makes the information propagation regionalized, and can assist in the defense against a virus to some extent. In the presence of a search engine, the nodes can search information across different regions, thus increasing the propagation paths of the virus. As a result, the network diameter is decreased faster and the virus propagation speed is increased.

4.5 Superiority of Our Model

In this section, we first simulate three classical epidemic models on Matlab, namely SI [12], SIS [13], and SIR [5]. Then, we compare these models with our virus propagation model to show the superiority of our model. We set the infected individual to pass the infectious disease to the susceptible individual with a probability of 0.3. The infected

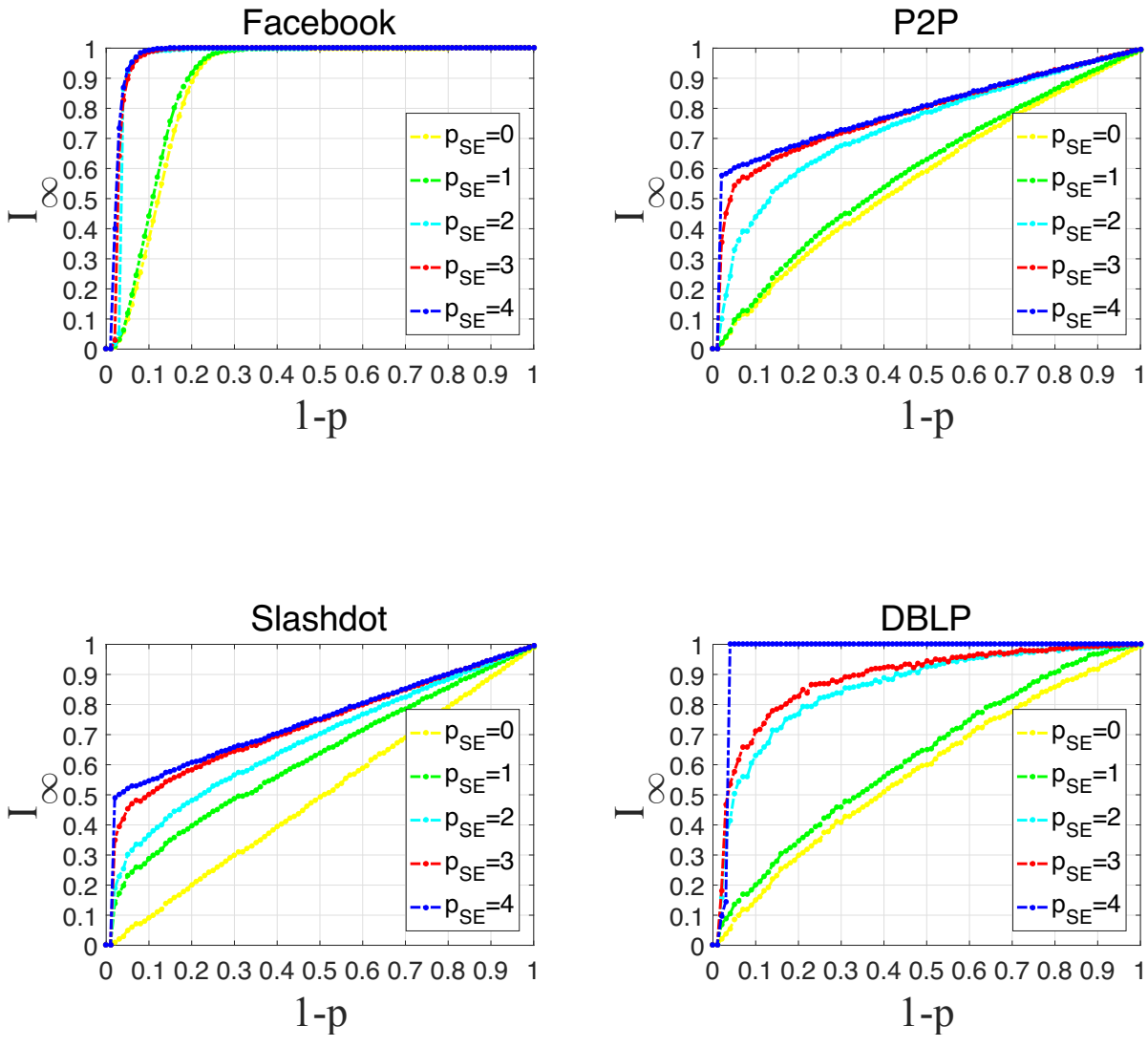
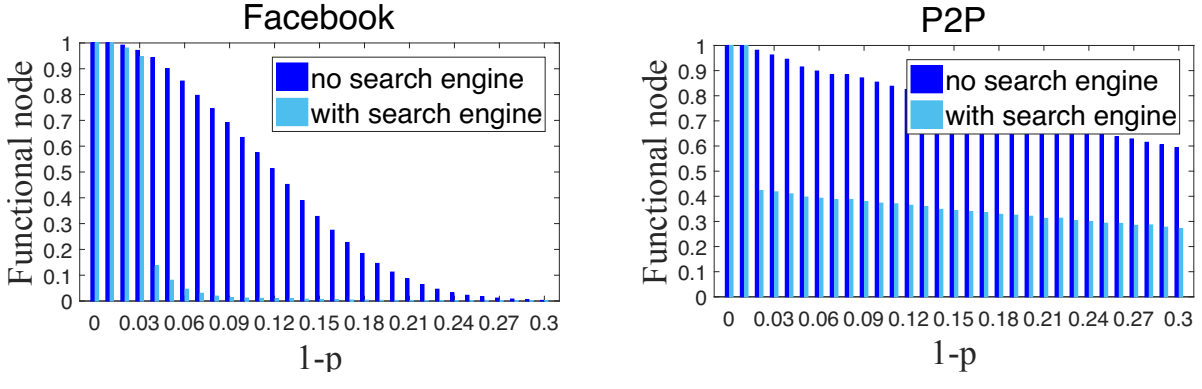


