

From the Dialectical Perspective: Modeling and Exploiting of Hybrid Worm Propagation

Tianbo Wang^{ID}, *Member, IEEE*, Huacheng Li, Chunhe Xia^{ID}, Han Zhang^{ID}, *Member, IEEE*, and Pei Zhang^{ID}

Abstract—The hierarchical network is the more effective platform, which provides multiple channels for various worm propagation. Thus, emerging worms can infect vulnerable hosts by scanning strategy and social media. However, the spread of scan-based worm is restrained due to uneven distribution of vulnerable hosts and NAT (Network Address Translation) technique. Meanwhile, topological dependency dictates to topology-based worm only infecting those hosts in social networks. To avoid their respective disadvantages, modern hybrid worm, which combines the above two propagation mechanisms, can implement efficient IP-address scanning by *enhanced combination-scanning* strategy, and spread more aggressively in social networks using *enhanced reinfection* mechanism. This paper presents a Hierarchical-Stochastic Propagation model to understand hybrid worm propagation. Inspired by hybrid worm, we design a new vaccine based on the Hierarchical-Measure Immunization strategy. For physical networking layer, we can estimate vulnerable-host distribution to find vulnerable hosts effectively through Maximum Likelihood estimation. For social networking layer, we use a novel propagation centrality measure to discover vital social nodes accurately. The experimental results show that our model can characterize the propagation mechanism of hybrid worms more comprehensively, and greatly outperforms state of the art models in terms of estimation accuracy. Meanwhile, our strategy is more effective to restrain the hybrid worm from spreading in networks.

Index Terms—Hybrid worms, propagation dynamics, defense strategy, modeling.

I. INTRODUCTION

HYBRID worms and their variants are at the top of network security threats due to their high infectivity and destructiveness [1]. And then they lead to a series of

problems such as network downtime, financial loss, and information disclosure. Compared with traditional scan/topology-based worms, hybrid worms (e.g., Nimda, Stuxnet and Flame) can utilize multiple approaches to hit victims through Email, probing vulnerable hosts and USB devices and so on. Generally, hybrid worms spread throughout the hierarchical network. For social networking layer, malware utilizes the friends' relationship to spread hop-by-hop through mutual communication. For physical networking layer, compromised hosts utilize efficient scanning strategies to discover system vulnerabilities and implant malicious codes.

A. Motivation

Modern hybrid worm is a more destructive and contagious malware, which combines efficient scanning strategies with social networks. Thus, **this kind of worm becomes a new trend for attackers** from two perspectives: (1) *Network heterogeneity for promoting worm mutation*. Vulnerable hosts are not uniformly distributed over the IPv4 space. For example, Fig. 1(a) indicates that the vulnerable-host distribution follows a power law. According to the set of unique scanning sources from CAIDA's /8 network telescope, Fig. 1(b) shows that there is a more pronounced heavy-tailed effect for infected population distribution. The above results clearly indicate that the vulnerable population distribution is far from being uniformly distributed. Additionally, with the help of the NAT technology, the study [2] shows that 60% of the hosts infected by Code Red II have IP masquerading through private network addresses. Fig. 1(c) indicates that the NAT technique and the non-uniform distribution can restrain worm propagation. Consequently, hackers consider the combined strategy (i.e., *enhanced combination-scanning*) to improve the efficiency of scan-based worms, as shown in Section III-B. Specifically, importance sampling can select subnets with more vulnerable hosts based on infection information, and local-preference-based scanning strategy can make worms quickly spread in the same NAT subnet. However, the NAT mechanism makes some victims "invisible" to inter-area hosts, which reduces the efficiency of scanning worms spreading across different NAT subnets. Thus, topology-based worms spread through social relationships to solve the NAT blocking problem. For example, the email-based method with *enhanced reinfection* mechanism, is more aggressive spreading in social networks [3]. However, due to the dependence on the social topology, all propagation paths must be subgraphs of the social

Manuscript received 8 November 2021; revised 10 July 2022, 3 November 2022, and 17 January 2023; accepted 9 February 2023. Date of publication 20 February 2023; date of current version 27 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61902013 and in part by the Beihang Youth Top Talent Support Program under Grant YWF-22-L-1038. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thorsten Strufe. (Corresponding author: Tianbo Wang.)

Tianbo Wang is with the School of Cyber Science and Technology, Beihang University, Beijing 100191, China, and also with the Shanghai Key Laboratory of Computer Software Evaluating and Testing, Shanghai 201112, China (e-mail: wangtb@buaa.edu.cn).

Huacheng Li and Chunhe Xia are with the Beijing Key Laboratory of Network Technology, Beihang University, Beijing 100191, China (e-mail: leehc@buaa.edu.cn; xch@buaa.edu.cn).

Han Zhang is with the INSC&BNRist, Tsinghua University, Beijing 100084, China (e-mail: zhhan@tsinghua.edu.cn).

Pei Zhang is with the School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhangpei@bupt.edu.cn).

Digital Object Identifier 10.1109/TIFS.2023.3246765

1556-6021 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

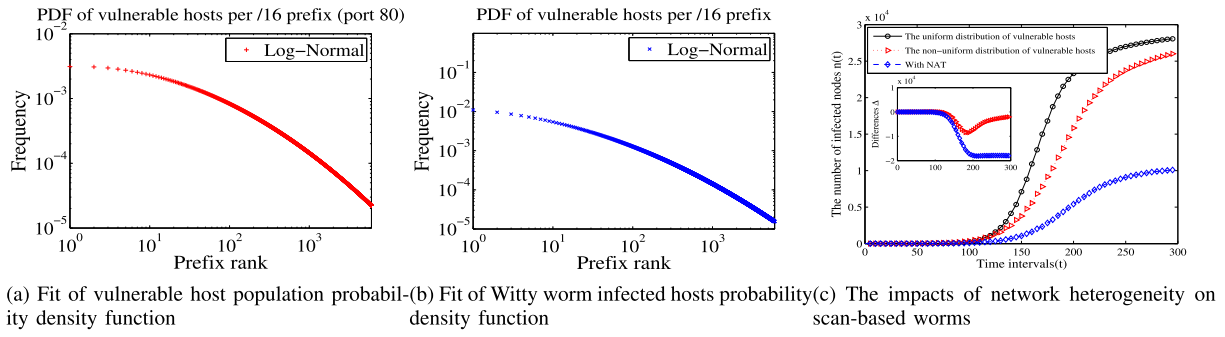


Fig. 1. Nonuniform distribution of web servers and their impacts on worm propagation. In the subfigure (c), All cases are for $N = 29,000$, $s = 5$, 5 autonomous systems, 22 NAT subnets. Experimental environment setup is used in Section V-A. Note: users and hosts follow the one-to-one correspondence relationship.

topology. Therefore, to take advantage of two kinds of worms, the hybrid worm is introduced in Section II-A. (2) *Hybrid propagation for improving infection efficiency*. Modern hybrid worm can spread through various propagation mechanisms. Worm.Nimda can send malicious codes by searching email addresses from email clients and HTML files and infect IIS servers by scanning specific system vulnerabilities randomly. Worm.Viking.m can spread over the Internet by means of email attachments with malware and infect vulnerable hosts by sending ICMP probe packets. We can see that hybrid worm variants have the common characteristic that they spread in the hierarchical network through scanning vulnerable resources and sending malicious emails. Additionally, *imprecise models for hybrid worm propagation*. Existing models can be roughly divided into both scan-based models [4], [5], [6], [7] and topology-based models [8], [9], [10], [11]. The former results in a “homogenous mixing” phenomenon by differential or difference methods with the mean-field theory and fails to analyze individual dynamics. Meanwhile, the latter only characterizes the worm propagation based on the contact relationship, and cannot characterize the infection process of scanning behavior between non-neighbors. Thus, those models focus on the one-sided modeling of traditional worms in the physical network or the social network, and cannot characterize hybrid worm propagation.

On the other hand, we can see that the hybrid worm has the *strong propagation ability and executes a task automatically in accordance with attacker's intention*. Meanwhile, the infected server and clients can finish the learning task cooperatively by exchanging information. Thus, we should develop the corresponding vaccine to find vulnerable hosts and vital users quickly and effectively in the hierarchical network. It provides support for patching target hosts and restraining malicious codes dissemination.

In this paper, to address those aforementioned problems, we present a **Hierarchical-Stochastic Propagation (HSP)** model for understanding modern hybrid worm. Meanwhile, inspired by the hybrid worm propagation, we present a novel vaccine by exploiting a **Hierarchical-Measure Immunization (HMI)** strategy. It can identify vulnerable hosts and vital social nodes to improve the robustness of the target network. Thus, our work can predict the propagation trend and suppress the potential threat from the hybrid worm in reality.

B. Contributions

The major contributions can be summarized as follows:

- We present a novel HSP model for analyzing hybrid worm propagation. This model characterizes the propagation dynamics of multi-behaviors (i.e., combination of scanning propagation and social network propagation).
- We present an HMI defense strategy based on the propagation model. This method effectively immunizes the target network through estimation of vulnerable-host group distribution and measure of vital social nodes.
- The experimental results verify the effectiveness of our HMI defense strategy and the accuracy of our HSP model. And we can understand the hybrid worm and the vaccine clearly and comprehensively.

The remainder of this paper is organized as follows: Section II describes the problems in characterizing the hybrid worm and the vaccine from the propagation and defense perspectives, respectively. And then we propose the HSP model for hybrid worms in Section III. Meanwhile, Section IV explains the details of the HMI defense strategy. In Section V, the experiments validate the rationality of the HSP model and examine the defense effects of the HMI strategy. Related work is summarized in Sections VI. We give the conclusion in Section VII.

II. PROBLEM STATEMENT

Fig. 2 is the illustration of modern hybrid worm propagation. Suppose that there is a network that consists of g groups (i.e., a target **group** means a routing prefix or a network prefix corresponding to specific IP address space). For group m , the prefix of subnet by global routing is $/\rho_m$, the IP address scale can be expressed as $\Omega_m = 2^{32-\rho_m}$. Additionally, group m contains n NAT subnets with different prefixes $/\rho_{n|m}$, and the scale of each NAT subnet is $\Omega_{n|m} = 2^{32-\rho_{n|m}}$. This paper gives major notations in Table I.

In this work, we first consider that modern Internet structure has heterogeneity, which indicates vulnerable hosts are non-uniformly distributed over target networks, and NAT technology is widely used in the Internet. Second, we assume that a probability distribution of vulnerable hosts is obtainable. That is to say, some Bots in a Botnet have a high Internet bandwidth capacity and routable IP addresses in the public

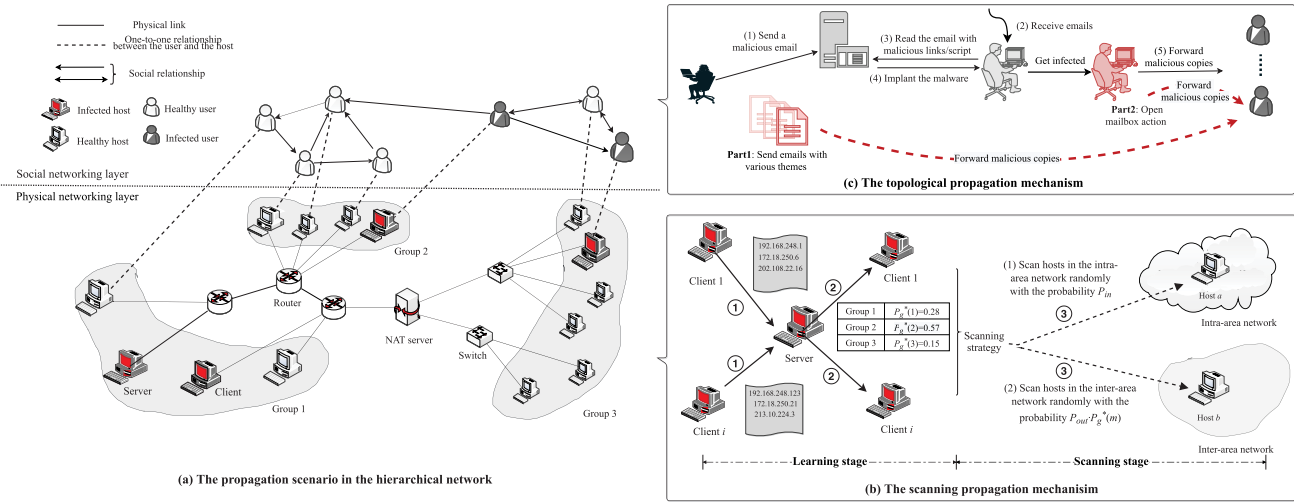


Fig. 2. Illustration of hybrid worm spreading in the real world. Subfigure (a) illustrates the structure of the hierarchical network. Subfigure (b) illustrates the enhanced combination-scanning mechanism in physical network layer. Subfigure (c) illustrates the enhanced reinfection mechanism in social network layer. Note: social networking layer and physical networking layer mean two different propagation vectors, respectively, and they are different from the hierarchical structure in the OSI model.

TABLE I
MAJOR NOTATIONS USED IN THIS PAPER

Symbol	Explanation
N	Number of susceptible hosts in hierarchical networks.
g	Number of groups in hierarchical networks.
N_m	Number of susceptible hosts in group m .
$N_{n m}$	Number of susceptible hosts in NAT subnet n of group m .
$N_{n m}^t$	Number of susceptible hosts in NAT subnet n of group m at time t .
Ω	Number of hosts in hierarchical networks.
Ω_m	Number of global routing hosts in the group m .
$\Omega_{n m}$	Number of hosts in NAT subnet n of group m .
$\Omega_{/\rho}$	Number of hosts in the $"/\rho$ prefix intra-area network.
s	Number of an infected host sending scans per unit time.
$P_g(m)$	Actual vulnerable-host proportion in group m , i.e., $\frac{N_{n m}}{N_m}$.
$\overline{P_g(m)}$	Estimated vulnerable-host proportion in group m , i.e., $\frac{\overline{N_{n m}}}{\overline{N_m}}$.
I_t	Number of infected hosts in hierarchical networks at time t .
I_m^t	Number of global routing infected hosts in group m at time t .
$I_{n m}^t$	Number of infected hosts in NAT subnet n of group m at time t .
U_m	The set of global routing hosts in group m .
$U_{n m}$	The set of hosts in NAT subnet n of group m .
U_i^{-1}	The host set except host i in hierarchical networks.

network. Thus, we allow them to act as worm servers to contact infected hosts and collect data about the knowledge of the group distribution. Third, worm codes can be customized for specific tasks by the attacker. Additionally, we assume that our required global knowledge of connectivity information can be obtainable/available. Finally, we assume that our proposed immunization strategy is suitable for common user scenario regardless of special scenario (e.g., healthcare and industrial control). Meanwhile, our proposed vaccine has the characteristics of controllability, confrontation, openness and security

A. Problem From a Propagation Perspective

Scan-based propagation behavior: As shown in Fig. 2(b), the propagation details in physical network layer are as follows:

- **Learning stage.** In phase ①, All infected hosts (clients) adopt the uniform strategy to scan and collect IP

addresses of susceptible hosts, and send address records back to the worm server until meeting sample size requirements. In phase ②, group distribution $P_g(m)$ is estimated by the worm server through the collected data. Finally, group scanning distribution $P_g^*(m)$ obtained by $P_g(m)$ is delivered to each infected host. It means that group scanning distribution is delivered by the worm server. The main task of the worm server is to collect IP addresses and calculate group scanning distribution, and the purpose is the later infected hosts can scan IP-address space faster by taking advantage of the knowledge of vulnerable-host distribution.

- **Scanning stage.** Once receiving the distribution $P_g^*(m)$, all infected clients adopt the enhanced combination-scanning strategy as follows: in phase ③, they will scan susceptible hosts of the local (i.e., intra-area) network and global (i.e., inter-area) network with probabilities p_{in} and $p_{out} = 1 - p_{in}$, respectively. For the local scanning, all hosts in the prefix $/\rho_{n|m}$ or $/\rho_m$ network are compromised with the same likelihood. For the global scanning, each susceptible host in the $/\rho_m$ network is hit by infected hosts based on the probability of $P_g^*(m)$.

Topology-based propagation behavior: As shown in Fig. 2(c), the propagation details in social network layer are as follows: the attacker sends malicious emails to some accounts on the target mail server. When there are new messages in the mailbox, the user will check and read them. If the user accesses their attachments with malicious links/script, the victim is implanted by the malware. Subsequently, the victim will forward malicious copies through the contact list. In the enhanced reinfection mechanism, there are two parts to improve propagation capacity. *Part 1* indicates that the malicious email exploits various themes to trick users into read them, and then increase their infection possibility. *Part 2* indicates that the victim can forward malicious copies by opening the mailbox except the traditional infection behavior of accessing malicious attachments.

B. Problem From a Defense Perspective

Modern Internet structure has heterogeneity, which indicates network services and vulnerable hosts are highly non-uniform, and it means that all intervals of the same length on the distribution's probability are not equally probable. This implies that vulnerable hosts are not uniform over the IP-address space. Meanwhile, NAT technology is widely used in the Internet. Meanwhile, the complexity of internal and external network environments has asymmetry [12], [13], which indicates the complexity of local network or NAT subnet is far lower than the complexity of the external network.

The social network is an information dissemination and worm spreading platform. To identify the influential nodes and restrain the propagation of those risks, researchers have introduced a number of vital node methods (e.g., node degree, closeness and betweenness [9], [14], [15]). However, these measures are unknown during the information dissemination. Therefore, for defense immunization in hierarchical networks, we intend to design a novel propagation strategy of vaccines in hierarchical networks to solve the problems:

- How to quickly scan more vulnerable hosts in order to immunize them in the physical network layer?
- How to accurately measure social nodes in order to discover vital nodes in the social network layer?

III. MODELING HYBRID WORM PROPAGATION

For studying the hybrid worm, in contrast with previous model, we consider the stochastic difference equation to characterize hybrid spreading behaviors through the idea of divide and conquer to characterize from the perspective of individual state transition.

A. Modeling Hybrid Propagation Process

Let the variable X_i^t denote the state of host i at time t . Note: we study hybrid spreading behaviors in Section II-A. And there is the following assumption: users and hosts follow a one-to-one correspondence relationship in the hierarchical network.

$$X_i^t = \begin{cases} \text{Sus.}, & \text{susceptible} \\ \text{Imm.}, & \text{immune} \\ \text{Act.}, & \text{active} \\ \text{Dor.}, & \text{dormant} \end{cases} \quad (1)$$

The variable X_i^t becomes one of four states (i.e., *susceptible*, *immune*, *active*, *dormant*) at time t , where susceptible state and immune state belong to *healthy* (*Hea.*) state. Active state and dormant state belong to *infected* (*Inf.*) state.

The graph of node state transition is provided in Fig. 3. Every node is presumed to be the *susceptible* state at the beginning. If the worm propagates by sending out malicious emails or scanning IP addresses to compromise any vulnerable target host i , node i becomes the *infected* state with infection probability $v(i, t)$. For the topology-based behavior, *infected* state includes two sub-states: *active* and *dormant*. If a user does not operate an infected host, it will become the *dormant*

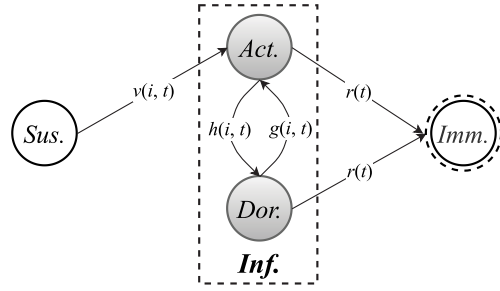


Fig. 3. Node state transition in hierarchical networks.

state with the probability $h(i, t)$. When a user is opening an email with a malicious payload or checking the infected mailbox, the dormant node becomes the *active* state with the probability $g(i, t)$, and then can send out malicious copies. For the scan-based behavior, the *infected* state only includes one sub-state: *active*, and it means that the active hosts can always scan vulnerable hosts. An arbitrary node can become the *immunized* state with the recovery probability $r(t)$.