Fig. 4: Infection density with different searching ratio in different networks. (The searching ratio p_{SE} is set from small to large. The numbers 0∼4 represent different levels of the searching ratio, where the level is 0 if there is no search engine, and 4 if the search engine is widely used by the users in social networks.)



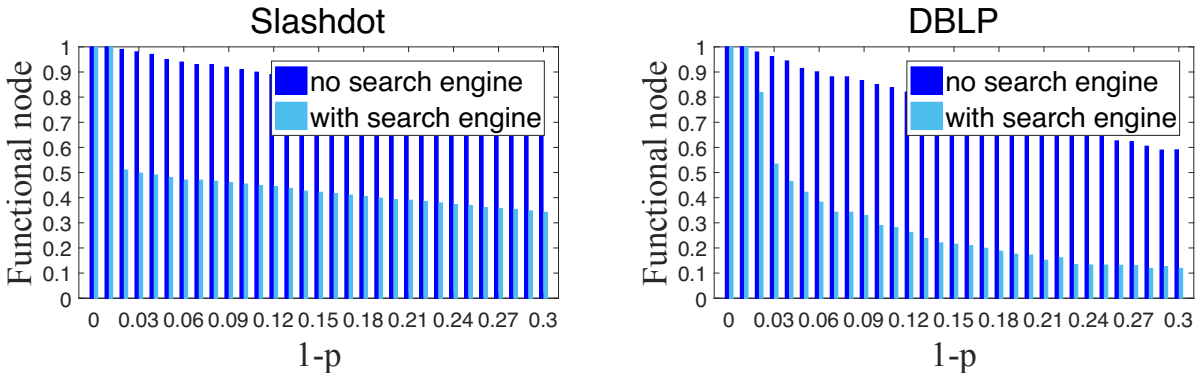


Fig. 5: Functional node density with and without a search engine in different networks.