The susceptible and infected states can be denoted by 0 and 1, respectively. Assume that there is a hierarchical network with N nodes. Let $n(t)$ denote the expected number of infected nodes at time t as follows:

$$\begin{aligned} n(t) &= E \left[\sum_{i=1}^N X_i^t \right] = \sum_{i=1}^N E [X_i^t] = \sum_{i=1}^N P(X_i^t = 1) \\ &= \sum_{i=1}^N P(X_i^t = \text{Inf.}) \end{aligned} \quad (2)$$

We can calculate the value of $P(X_i^t = \text{Inf.})$ from Fig. 3,

$$\begin{cases} P(X_i^t = \text{Inf.}) \\ = 1 - P(X_i^t = \text{Imm.}, X_i^{t-1} \in \{\text{Hea.}, \text{Inf.}\}) - P(X_i^t = \text{Sus.}, X_i^{t-1} = \text{Sus.}) \\ = 1 - P(X_i^t = \text{Imm.}, X_i^{t-1} \in \{\text{Sus.}, \text{Inf.}\}) - P(X_i^t = \text{Imm.}, X_i^{t-1} = \text{Imm.}) \\ - P(X_i^t = \text{Sus.}, X_i^{t-1} = \text{Sus.}) \\ = 1 - \underbrace{R_i(t)}_{\text{Immune probability}} - P(X_i^{t-1} = \text{Imm.}) - \underbrace{E_i(t)}_{\text{Susceptible probability}} \\ P(X_i^t = \text{Imm.}) = P(X_i^t = \text{Imm.}, X_i^{t-1} \in \{\text{Hea.}, \text{Inf.}\}) \\ = R_i(t) + P(X_i^{t-1} = \text{Imm.}) \\ P(X_i^t = \text{Sus.}) = P(X_i^t = \text{Sus.}, X_i^{t-1} \in \{\text{Hea.}, \text{Inf.}\}) \\ = P(X_i^t = \text{Sus.}, X_i^{t-1} = \text{Sus.}) = E_i(t) \\ = \underbrace{P(X_i^t = \text{Sus.} | X_i^{t-1} = \text{Sus.})}_{\text{Maintaining probability}} P(X_i^{t-1} = \text{Sus.}) = \underbrace{e(i, t)}_{\text{Maintaining probability}} \cdot P(X_i^{t-1} = \text{Sus.}) \end{cases} \quad (3)$$

Once we have the values of $R_i(t)$ and $e(i, t)$, we can obtain the value of $P(X_i^t = \text{Inf.})$ by equation (3).

Suppose that the recovery probability $r(t)$ denotes the possibility of hosts being immunized at time t . The infection probability $v(i, t)$ denotes the possibility of host i being

compromised and infected by hybrid worms at time t .

$$\begin{cases}
 \textbf{Recovery probability:} \\
 r(t) = \frac{R_i(t)}{P(X_i^{t-1} \in \{Sus., Inf.\})} \\
 = P(X_i^t = Imm. | X_i^{t-1} \in \{Sus., Inf.\}) \\
 \textbf{Maintaining probability:} \\
 e(i, t) = P(X_i^t = Sus. | X_i^{t-1} = Sus.) \\
 = 1 - r(t) - v(i, t) \\
 \textbf{Infection probability:} \\
 v(i, t) = P(X_i^t = Inf. | X_i^{t-1} = Sus., X_{U_i}^{t-1} = x_{U_i}^{t-1})
 \end{cases} \quad (4)$$

where the recovery probability $r(t)$ depends on the effectiveness of the AV (Antivirus) software (i.e., how long AV manufacturers can repair all network vulnerabilities), or the availability of AV software (i.e., how long it takes AV manufacturers to release their updates against a worm). Collecting release information of AV update for each malware sample is very difficult and sometimes impossible. Many AV providers do not share publicly this information. Thus, this paper uses several typical recovery probabilities in Section V-D.4.

Once obtaining the value of $v(i, t)$, we calculate the value of $P(X_i^t = Inf.)$ by the iteration of the equations (3) and (4).

The hybrid worm spreading in hierarchical networks has two scan-based and topology-based behaviors. And then let $v_s(i, t)$ and $v_e(i, t)$ denote the infection probabilities of the two behaviors, respectively. *The independence assumption*: there is concurrency between the scan-based behavior and the topology-based behavior for one host, and different hosts are relatively independent individuals. Thus, we consider the independence between $v_s(i, t)$ and $v_e(i, t)$. There is the hierarchical property of hybrid worms spreading throughout hierarchical networks. Thus, we calculate the infection probability $v(i, t)$ of hybrid behaviors,

$$v(i, t) = 1 - [1 - v_s(i, t)][1 - v_e(i, t)] \quad (5)$$

Next, we model scanning and topological propagation processes to obtain the values of $v_s(i, t)$ and $v_e(i, t)$, respectively.

B. Scan-Based Propagation Process $v_s(i, t)$

To improve the scanning effectiveness, the enhanced combination-scanning method of the hybrid worm introduces the probability of the worm scanning the specific group m (i.e., the group scanning distribution $P_g^*(m)$) to help worms scan some network segments containing more vulnerable hosts. In addition, to characterize scan-based propagation process, we should consider the influence of the topological behavior. Next, we model the two above factors in details.

1) *Calculating Group Scanning Distribution $P_g^*(m)$* : Given that a worm scans Ω_m hosts in group m with the same likelihood, and then there are two steps in the worm hitting host i in group m : First, the worm locates the group m of host i with the probability $P_g^*(m)$; Next, the worm scans host i in group m with the probability $\frac{1}{\Omega_m}$. Thus, the probability of a susceptible host i being hit is $P_h(i) = P_g^*(m) \cdot \frac{1}{\Omega_m}$.

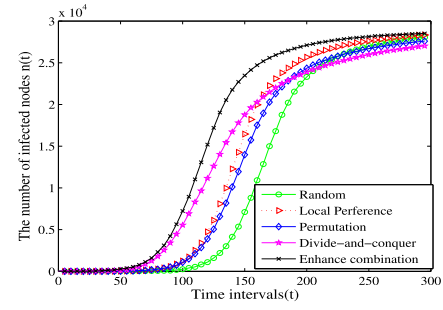


Fig. 4. The comparison among different scanning strategies. All cases are for $N = 30,000$, $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$. Experimental environment setup is used in Section V-A.

Let X_i denote the number of scans when host i is first hit, which follows a geometric distribution. Thus, we have the expectation of X_i is $E[X_i] = P_h(i)^{-1} = \frac{\Omega_m}{P_g^*(m)}$.

Suppose the average number of scans until an arbitrary susceptible host i is first compromised, denoted by $\mathbb{E}_i$. In order to improve scanning efficiency, we should minimize $\mathbb{E}_i$ as follows:

$$\begin{cases}
 \text{minimize } \mathbb{E}_i = \sum_{m=1}^g P_g(m) \cdot \frac{\Omega_m}{P_g^*(m)} \\
 \text{subject to } \sum_{m=1}^g P_g^*(m) = 1 \quad \text{and} \quad \sum_{m=1}^g P_g(m) = 1
 \end{cases} \quad (6)$$

Thus, *group scanning distribution $P_g^*(m)$* is obtained from Lagrange Multipliers

$$P_g^*(m) = \frac{\sqrt{\Omega_m \cdot P_g(m)}}{\sum_{n=1}^g \sqrt{\Omega_n \cdot P_g(n)}} \quad (7)$$

Because N and N_m are not known beforehand in the real world, we use an unbiased estimator $P_g(m) = \frac{C_m}{C}$ to calculate $P_g^*(m)$, where C_m is the number of worm clients in group m , and C is the total number of worm clients.

Hence, the **enhanced combination-scanning** method adopts two steps to hit vulnerable hosts in the target network:

- **Step 1**: Probability p_{in} to hit some vulnerable hosts by scanning an IP address in the “/ρ” prefix intra-area network;
- **Step 2**: Probability $p_{out} \cdot P_g^*(m)$ to hit some vulnerable hosts by scanning an IP address in the group m of the inter-area network.

As shown in Fig. 4, we can see that enhanced combination-scanning improves worm’s spreading ability in the case of vulnerable hosts non-uniformly distributed over target networks.

2) *The Influence of the Topology-Based Behavior*: The scan-based/topology-based behaviors generally belong to two different threads, and there is concurrency between them. Thus, when calculating infection probability $v_s(i, t)$ for the scan-based behavior, we should exclude the infected hosts being used to check the mailbox.

Suppose that the variable Y_j^t denotes whether the user is checking mails through host j in host set $U \subset U_i^{t-1}$ at time t . Thus, we have I_U^t is the expected number of infection hosts in host set U by which corresponding users are checking mails

at time t .

$$\begin{aligned} \bar{I}_U^t &= E[\sum_{j \in U} \bar{Y}_j^t] = \sum_{j \in U} E[\bar{Y}_j^t] = \sum_{j \in U} P(\bar{Y}_j^t = 1) \\ &= \sum_{j \in U} P(X_j^t = \text{Inf.}) \cdot P(\text{open}_j^t = 1) \end{aligned} \quad (8)$$

Thus, we should exclude them when calculating the number of infected hosts for the probability $v_s(i, t)$ later.

3) *Calculating Infection Probability $v_s(i, t)$ for the Scan-Based Behavior:* The IP address of a host is either public or private. Consequently, there are two cases that a host is compromised in the propagation process based on enhanced combination-scanning strategy (i.e., Step 1 and Step 2). *The independence assumption:* there is concurrency between the intra-area scanning behavior and the inter-area scanning behavior for one host, and different hosts are relatively independent individuals. Thus, we consider the probabilities of host i not being hit in scanning Steps 1 and 2 to be independent of each other.

- *Case 1 (public IP):* In scanning step 1, infected hosts in the “/ρ” prefix intra-area network can implement $\alpha_1 = p_{in} \cdot s \cdot (I_m^{t-1} - I_{U_m}^{t-1})$ local preference scans, and then the victim i in the “/ρ” prefix is compromised with the probability $\frac{1}{\Omega_{/\rho}}$. Thus, the probability of host i not being hit is $(1 - \frac{1}{\Omega_{/\rho}})^{\alpha_1}$. Again, in scanning step 2, infected hosts in the entire network can implement $\alpha_2 = p_{out} \cdot P_g^*(m) \cdot s \cdot (I_{t-1} - I_U^{t-1})$ remote scans, and then a susceptible host i is hit with the probability $\frac{1}{\Omega_m}$. Thus, the probability P_{case1} of host i not being hit is $(1 - \frac{1}{\Omega_m})^{\alpha_2}$. And we obtain the probability of host i being hit,

$$P_{case1} = (1 - NAT_{in}) \cdot [1 - (1 - \frac{1}{\Omega_{/\rho}})^{\alpha_1} \cdot (1 - \frac{1}{\Omega_m})^{\alpha_2}] \quad (9)$$

- *Case 2 (private IP):* Similar to case 1, in scanning step 1, infected hosts in the “/ρ” prefix NAT subnet perform $\beta_1 = p_{in} \cdot s \cdot (I_{n|m}^{t-1} - I_{U_{n|m}}^{t-1})$ local preference scans. Thus, the probability of host i not being hit is $(1 - \frac{1}{\Omega_{/\rho}})^{\beta_1}$. Again, in scanning step 2, infected hosts in the same NAT subnet perform $\beta_2 = p_{out} \cdot P_g^*(m) \cdot s \cdot (I_{n|m}^{t-1} - I_{U_{n|m}}^{t-1})$ remote scans. Thus, the probability of host i not being hit is $(1 - \frac{1}{\Omega_m})^{\beta_2}$. And we obtain the probability of host i being hit,

$$P_{case2} = NAT_{in} \cdot [1 - (1 - \frac{1}{\Omega_{/\rho}})^{\beta_1} \cdot (1 - \frac{1}{\Omega_m})^{\beta_2}] \quad (10)$$

where NAT_{in} means whether the host is in NAT subnet.

Thus, based on the above analysis, we obtain **infection probability $v_s(i, t)$ for the scan-based behavior**,

$$v_s(i, t) = P_{case1} + P_{case2} \quad (11)$$

C. Topology-Based Propagation Process $v_e(i, t)$

In order to solve NAT blocking problem for the scan-based propagation process, the topological behavior introduces enhanced reinfection mechanism.

Given a social networking layer with M users, the propagation relationship between these users can be described as the following M by M square matrix

$$\begin{pmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & p_{ij} & \vdots \\ p_{M1} & \cdots & p_{MM} \end{pmatrix} p_{ij} \in [0, 1] \quad (12)$$

where p_{ij} is the infection probability from user i to user j , which is the probability of user j reading a malicious email from user i .

From the perspective of the short-term user behavior, users may check their mailboxes or not, and there is a lot of randomness. In order to eliminate the impact of randomness on modeling, we consider that users check their mailboxes periodically from the perspective of the long-term user behavior and statistics. Let random variable open_i^t discriminate whether the user i is opening the mailbox at time t , where $\text{open}_i^t = 1$ indicates that user i is checking the mailbox at time t . Thus, we have

$$P(\text{open}_i^t = 1) = \begin{cases} 1, & t \bmod T_{check}^i = 0 \\ q_n, & t = t' \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where T_{check}^i represents the checking period, and $t = t'$ means that the mail application will notify user i when the mailbox receives a new mail at time t' . q_n denotes the probability that user i checks emails after receiving a notification.

Meanwhile, for the periodic behavior of checking emails, let variable ϵ represent an arbitrary time of every newly arrived email between the user checking the last mailbox and the current time t .

$$\begin{cases} t - T_{check}^i \leq \epsilon < t, & \text{if } \text{open}_i^t = 1 \\ t - (t \bmod T_{check}^i) \leq \epsilon < t, & \text{otherwise} \end{cases} \quad (14)$$

Let random variable trigger_i^t denote whether the specific action of checking the mailbox is triggered. We have $\text{trigger}_i^t = 1$ if the action of user i happens at time t . Otherwise, $\text{trigger}_i^t = 0$. Thus, we obtain the probability of the user i 's action being triggered,

$$P(\text{trigger}_i^t = 1) = \begin{cases} 1, & \text{open}_i^t = 1 \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

In order to characterize topology-based propagation process, we introduce the function $s(i, t)$ to characterize the behavior of user i reading malicious emails from neighbors. Meanwhile, we should consider the influence of the scanning behavior. Next, we model the two above aspects in details.

1) *Calculating the Probability of Reading Malicious Emails $s(i, t)$:* The probability $s(i, t)$ that user i is reading malicious emails sent by infected neighbor j is calculated as,

$$s(i, t) = 1 - \prod_{j \in \tilde{U}_i} [1 - p_{ji} \cdot P(X_j^\epsilon = \text{Act.})] \quad (16)$$

where $P(X_j^\epsilon = \text{Act.})$ denotes that infected node j sends out a malicious copy at time ϵ , where the relation ϵ and t is the equation (14). $\tilde{U}_i$ denotes the neighbor set of user i in the

contact list. The equation (16) means that user i is reading newly arrived malicious emails at time ϵ , after the user i 's checking time t .

For *enhanced reinfection* case, we introduce virtual nodes to characterize the k th infection event caused by triggering specific action or reading k th malicious email at time t . And then it requires $\widetilde{U}_i$ to be extended into three subsets: $\widetilde{U}_{i|N}$, $\widetilde{U}_{i|R}$ and $\widetilde{U}_{i|E}$. Thus, we need to revise the equation (16),

$$s(i, t) = 1 - \prod_{j \in \widetilde{U}_{i|N} \cup \widetilde{U}_{i|R} \cup \widetilde{U}_{i|E}} [1 - p_{ji} \cdot P(X_j^\epsilon = Act.)] \quad (17)$$

First, the subset $\widetilde{U}_{i|N}$ denotes the real neighbors, in which these neighbors get infected by user i reading the first malicious email. According to Fig. 3, the function $g(i, t) = P(open_i^t = 1) \cdot s(i, t)$ means that the user i 's host in the dormant state can become the active state when user i is checking the mailbox and reading malicious emails at time t . Because $h(i, t)$ is the opposite of $g(i, t)$, the function $h(i, t) = 1 - g(i, t)$. Thus, we have

$$\begin{aligned} P(X_j^\epsilon = Act.) &= v_e(j, \epsilon) \cdot P(X_j^{\epsilon-1} = Sus.) \\ &+ g(j, \epsilon) \cdot P(X_j^{\epsilon-1} = Dor.) \\ &+ [1 - h(i, \epsilon) - r(t)] \cdot P(X_j^{\epsilon-1} = Act.) \end{aligned} \quad (18)$$

Second, the subset $\widetilde{U}_{i|R}$ denotes the virtual neighbors, in which these neighbors get infected by user i reading more than one malicious email. Let $\eta_{ji}(\epsilon)$ represent the probability that user i is infected by neighbor j at time ϵ .

$$\begin{cases} \eta_{ji}(\epsilon) = P(X_j^\epsilon = Act.) \cdot P(open_i^\epsilon = 1) \cdot p_{ji} \cdot [1 - r(\epsilon)] \\ \overline{\eta_{ji}(\epsilon)} = 1 - \eta_{ji}(\epsilon) \end{cases} \quad (19)$$

To obtain the probability of reading the k th ($k > 1$) email sent by infected neighbor j , we consider node j (i.e., the k th virtual node, denoted by j_k) being infected by $C_{||\widetilde{U}_j||}^k$ infected neighbors. Thus, we obtain

$$P(X_{j_k}^\epsilon = Act.) = \sum_{z \in U_j^k} [\prod_{x \in z} \eta_{xi}(\epsilon) \cdot \prod_{y \in \widetilde{U}_j - z} \overline{\eta_{yi}(\epsilon)}] \quad (20)$$

where the set U_j^k consists of $C_{||\widetilde{U}_j||}^k$ elements, in which each element is the set of k infected nodes selected from set $\widetilde{U}_j$ randomly.

Third, the subset $\widetilde{U}_{i|E}$ denotes the infection event caused by the infected node performing some specific actions. According to the equation (15), we can calculate $P(X_j^\epsilon = Act.)$ for user j triggering a specific action,

$$P(X_j^\epsilon = Act.) = P(trigger_j^\epsilon = 1) \cdot P(X_j^\epsilon = Inf.) \quad (21)$$

2) *The Influence of the Scan-Based Behavior*: Suppose that there are not malicious emails in user j 's mailbox, but user j 's host infected by the scan-based behavior can search the contact list and send malicious copies to neighbors at checking time ϵ . Thus, when calculating the probability

$P(X_j^\epsilon = Act.)$, we should consider those users' hosts infected by the scan-based propagation process without malicious emails in their mailbox at checking time ϵ . And then we can obtain $P(X_j^\epsilon = Act.)$ that user j can send out malicious copies without malicious emails in the mailbox,

$$P(X_j^\epsilon = Act.) = P(X_j^{\epsilon-1} = Sus.) \cdot v_s(j, \epsilon) \cdot P(open_i^\epsilon = 1) \quad (22)$$

3) *Calculating Infection Probability $v_e(i, t)$ for the Topology-Based Behavior*: There are three preconditions for each infection in the topology-based propagation process: (1) user i is checking the mailbox at time t ; (2) user i reads those malicious emails at time t ; (3) user i 's host has not been immunized at time t . Thus, we obtain **infection probability $v_e(i, t)$ for the topology-based behavior**,

$$v_e(i, t) = P(open_i^t = 1) \cdot s(i, t) \cdot [1 - r(t)] \quad (23)$$

IV. DEFENSE STRATEGY ORIENTED TOWARDS HIERARCHICAL NETWORKS

Inspired by the propagation process of the hybrid worm, we design a new **vaccine** based on the preimmunization HMI strategy (i.e., this vaccine with our proposed defense strategy pre-patches vulnerable hosts and immunizes them) in the hierarchical networks to defend against the malicious worm propagation. Similar with worms, our vaccine have four main components as follows:

- *Intrusion engine*: The code segment of getting access permission of the target system (e.g., buffer overflow, file share, and misconfiguration).
- *Propagation engine*: The code segment of transmitting malicious codes to the target host (e.g., exploit FTP, HTTP, and Email).
- *Scanning engine*: The code segment of searching target hosts by exploiting the resources of victims.
- *Payload*: The code segment of performing a specific function.