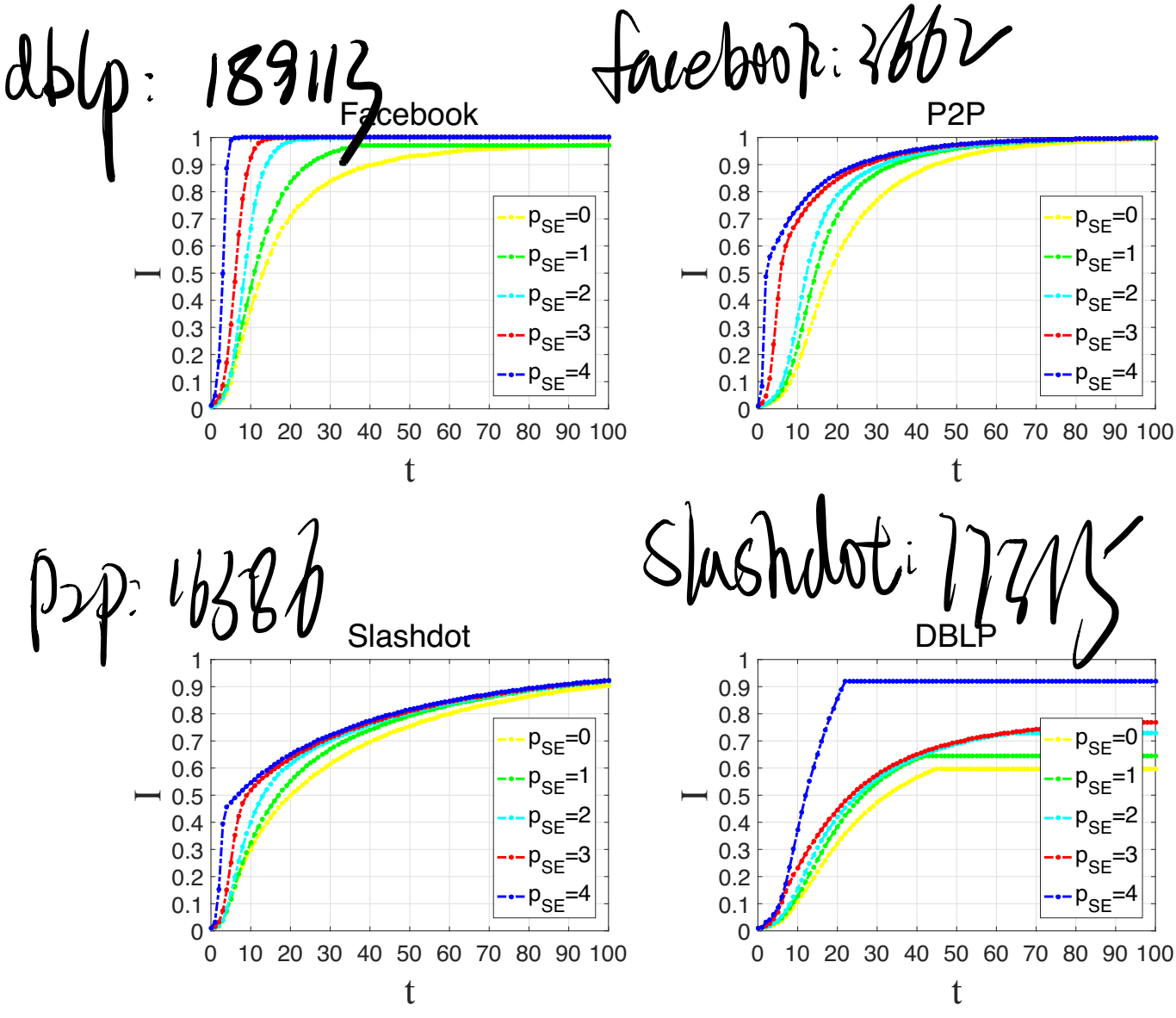


Fig. 6: Infection density with different searching possibilities in different networks. (The searching ratio p_{SE} is set from small to big. The numbers 0∼4 represent different levels of the searching ratio, where a level is 0 if there is no search engine, and 4 if the search engine is widely used by the users in social networks.)