Based on the above analysis, *our proposed defense strategy is deployed in propagation and scanning engines*, and then we package multiple application vulnerabilities or Oday vulnerabilities into the intrusion engine. The patch program is placed in the payload. In this way, we can automatically and quickly immunize hosts in heterogeneous networks. However, in order to implement this strategy better, we should consider these questions:

- **Question 1**: *How to calculate the vulnerable-host proportion in the specific group accurately by scanning target networks?* As shown in Fig.5 (a), We can see that group distribution $P_g(m) = \frac{N_m}{N}$ can capture the statistics of vulnerable hosts in group m . And then we can increase the scanning speed by group scanning distribution $P_g^*(m)$ in the equation (7).
- **Question 2**: *How to discover the more vital nodes in social networks?* As shown in Fig.5 (b), from the propagation perspective, in infection paths, the number of cascades that red node 1 infected by other nodes (e.g., gray nodes) and infecting other nodes (e.g., gray nodes)

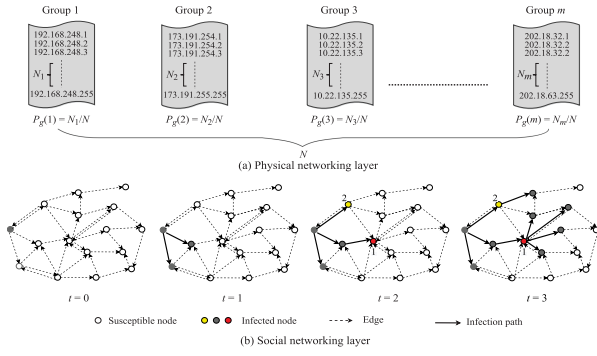


Fig. 5. The vaccine defense scenarios. Subfigure (a) illustrates an example of susceptible hosts distribution in physical networking layer. Subfigure (b) illustrates an example of worm spreading in social network layer.

is more than the yellow node 2's cascading number. And then we can see that the vital node (e.g., the red node 1) can influence worm spreading in social networks more efficiently than the normal node (e.g., the yellow node 2). Thus, we can propose a propagation centrality measure to quantify the propagation influence of an arbitrary node. It explains that vital nodes have more large impacts on the information cascade in reality.

To answer the two questions, we present the following methods in Sections IV-A and IV-B.

A. Estimation of Group Distribution $P_g(m)$ in the Physical Network Layer

In this section, we derive the computation of the group distribution for developing a faster scanning strategy. According to the definition of the group distribution $P_g(m)$ in TABLE I, once we estimate any vulnerable-host population $\widehat{N}_m$, we can obtain estimator $\widehat{P}_g(m)$ of group distribution.

1) *Calculating Vulnerable-Host Population $\widehat{N}_m$ in Group m :* Suppose that $N_c(m)$ is the number of worm clients¹ of the group m . $N_s(m)$ is the sample size that is collected from a target group m . Therefore, the lower limit of the sample size collected by each worm client in group m is $N_{s/c}(m) = \lceil (N_s(m)/(N_c(m))) \rceil$. Let $x_{j \rightarrow k}^m (j = 1, \dots, N_c(m); k = 1, \dots, N_{s/c}(m))$ denote whether the k th host of group m is vulnerable, which is scanned by the j th worm client,

$$x_{j \rightarrow k}^m = \begin{cases} 0, & \text{host } k \text{ is not vulnerable} \\ 1, & \text{otherwise} \end{cases} \quad (24)$$

That is, it follows a Bernoulli distribution. Given that $P_g(m, b)$ and $\widehat{P}_g(m, b)$ denote actual and estimated vulnerable host proportions in the sub-group b for the group m , respectively. Thus, the probability mass function is

$$f_X[x_{j \rightarrow k}^b | P_g(m, b)] = P_g(m, b)^{x_{j \rightarrow k}^b} \cdot [1 - P_g(m, b)]^{1-x_{j \rightarrow k}^b} \quad (25)$$

Through calculating the likelihood function $L[P_g(m, b)] = \prod_{k=1}^{N_{s/c}(m)} f_X[x_{j \rightarrow k}^b | P_g(m, b)]$, we obtain the estimator of $P_g(m, b)$ as

¹the corresponding IP area that the client worm scans in the group m is called a sub-group for group m .

$$\widehat{P}_g(m, b) = \frac{1}{N_{s/c}(m)} \cdot \sum_{k=1}^{N_{s/c}(m)} x_{j \rightarrow k}^m \quad (26)$$

Thus, we have

$$\begin{aligned} P_g(\widehat{m}, \Omega_m) &= \frac{\sum_{b=1}^{N_c(m)} \widehat{P}_g(m, b)}{N_c(m)} = \frac{\sum_{j=1}^{N_c(m)} \sum_{k=1}^{N_{s/c}(m)} x_{j \rightarrow k}^m}{N_c(m) \cdot N_{s/c}(m)} \\ &= \frac{\sum_{j=1}^{N_c(m)} \sum_{k=1}^{N_{s/c}(m)} x_{j \rightarrow k}^m}{N_s(m)} \end{aligned} \quad (27)$$

where $P_g(\widehat{m}, \Omega_m)$ (i.e., $\frac{\widehat{N}_m}{\Omega_m}$) is the estimated vulnerable-host proportion of in Ω_m address space. Finally, we obtain the estimator of vulnerable-host population $\widehat{N}_m$ as

$$\widehat{N}_m = P_g(\widehat{m}, \Omega_m) \cdot \Omega_m \quad (28)$$

According to the equations (27) and (28), once we calculate the value of $N_s(m)$, the above value of $\widehat{N}_m$ is obtained.

2) *Calculating Sample Size Collected in Group m $N_s(m)$:* We see that the random variable $X = \sum_{j=1}^{N_c(m)} \sum_{k=1}^{N_{s/c}(m)} x_{j \rightarrow k}^b$ follows the binomial distribution. According to central limit theorem, we have

$$\begin{aligned} \lim_{N_s(m) \rightarrow \infty} P \left\{ \frac{X - N_s(m) \cdot P_g(m, \Omega_m)}{\sqrt{N_s(m) \cdot P_g(m, \Omega_m) \cdot [1 - P_g(m, \Omega_m)]}} \leq x \right\} \\ = \lim_{N_s(m) \rightarrow \infty} P \left\{ \frac{\frac{X}{N_s(m)} - P_g(m, \Omega_m)}{\sqrt{P_g(m, \Omega_m) \cdot [1 - P_g(m, \Omega_m)]/N_s(m)}} \leq x \right\} \end{aligned} \quad (29)$$

And then, we choose the equation (29) to be $100(1-\alpha)$ percent confident that error $e = |P_g(\widehat{m}, \Omega_m) - P_g(m, \Omega_m)|$ is less than sampling error $Z_{\frac{\alpha}{2}} \cdot FA_{N_s(m) \rightarrow \infty}$, where $FA_{N_s(m) \rightarrow \infty} = \sqrt{P_g(m, \Omega_m) \cdot [1 - P_g(m, \Omega_m)]/N_s(m)}$. And we obtain

$$\lim_{N_s(m) \rightarrow \infty} P \left\{ e \leq Z_{\frac{\alpha}{2}} \cdot FA_{N_s(m) \rightarrow \infty} \right\} = 1 - \alpha \quad (30)$$

According to the equation (30), we have

$$N_s(m) \leq \left(\frac{Z_{\frac{\alpha}{2}}}{e} \right)^2 \cdot P_g(m, \Omega_m) \cdot [1 - P_g(m, \Omega_m)] \quad (31)$$

Because the bigger the sample size is, the smaller the error is. Meanwhile, when the value of $P_g(m)$ is 0.5, we can calculate an upper bound on the equation (31). That is to say, we can obtain the estimating error of smaller than e with at least $100(1-\alpha)$ percent confident.

$$\begin{aligned} \overline{N_s(m)} &= \left(\frac{Z_{\frac{\alpha}{2}}}{e} \right)^2 \cdot P_g(m, \Omega_m) \cdot [1 - P_g(m, \Omega_m)] \\ &= \left(\frac{Z_{\frac{\alpha}{2}}}{e} \right)^2 \cdot 0.25 \end{aligned} \quad (32)$$

In the real world, the sample size $N_s(m)$ is finite. We introduce a correction factor $CF = \sqrt{\frac{\Omega_m - N_s(m)}{\Omega_m - 1}}$ to the condition of $N_s(m) \rightarrow \infty$. And the equation (29) is revised as:

$$\lim_{N_s(m) \rightarrow \Omega_m} P \left\{ \frac{e}{FA_{N_s(m) \rightarrow \infty} \cdot CF} \leq Z_{\frac{\alpha}{2}} \right\} = 1 - \alpha \quad (33)$$

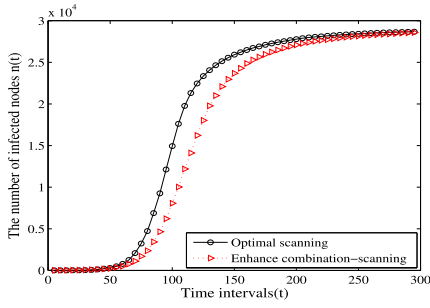


Fig. 6. The comparison between two scanning strategies. All cases are for $N = 30,000$, $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$. Experimental environment setup is used in Section V-A.

Thus, we see $e = Z_{\frac{\alpha}{2}} \cdot F A_{N_s(m) \rightarrow \infty} \cdot CF$ satisfies $100(1 - \alpha)$ percent confident, and combine the equation (32), the sample size of group m is obtained as

$$N_s(m) = \frac{\overline{N_s(m)} \cdot \Omega_m}{\overline{N_s(m)} + (\Omega_m - 1)} \quad (34)$$

Thus, based on $N_s(m)$ obtained from the equation (34) and $P_g(m, \Omega_m)$ from the equation (27), we can calculate the **estimator of group distribution** as

$$\widehat{P_g(m)} = \frac{\widehat{N_m}}{\widehat{N}} = \frac{P_g(m, \Omega_m) \cdot \Omega_m}{\sum_{n=1}^g P_g(n, \Omega_n) \cdot \Omega_n} \quad (35)$$

We divide the target scanning space into eight groups in the experiment. Table II shows that the precision of estimating $\widehat{P_g(m)}$ will greatly improve with the increase in confidence coefficient $1 - \alpha$ and the decrease of the error e . In the case of $e = 1\%$, $\alpha = 1\%$, the absolute estimation errors $\mathcal{E} = e(\widehat{P_g(m)})$ are close to 0.1%. Thus, the evaluation proves the effectiveness of our method.