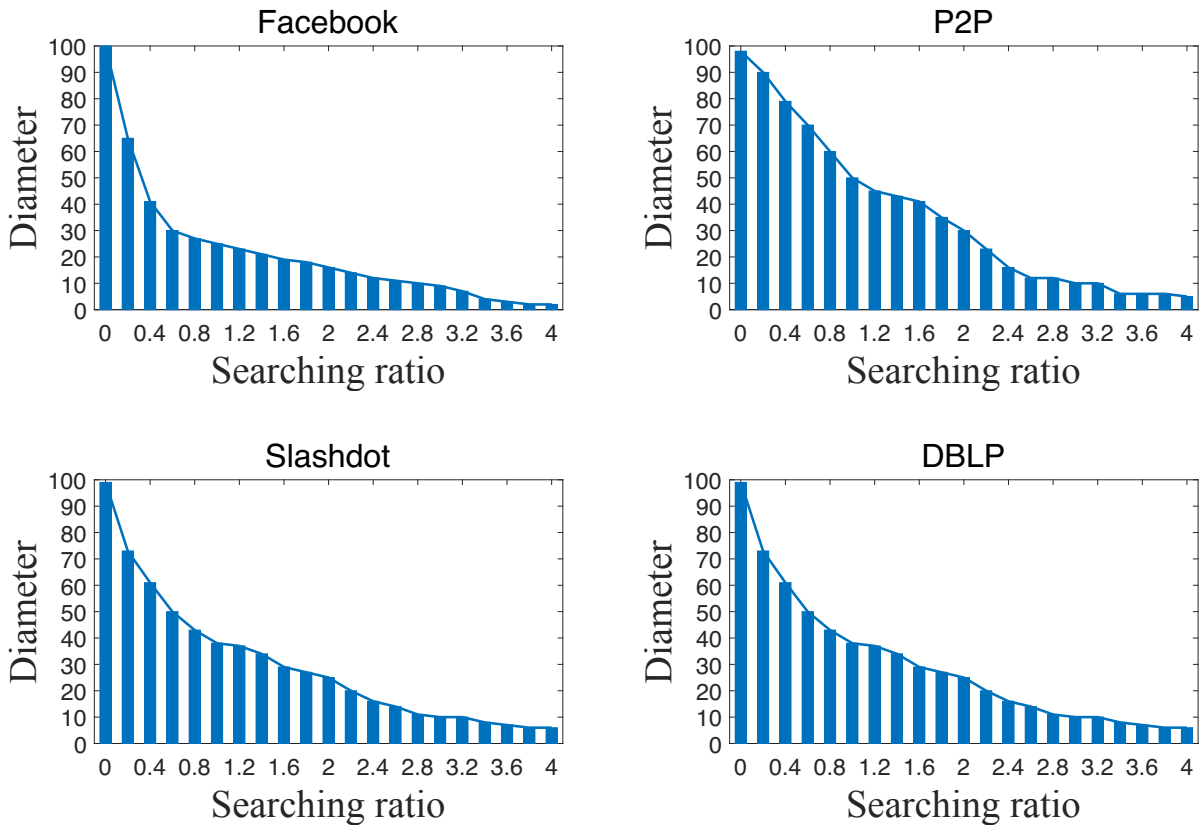


Fig. 7: Average network diameter with different searching possibilities in different networks.

individual in the SIS was cured with a probability of 0.1, and the infected individual in the SIR was converted to the removed state with a probability of 0.1, and the initial infection rate was $1 - p = 0.1$.

The crucial searching ratio experiment above demonstrates that I_∞ increases with the initial infection ratio $1-p$. Moreover, when the searching ratio is small, I_∞ increases continuously with $1 - p$. When the searching ratio is large, with the increase of $1 - p$, a jump phenomenon appears in which I_∞ changes discontinuously. In the simulated SI model, SIS model, and SIR model, we donnot observe this phenomenon. The SI model, SIS model, and the SIS model show as a straight line (Fig. 8). In real-world situations, the community structure may limit the virus propagation, so when the infection reaches a stable state, the infection density does not always remains unchanged. The three classical models cannot describe the first-order phase transition and the second-phase transition during virus propagation in real world.

Considering computer viruses, anti-virus software can isolate or delete infected files, but there is no permanent anti-virus program that can make computers immune to the same virus. It must be combined with operating system patches to avoid re-infection, so SIS model is suitable to describe computer viruses. But the SIS model does not take into account of search engines. When the parameters are the same, the changing rate of infection density with time is higher in the presence of search engine, and the infection density is higher when stable (Fig. 9).

In summary, the superiority of our virus propagation model can be summarized in three points:

- Many new characteristics appear as a result of the presence of a search engine. Our virus propagation model emphasizes the effects of the search engine, in contrast to current research, where specific propagation effects with a search engine are ignored. Moreover, we further discuss the relationship between the searching ratio and the Google PageRank algorithm, thus making our model more understandable and closer to the truth.
- Our model is rooted in percolation theory, which is widely used in many research areas. It initially comes from thermodynamics in which the van der Waals equation predicts a liquid-gas first-order phase transition line ending at a critical point characterized by a second-order transition. To our knowledge, we are the first to apply the percolation method to describe the virus propagation process in social networks with a search engine; we find that it becomes much easier to analyze the virus propagation characteristics using this approach.
- Our virus propagation model effectively describes the "jump phenomenon" of the infection density as the initial infection ratio increases. Studying the "jump phenomenon" is of great significance in defending against virus propagation. By controlling the initial infection ratio below a crucial value or controlling the number of immune nodes above a crucial value, we can avoid a large-scale virus outbreak. Most of the current research has not paid much

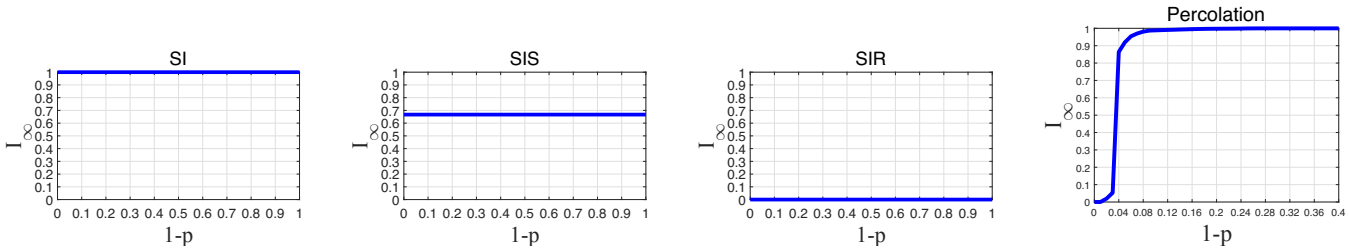


Fig. 8: Stable Infection density in different models.

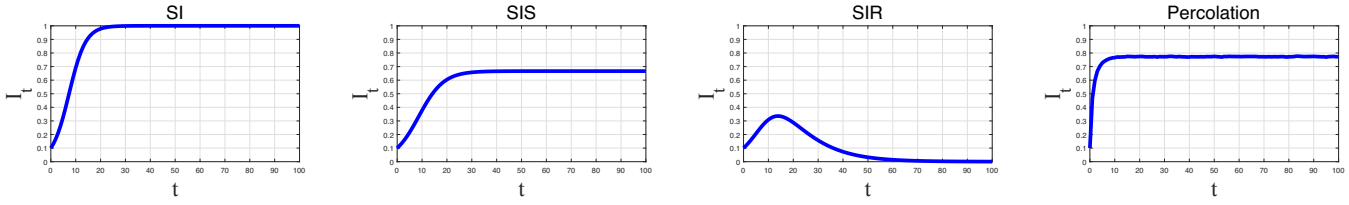


Fig. 9: Infection density in different models.

attention to this phenomenon.