Base on the above analysis, we present an **optimal scanning strategy** based on HSP model:

- In the learning stage, we give a new and unbiased estimator for group m distribution over time: $\widehat{P_g(m)} = \frac{\widehat{N_m} - I_m^t}{\sum_{n=1}^g (\widehat{N_n} - I_n^t)}$, and we have $P_g^*(m) = \frac{\sqrt{\Omega_m \cdot \widehat{P_g(m)}}}{\sum_{n=1}^g \sqrt{\Omega_n \cdot \widehat{P_g(n)}}$
- In the scanning stage, we continue the idea in Fig. 2 (b).
- Base on the enhanced combination-scanning strategy, we add the updating stage: the worm server calculates $\Delta P_g^t(m, n) = P_g^*(m) - P_g^*(n)$. If the numerical sign of $\Delta P_g^t(m, n)$ changes, it sends the current $P_g^*(m)$ to all corresponding infected hosts.

As shown in Fig. 6, it indicates that optimal scanning worm spreads faster than the enhanced combination-scanning worm. The main reason is that our proposed method can estimate $\widehat{P_g(m)}$ over time more accurately.

B. Measure of Vital Nodes in the Social Network Layer

We compute propagation centrality of nodes based on the HSP model in social networks. In fact, the centrality of the node denotes the ability of itself infecting other nodes. We give

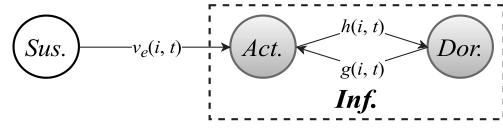


Fig. 7. State transition graph of a node in the social network layer.

the node state transition graph in Fig. 7. Thus, we have

$$\begin{cases} P(X_i^t = Inf.) = v(i, t) \cdot P(X_i^{t-1} = Sus.) + P(X_i^{t-1} = Inf.) \\ P(X_i^t = Sus.) = 1 - P(X_i^t = Inf.) \\ v_e(i, t) = s(i, t) \cdot P(open_i^t = 1) \\ s(i, t) = 1 - \prod_{j \in \tilde{U}_i} [1 - p_{ji} \cdot P(X_j^e = Act.)] \\ P(X_i^t = Act.) = v_e(i, t) \cdot P(X_i^{t-1} = Sus.) \\ \quad + g(i, t) \cdot P(X_i^{t-1} = Dor.) \\ \quad + [1 - h(i, t)] \cdot P(X_i^{t-1} = Act.) \end{cases} \quad (36)$$

1) *Calculating Influence of a Node $E(\mathcal{J}_{i|r})$* : Suppose that an epidemic incident occurs on node r in the social network. we consider the influence of node i on this propagation process, $\mathcal{J}_{i|r}$, as the sum of the expected number of the subsequent infected nodes affected by node i (i.e., the outflowing influence $\mathcal{J}_{i \rightarrow |r}$) and the expected number of the previous infected nodes affecting node i (i.e., the inflowing influence $\mathcal{J}_{\rightarrow i|r}$). The former and latter explain the cascading possibility of node i infected by other nodes and the cascading ability of node i infecting other nodes, respectively. Thus, we have

$$E(\mathcal{J}_{i|r}) = E(\mathcal{J}_{i \rightarrow |r}) + E(\mathcal{J}_{\rightarrow i|r}) \quad (37)$$

Due to node i getting infected at any time, we obtain

$$E(\mathcal{J}_{i \rightarrow |r}) = \sum_{t=1}^{\infty} [P(X_i^t = Act.) \cdot E(\mathcal{J}_{i \rightarrow |r}^t)] \quad (38)$$

$$E(\mathcal{J}_{\rightarrow i|r}) = \sum_{t=1}^{\infty} [P(X_i^t = Act.) \cdot E(\mathcal{J}_{\rightarrow i|r}^t)] \quad (39)$$

In social networking layer, the epidemic spreading is neighborhood spreading. Thus, the outflowing and inflowing influences of node i are derived from its neighbors. Thus,

$$E(\mathcal{J}_{i \rightarrow |r}^t) = \sum_{j \in \tilde{U}_i} \zeta_{ij}^t [E(\mathcal{J}_{j \rightarrow |r}^{t+1}) \cdot P(X_j^{t+1} = Act.)] \quad (40)$$

$$E(\mathcal{J}_{\rightarrow i|r}^t) = \sum_{j \in \tilde{U}_i} \zeta_{ji}^t [E(\mathcal{J}_{\rightarrow j|r}^{t-1}) \cdot P(X_j^{t-1} = Act.)] \quad (41)$$

where ζ_{xy}^t denotes the ratio of the contribution of node x infecting node y at time t . Thus, we have

$$\zeta_{xy}^t = \frac{p_{xy} \cdot P(X_x^t = Act.)}{\sum_{k \in N_{B_y}} [p_{ky} \cdot P(X_k^t = Act.)]} \quad (42)$$

2) *Calculating Propagation Centrality $V_{pc}(i)$* : The above analysis fixes the position of the starting node in the spreading process. Actually, epidemics have randomness in the network. Thus, the propagation centrality of node i should be averaged

TABLE II

THE VULNERABLE-HOST GROUP DISTRIBUTION, WHERE $\mathcal{E} = e(\widehat{P_g(m)}) = |P_g(\hat{m}) - P_g(m)|$

m	$e = 1\%, \alpha = 1\%$				$e = 1\%, \alpha = 5\%$				$e = 5\%, \alpha = 1\%$				$e = 5\%, \alpha = 5\%$			
	$N_s(m)$	$P_g(m)$	$P_g(m)$	$\mathcal{E}$	$N_s(m)$	$P_g(m)$	$P_g(m)$	$\mathcal{E}$	$N_s(m)$	$P_g(m)$	$P_g(m)$	$\mathcal{E}$	$N_s(m)$	$P_g(m)$	$P_g(m)$	$\mathcal{E}$
0	8256	0.122	0.121	0.001	6055	0.121	0.121	0.000	639	0.127	0.121	0.006	375	0.115	0.121	0.006
1	8256	0.126	0.124	0.002	6055	0.122	0.124	0.002	639	0.137	0.124	0.013	375	0.150	0.124	0.026
2	8256	0.129	0.126	0.003	6055	0.134	0.126	0.008	639	0.130	0.126	0.004	375	0.139	0.126	0.013
3	8256	0.122	0.125	0.003	6055	0.125	0.125	0.000	639	0.124	0.125	0.001	375	0.099	0.125	0.026
4	8256	0.085	0.087	0.002	6055	0.085	0.087	0.002	639	0.076	0.087	0.011	375	0.099	0.087	0.012
5	8256	0.112	0.111	0.001	6055	0.113	0.111	0.002	639	0.080	0.111	0.031	375	0.084	0.111	0.027
6	8256	0.151	0.153	0.002	6055	0.154	0.153	0.001	639	0.157	0.153	0.004	375	0.161	0.153	0.008
7	8256	0.154	0.154	0.000	6055	0.148	0.154	0.006	639	0.169	0.154	0.015	375	0.153	0.154	0.001

* Note: we use synthetic physical topology in Section V-A

over all possible starting position as follows:

$$V_{pc}(i) = \frac{1}{M} \cdot \sum_{r=1}^M E(\mathcal{J}_i|_r) \quad (43)$$

We carry the experiment to simulate email worm spreading (e.g., Mydoom, Nyxem) in the Facebook, and then obtain the node influence containing the previous infection scale infecting this node and the following infection scale infected by this node. The above results are considered as the benchmark to validate the effectiveness of the propagation centrality. We sort influential nodes exposed by different measures (i.e., betweenness, closeness, communicability [16], current_flow_betweenness, current_flow_closeness [17], degree, eigenvector [18], load [19], epidemic betweenness [20] and propagation centrality), respectively. We fetch the sorted nodes for each method and the simulation results at a sampling ratio $f\%$. The intersection of their respective sorted results can evaluate the effectiveness of various measures. As shown in Fig. 8, we can see that the red solid line of our propagation centrality achieves higher intersection ratios than other measures. And it indicates that our presented measure can identify vital nodes more accurately.

Based on the above analysis, we present a **preimmunization HMI strategy** as follows:

- For vulnerable hosts in physical network layer, we release the vaccine by the optimal scanning strategy to achieve fast patch dissemination.
- For nodes in social network layer, we exploit the vaccine to sort the influential nodes by the propagation centrality. According to a realistic budget, we immunized the top $f\%$ vital nodes.

V. MODEL VALIDATION

In this field, all existing research adopts simulation to evaluate analytical models [4], [5], [7], [21], [22], [23]. Thus, we use C++ and Matlab R2014a to develop a simulation environment for the hybrid worm.

A. Experimental Setup and Assumptions

Firstly, the hierarchical network includes physical networking and social networking layers. The social network with 45,814 nodes is based on the realistic Facebook topology [24]. The physical network with IPv4 addresses is synthetic which

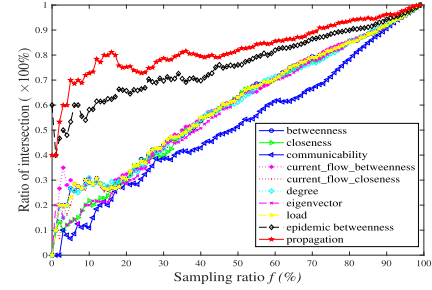


Fig. 8. The intersection percentage of vital nodes between different measures and simulations. Experimental environment setup and Facebook topology are used in Section V-A.

consists of 41 NAT subnets and 14 autonomous systems (AS). We assume that the vulnerable-host scale $N = 63150$ and the scan rate $s = 5$ [4].

Secondly, in order to align experiment parameters consistent with a series of experiments, according to parameter settings of previous works [4], [9], [21], [25], we assume that the infection probability p_{ij} and the checking time T_{check}^i follow Gaussian distributions $N(0.5, 0.2^2)$ and $N(40, 20^2)$, respectively. Note: Gaussian distribution generator may produce some values without practical meaning, such as $p_{ij} < 0$ and $T_{check}^i < 1$. Thus, we replace these values with the realistic minimum values. If $p_{ij} < 0$ and $T_{check}^i < 1$, we let $p_{ij} = 0$ and $T_{check}^i = 1$.