5 Related Work

In recent years, the social network has become an important network. Social networks can be applied in many different fields, ranging from smart industrial applications to agricultural optimization to the intelligent planning of cities, health services, energy consumption, and environmental protection [14]. The search engine, which is primarily designed so that social network users can acquire interesting information, is becoming more tightly coupled with social networks as the volume of users increases [15]. However, search engines serve as a "double-edged sword." On one hand, the search engine assists with information dissemination. On the other, it exacerbates security threats caused by viruses. Search engines supply a highly effective means of information retrieval. But a search engine is also a platform for spreading information. Because of these features, the propagators of malicious code have kept in step with search engines. For example, the search engine poisoning (SEP) [16] [17] was applied by malicious softwares that published vicious and fake pages to push the page ranking higher and attract more accesses. With the development of social networks, attackers are writing special code to attack others and spread the virus via search engines.

The virus propagation power of the search engine has been recognized recently. M Egele et al. [18] have pointed out that attackers use different techniques to boost their pages to higher ranks, since a better position in the rankings directly and positively affects the number of visits to a site. In the worst case, these pages pose a threat to Internet users by hosting malicious content and launching drive-by attacks against unsuspecting victims. When successful, these drive-by attacks then install malware on the victims' machines. E. Levy [19] has discovered that worms transmitted via email have had great success propagating themselves because they find their next targets either by raiding a user's email address book or by searching through the user's mailbox. Such addresses are almost certain to be valid, permitting the worm to hijack the user's social web and exploit trust relationships.

Much of the work focusing on defense against virus propagation has been based on virus features and network environments. Various virus propagation models make it simple and convenient to analyze the virus propagation features in various scenarios. G. Yan et al. [20] present a detailed analytical model characterizing the propagation dynamics of Bluetooth worms which captures not only the behavior of Bluetooth protocol but also the impact of mobility patterns on the Bluetooth worm propagation. Based on the characteristics of an epidemic virus, Zhiqiang et al. [21] propose a VEIS epidemic virus propagation model based on partial immunization of vulnerable people. This model analyzes the existence and stability of equilibrium points by using the mean-field theory and qualitative theory of differential equations, which turns out to be a stable and effective model. Sellke et al. [22] present a (stochastic) branching process model for characterizing the propagation of Internet worms. The model is developed for uniform scanning worms and then extended to preference scanning worms. This model leads to the development of an automatic worm containment strategy that prevents the spread of a worm beyond its early stage. With respect to network structures, the research work in [23] [13] indicates that network topology has an important impact on malware propagation. In [23], a virus model is presented based on an epimedological model that predicts how a virus propagates in a P2P network; this work also addresses issues related to virus propagation in a P2P network. In [13], a modified SIS epidemic model is proposed to study the dynamics of virus spread in wireless sensor networks. The sensor nodes are attacked by viruses and initially only a small number of nodes are infected. The virus spreads itself to its neighbor nodes by piggybacking on normal data via regular communications. The infected neighbor nodes repeat the process to their respective neighbour nodes in turn. Each sensor node is installed with an antivirus program, which can be periodically refreshed to recover the infective nodes

to the susceptible group. Cai Fu et al. [27] point out that a search engine generates a propagation wormhole effect by delivering virtual virus sources to each community and creating short-cuts between distant nodes. They design a positive feedback model and determine quantitatively how the search engine influences virus propagation using virus propagation dynamics.