Thirdly, we simulate the hybrid worm propagation as a discrete process, and then construct the hierarchical network where each social node corresponds to one physical node. On the other hand, we deploy social events for user simulation (i.e., the social networking layer) and scanning strategies for host simulation (i.e., the physical networking layer), respectively. According to the transition diagram, the state of each host switches among the four states: susceptible, active, dormant, and immune. An infected host can send a sequence of probes to infect other hosts through scanning strategies, certain trigger action or user reading malicious emails. Finally, we consider the average number of infected nodes in simulation as the benchmark in order to evaluate our model.

B. Realistic Infection Samples

Generally, the spread of worms in the real world cannot be traced directly. Even worse, hybrid worms involve users'

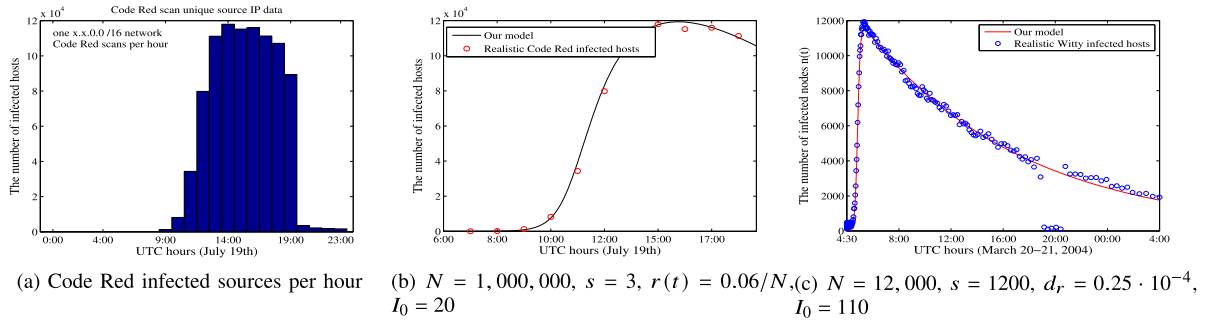


Fig. 9. Comparisons between our revised model and two types of worms.

privacy. Therefore, available worm datasets in the real world are not provided for the experimental evaluation. However, in order to validate the accuracy and rationality of the HSP model, we adopt “the idea of controlling variables” to verify partial characteristics of our model.

1) *Real-World Data of the Scan-Based Worms*: The data 1 in Fig. 9(a) by Eichman [26] includes the scale of Code Red infected sources sending out worm scans per hour on July 19th. In order to characterize the Code Red worm, we revised our HSP model,

$$v(i, t) = v_s(i, t) = [1 - (1 - \frac{1}{\Omega})^{s \cdot I_{t-1}}] \cdot [1 - r(t)] \quad (44)$$

As shown in Fig. 9(b), we compare the solid line from our model with the realistic curve of the infection scale from July 19th 6:00 to 17:00 UTC, and can see that they are matched quite well.

The data 2 of Witty worm indicates the scanning number during each 1000s from observation records of IMS on three “/24” blackhole networks [27]. Witty uniformly scanned the Internet. Meanwhile, researchers do not consider the recovery process, but they introduce the destruction rate d_r to denote the ratio of infected hosts destroyed by Witty worm. Thus, we replace $r(t)$ with d_r in the equation (44). Fig. 9(c) shows our model could characterize the propagation of Witty worm very well by comparing the experimental result with the Witty trace.

2) *Real-World Data of the Topology-Based Worm*: The data 3 of Cckun worm is observed by CNCERT/CC [28], which was detected by the Mobile Service Providers (MSPs) of China on Sep. 22, 2013. In order to compare with this dataset, we revise the equations (13) and (23) to characterize the diurnal behavior at different time zones,

$$\bar{P}(\text{open}_i^t = 1) = \begin{cases} 1, & t \bmod T_{check}^i = 0 \text{ and } U^{-1}[0, 1] \in \text{act}(t) \\ q_n, & t = t' \\ 0, & \text{otherwise} \end{cases} \quad (45)$$

$$v(i, t) = v_e(i, t) = s(i, t) \cdot \bar{P}(\text{open}_i^t = 1) \quad (46)$$

where $U^{-1}[0, 1]$ is the inverse function of uniform distribution $U[0, 1]$, $\text{act}(t)$ represents the diurnal behavior of message sending and T_{check}^i follows the zeta distribution with parameter $\tau = 2.28$. The setting of related parameters is the same as paper [28]. Fig. 10 shows that our model can characterize the infection trend of Cuckun’s worm accurately.

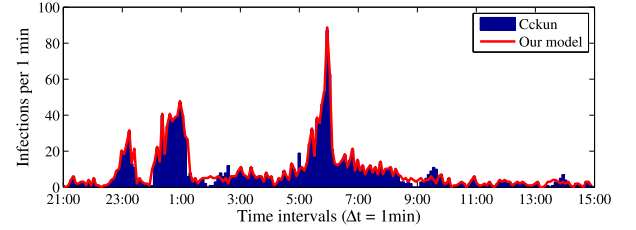


Fig. 10. Infection incremental of Cckun’s worm. $N = 400,000$, $I_0 = 10$.

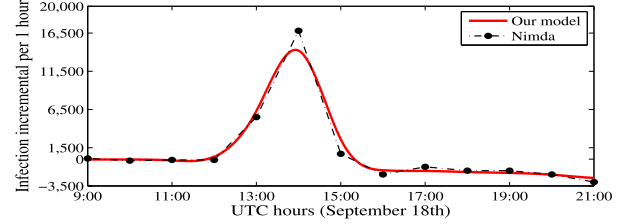


Fig. 11. Hourly infection incremental of Nimda worm on Sept. 18th. The scan rate $s = 10$, $p_{in} = 0.75$, $p_{out} = 0.25$.

3) *Real-World Data of the Hybrid Worm*: On September 18th, SANS Internet Storm Center received massive reports of network slowdowns and HTTP probing. And the data 4 of Nimda worm observed by CAIDA [29] indicates that the attack frequency increased dramatically and reached a peak at approximately 14:00 GMT. Fig. 11 depicts the hourly infections for the number of particular activity sources on the day Nimda is first reported. In view of Nimda mainly infects other devices through Email and the local-preference scanning strategy, we revise the equations $v_e(i, t)$ and $v_s(i, t)$ as follows:

$$\begin{cases} v_e(i, t) = s(i, t) \cdot P(\text{open}_i^t = 1) \\ v_s(i, t) = [1 - (1 - \frac{1}{\Omega})^{p_{out} \cdot s \cdot I_{t-1}}] \cdot (1 - \frac{1}{\Omega_m})^{p_{in} \cdot s \cdot I_m^{t-1}} \end{cases} \quad (47)$$

Fig. 11 indicates that the prediction of our model fits the infection curve of Nimda worm very well.

C. Analyzing the Performance of Our HSP Model

Next, we validate the accuracy of our HSP model and the impact of some factors.

1) *Comparison Our Model With Previous Models*: We mainly study worm propagation in excluding the impact of the recovery process ($r(t) = 0$). And then let checking time T_{check}^i and infection probability p_{ij} follow Gaussian distributions $N(40, 20^2)$ and $N(0.5, 0.2^2)$ [4], [9], [21], and [25], respectively. Additionally, we use two different recovery functions $r(t)$ to describe the recovery process as follows: (1) ‘Constant

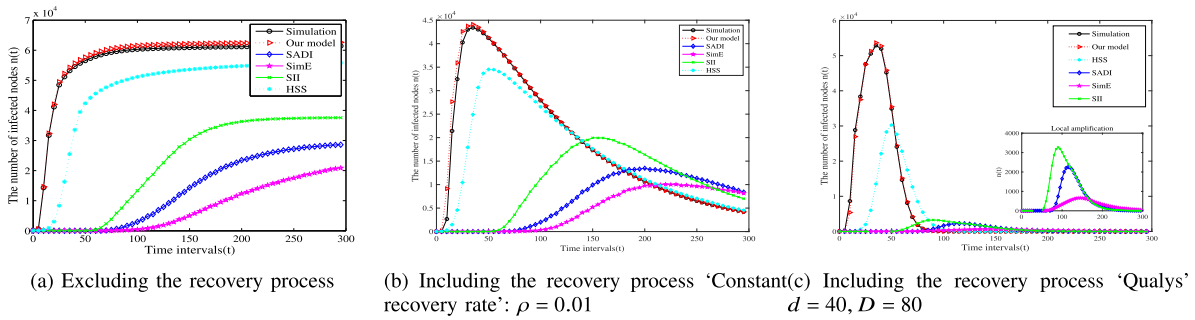


Fig. 12. Comparison the proposed model with other models. Checking time: $T_{check}^i \sim N(40, 20^2)$; Infection probability $p_{ij} \sim N(0.5, 0.2^2)$; The scan rate $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$.

recovery rate' [8]: $r(t) = \rho$; (2) 'Qualys' [30]: is obtained by the statistics of Qualys Inc.

$$r(t) = \begin{cases} 1 - 0.5 \frac{t-d}{D}, & t \geq d \\ 0, & t < d \end{cases} \quad (48)$$

where d is the time span from the malware appearing on the Internet to it is detected by scientists. D is the half-life period, which represents the time that the infection scale decreased by 50% [10].

Note: In the experiment, the parameter settings of the same factors (e.g., $T_{check}^i \sim N(40, 2^2)$, $p_{ij} \sim N(0.5, 0.2^2)$, $s = 5$, $p_{in} = 0.7$, and $p_{out} = 0.3$) involved in characterizing spreading behaviors are the same in various models. And it can explain we set various model parameters in a fair manner.

Excluding the recovery process, Fig. 12 (a) shows that previous models [9], [21], [25], [31] underestimate the infection scale. The primary reason is that they cannot model the multiple behaviors of hybrid worms. However, the HSP model can characterize hybrid worm spreading in the hierarchical network accurately. On the other hand, including the recovery process, Fig. 12 (b) and (c) show that our model has a greater accuracy than previous models [9], [21], [25], [31] in characterizing hybrid worm propagation in the two different recovery processes.