Although virus propagation has received more research attention recently, and many researchers have recognized the virus propagation power of the search engine and proposed various models to calculate the virus propagation process, the quantified relation between the search engine and virus propagation is still largely unexamined. If we can fully understand the relationship between the search engine and virus propagation, determine the propagation threshold, and build the search engine propagation model, it will help to master the spreading principle and the degree of damage, develop defensive policies, and even optimize network structure to prevent propagation. In this paper, we attempt to fill this research gap in virus propagation.

The research results of Buldyrev et al. [24] have inspired our work. They have pointed out that modern systems are coupled together. Failure of nodes in one network may lead to failure of dependent nodes in other networks. This may happen recursively and can lead to a cascade of failures. This scene is quite similar to the virus propagation process in which network nodes are infected by their infected neighbors, and the newly-infected nodes go on infecting their neighbours. This process continues until the virus propagation reaches a stable stage. In the presence of a search engine, nodes connect with each other by the search engine with some likelihood. Consequently, when nodes are infected by a virus they cause other nodes to also become infected without being limited by the community structure. If the users that use the search engine change over time, the initial small-scale infection may trigger a recursive process of virus spread that can completely cover the whole network.

Our virus propagation model is rooted in percolation theory, which is widely used in many research areas. It initially comes from thermodynamics, in which the van der Waals equation predicts a liquid-gas first-order phase transition line ending at a critical point characterized by a second-order transition [25]. Similarly, interactions between networks give rise to a first-order percolation phase transition line that changes at the critical point to a second-order transition, as the coupling strength between the networks is reduced [3]. Unlike the current research, we study the virus propagation in the presence of a search engine based on percolation theory. We reveal that increasing the searching ratio leads to a change from a second to first percolation transition in the virus propagation process.

6 Conclusion

In the past few years, a wide variety of viruses have caused countless security threats. In the presence of a search engine, the situation becomes even worse, since a search engine assists the virus dissemination for social networks. Studying the characteristics of virus propagation in the presence of a search engine is therefore of significance. However, most traditional virus propagation models, such as SI, SIS, and SIR, ignore the greater number of propagation paths added by a search engine and the positive feedback effect brought by a search engine.

In this paper, we reveal that virus propagation in social networks is driven by the search engine, and propose a virus percolation model that takes into the search engine into consideration. Based on the model, we reveal that increasing the searching ratio leads to a change from a second to first percolation transition in the virus propagation process. Furthermore, we calculate the crucial searching ratio in both a Poisson distribution network and a power law distribution network. Then, we quantitatively characterize how a search engine influences the metrics of the virus propagation process, such as the infection density, network diameter, jump point, and giant components. We mathematically and theoretically prove that the search engine increases the infection density, shrinks the network diameter, and makes a social network more vulnerable. Moreover, we introduce the Google PageRank algorithm to analyze the positive feedback effect of the search engine.

In the future, we will continue to explore the specific effects of the search engine by studying more characteristics of virus propagation. At the same time, we will further study methods for applying our research results in practice, particularly in the area of virus defense, our ultimate objective.

References

[1] "Facts about Google and Competition", Archived from the original on 4 November 2011, Retrieved 12 July 2014.

[2] Brin S and Page L, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems 30: ISSN 0169-7552, doi: 10.1016/S0169-7552(98)00110-X, 1998.

[3] Roni Parshani, Sergey V. Buldyrev, and Shlomo Havlin, "Interdependent networks: reducing the coupling strength leads to a change from a first to second order percolation transition," in *American Physical Society*, 2010.

[4] Wenzhi Chen, "A mathematical model of ebola virus based on SIR model," in *2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration*, 2015.

[5] M. E. J. Newman, "Spread of epidemic disease on networks," in *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 2002.

[6] J. Shao, S.V. Buldyrev, R. Cohen, M. Kitsak, S. Havlin, and H. E. Stanley, "Fractal boundaries of complex networks," in *Epl*, 2008, 84 (4): 605-609.

[7] Cai Fu, Chenchen Peng, and Xiao-Yang Liu, "Search engine drives the evolution of social networks," ACM TUR-C '17: Proceedings of the ACM Turing 50th Celebration Conference - China, 2017.

[8] Facebook: https://snap.stanford.edu/data/egonets-Facebook.html.

[9] P2P: http://snap.stanford.edu/data/p2p-Gnutella31.html.

[10] Slashdot: http://snap.stanford.edu/data/soc-Slashdot0811.html.

[11] DBLP: https://snap.stanford.edu/data/com-DBLP.html.

[12] V. M. Eguiluz and K. Klemm, "Epidemic threshold in structured scale-free networks," in *Physical review letter*, 2002.

[13] S. Tang, D. Myers, and J. Yuan, "Modified SIS epidemic model for analysis of virus spread in wireless sensor networks," IJWMC, vol. 6, no. 2, 2013.

[14] Angela Ullrich, "The Social Internet of Things: When the forest calls the fire department," in *Social Entrepreneurship BMW Foundation*, April 04, 2016.

[15] Facebook: number of daily users worldwide: http://www.statista.com

[16] L. Lu, R. Perdisci, and W. Lee, "SURF: detecting and measuring search poisoning," in *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS 2011, Chicago, Illinois, USA, October 17-21, 2011, Y. Chen, G. Danezis, and V.

Shmatikov, Eds. ACM, 2011, pp. 467-476. [Online]. Available: http://doi.acm.org/10.1145/2046707.2046762.

[17] N. Leontiadis, T. Moore, and N. Christin, "A nearly four-year longitudinal study of search-engine poisoning," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, Scottsdale, AZ, USA, November 3-7, 2014, G. Ahn, M. Yung, and N. Li, Eds. ACM, 2014, pp. 930-941. [Online]. Available: http://doi.acm.org/10.1145/2660267.2660332.

[18] M. Egele, C. Kolbitsch, and C. Platzer, "Removing web spam links from search engine results," *Journal in Computer Virology*, vol. 7, no. 1, pp. 51-62, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11416-009-0132-6

[19] E Levy ,"Worm propagation and generic attacks," *IEEE Security and Privacy*, 2005,3 (2): 63-65.

[20] G. Yan and S. Eidenbenz, "Modeling propagation dynamics 1367 of bluetooth worms," *IEEE Trans*, Mob. 1368 Comput, vol. 8, no. 3, pp 353-368, 2009.

[21] Zhiqiang Sun, Liang Chen, and Qiaoling Chen ,"A VEIS computer virus propagation model based on partly immunization," *ICIIP'16 Proceedings of the 2016 International Conference on Intelligent Information Processing*, Article No. 34, December 2016.

[22] S. H. Sellke, N. B. Shroff, and S. Bagchi, "Modeling an automated containment of worms," *IEEE Trans*, Dependable Sec. Comput, vol. 5, no. 2, pp. 71-86, 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TDSC.2007.70230.

[23] Mansoor Ebrahim, S. M. Umar Talha, and Jahanzeb Ahmad, "Modeling virus propagation in pure peer-to-peer networks," *FIT '10: Proceedings of the 8th International Conference on Frontiers of Information Technology*, December 2010.

[24] Sergey V. Buldyrev, Roni Parshani, Gerald Paul, H. Eugene Stanley, and Shlomo Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, Vol 464|15, April 2010, doi: 10.1038/nature08932.

[25] Shlomo Havlin, H. Eugene Stanley, Amir Bashan, Jianxi Gao, and Dror Y. Kenett, "Percolation of interdependent network of networks," *Chaos, Solitons and Fractals*, 2015.

[26] Wasserman, Stanley, Faust, and Katherine, "Social network analysis in the social and behavioral sciences," in *Social Network Analysis: Methods and Applications*, Cambridge University Press. pp. 1âĂŞ27. ISBN 9780521387071, 1994.

[27] Cai Fu, Xiao-Yang Liu, Jia Yang, Laurence T. Yang, Shui Yu, and Tianqing Zhu, "Wormhole: The hidden virus propagation power of a search engine in social networks" *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2015.