2) *Comparison With Different Propagation Mechanisms in the Social Network:* We explain how different mechanisms affect the topology-based propagation of hybrid worms. The events of mailbox opening and message notification are associated with user checking emails and message notification, respectively. Fig. 13 indicates that message notification can accelerate the worm spreading, because users will check the message when they notice a message box pop-up. Meanwhile, the reason why reinfection can stimulate the worm spreading is that it can send out more malicious copies. However, the infection scale of notification is smaller than the infection scale of reinfection and mailbox opening trigger, because it does not increase the number of messages, but just speeds up the frequency of reading messages. Additionally, the non-reinfection spreads the slowest. The reason is it will send only one malicious copy.

D. Evaluating the Defense Effects of Our HMI Strategy

We verify the defense effect of our HMI strategy in the section.

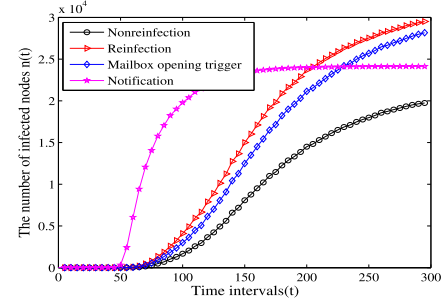


Fig. 13. Impacts of different mechanisms in social network layer. The scan rate $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$; Checking time: $T_{check}^i \sim N(40, 20^2)$; $q_n \sim N(0.5, 0.2^2)$.

1) *Defense Effect in Social Networking Layer:* We exploit the propagation centrality to find out the most vital nodes in social networks. We sort the nodes in ascending order through the equation (43), and then immunize the top $f = 10\%$, 20% , 30% nodes, as shown in Fig. 14(a). We can see that the propagation of the hybrid worm can be obviously restrained. Moreover, the inset of Fig. 14(a) shows that propagation centrality measure will almost stop worm spreading in the social networking layer when we set defense ratio $f = 30\%$.

2) *Defense Effect in Physical Networking Layer:* We use the vaccine based on optimal scanning strategy to immunize hosts in physical networks. Fig. 14(b) shows that optimal scanning strategy restrains the spreading of hybrid worms effectively through the difference Δ of infection peaks between the defense and no defense.

3) *Defense Effect for the HMI Strategy:* We use the vaccine based on the HMI strategy, as shown in Fig. 14(c). We can see that this strategy can restrain the infected number of the hybrid worm very well. Meanwhile, according to different time T_{start} of releasing the vaccine, Fig. 14(c) shows that the shorter deployment delay determines the smaller infection scale. This experimental result conducts security companies to improve their detection ability and release patches as early as possible.

4) *Comparison With Other Strategies:* To further investigate the defense performance of the HMI strategy, we consider the following three baselines: 'Ratio': $r(t) = \frac{n(t)}{\Omega}$, where its design philosophy is the recovery rate $r(t)$ equals the number of infected hosts $n(t)$ divided by the network size Ω ; 'Composite' [32]: $r(t) = r_0 \left[1 + \frac{i(t)}{\Omega}\right]^\gamma$, where $i(t)$ is the number of immune hosts, and r_0 is the initial recovery rate. The scale parameter γ aims to regulate the model sensitivity; 'Qualys': is obtained by the statistics of Qualys Inc. [30]

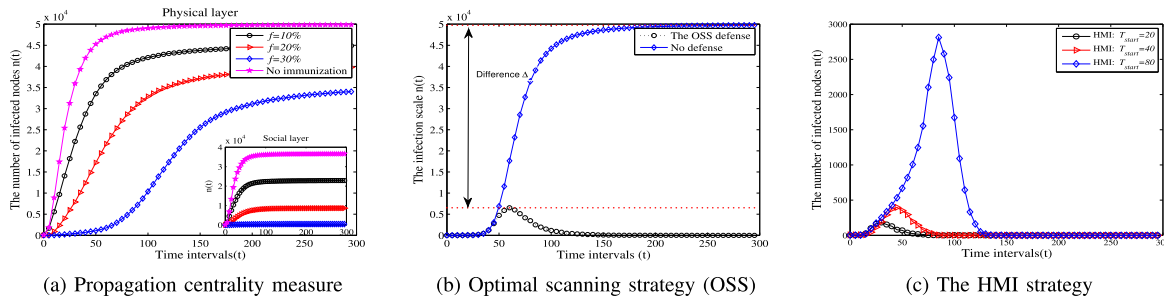


Fig. 14. Comparisons between our revised model and two types of worms. The scan rate $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$; Checking time: $T_{check}^i \sim N(40, 20^2)$; $p_{ij} \sim N(0.5, 0.2^2)$, $q_n \sim N(0.5, 0.2^2)$.

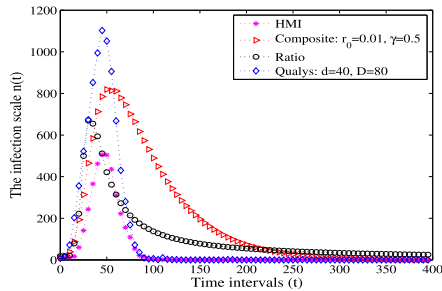


Fig. 15. Impacts of different recovery functions. The scan rate $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$; $T_{check}^i \sim N(40, 20^2)$; $p_{ij} \sim N(0.5, 0.2^2)$, $q_n \sim N(0.5, 0.2^2)$.

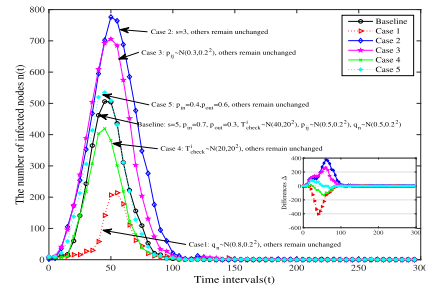


Fig. 16. Impacts of different parameters. Baseline: The scan rate $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$; $T_{check}^i \sim N(40, 20^2)$; $p_{ij} \sim N(0.5, 0.2^2)$, $q_n \sim N(0.5, 0.2^2)$.

and $r(t)$ is as shown in the equation (48). The parameter setting of our HMI strategy: The scan rate $s = 5$, $p_{in} = 0.7$, $p_{out} = 0.3$, $T_{check}^i \sim N(40, 20^2)$, $p_{ij} \sim N(0.5, 0.2^2)$ and $q_n \sim N(0.5, 0.2^2)$. Note: according to some previous works [8], [10], [21], [32], the recovery process is usually represented as the recovery function because of specific repaired data obtained by us hardly. And their measure of defense effectiveness means which strategy restrains the infection scale of worms more obviously in the same temporal span has more better defense effect.

Compared with other strategies, Fig. 15 shows that ‘Qualys’ is the worst defense effect at the beginning, because hybrid worm can spread freely before the time span d from the malware appearing on the Internet to it is detected by scientists. ‘Ratio’ is better defense effect that ‘Composite’ because of greater recovery growth rate. However, we can see that our defense strategy can significantly suppress worm spreading, and it can bring worm propagation to a halt faster. Obviously, it proves the effectiveness and advancement of our HMI strategy.

5) *Sensitivity Analysis of Our HMI Model*: HMI model with different q_n , s , p_{ij} , T_{check}^i , p_{in} and p_{out} values are investigated. We have Case1 (what if message notification becomes timelier): $q_n \sim N(0.8, 0.2^2)$; Case2 (what if infected hosts scan less IP address every time): $s = 3$; Case3 (what if the network is less vulnerable): $p_{ij} \sim N(0.3, 0.2^2)$; Case4 (what if users check messages more frequently): $T_{check}^i \sim N(20, 20^2)$; Case5 (what if infected hosts scan vulnerable hosts in the intra-area less possibly): $p_{in} = 0.4$ and $p_{out} = 0.6$. As shown in Fig. 16, compared with the baseline, Cases1, 2, 3 and 5 are positive correlation with the defense effect and it means that their weaker impacts lead to the less effective defense. Case4 is negative with the defense effect and it means that their weaker

impacts lead to the better defense. Meanwhile, we can see that the HMI model is most sensitive to parameter q_n , which directly affect the defense effect.

VI. RELATED WORK

A. Worm Propagation Through Scan and Topology

For decades, researchers focus on studying the propagation process of *scan-based* and *topology-based* worms. The former exploits various scanning strategies to detect vulnerable hosts [4], [33]. Recently, several scan-based worms with new features emerged. Chen and Ji [23] introduced an important-scanning strategy based on the vulnerable-host distribution to improve scanning efficiency. The camouflaging worm [5] hide its tracks by manipulating scan traffic volume over time intelligently. Yu et al. [6] introduced game-theoretical formulations to study the propagation and countermeasure of self-disciplinary worms. Manna et.al [7] comprehensively studied the propagation of permutation-scanning worms, and investigated the impact of dynamic topology on the model correctness.

The topology-based worms transfer themselves among different terminals via social links. With the deep-going of social engineering and social networks, many researchers turned their attention to modeling propagation behaviors of social worms. Chen and Ji [8] proposed a novel model based on probabilistic graph, which uses the independent and Markov models to approximate the spatially dependent propagation process. Zou et al. [9] make comparative studies on the spread of email worms in different types of network. Wen et al. [10] proposed the SII model, which aims to study the impacts of temporal dynamics and spatial dependence on social worms. Meanwhile, Wen et al. [21] introduced the reinfection mechanism to address the previous underestimation problem of the

infection scale. Wang et al. [11], [31] considered the impacts of social behavior and user mobility on social worms.

However, existing scan-based models are mainly developed by difference and differential equations, which cannot cover all propagation situations. Differential-based models primarily focus on the changes of infection scale from a continuous-time perspective, but cannot characterize discrete propagation events. Difference-based models model the propagation process from the discrete-time perspective, while they cannot describe random user behaviors. Despite the above problems can be solved by topological models using stochastic discrete-time methods, they only concentrate on social worms and do not involve scanning worms. Finally, some scholars characterized several hybrid worms [25], [34] preliminarily. However, in their research, propagation dynamics and user behaviors can not be analyzed from an analytical approach. Additionally, some researchers studied IoT worm propagation, such as wireless IoT worms [35], [36], [37], and VANETs (vehicular ad hoc networks) worms [38], [39]. However, because these models adopt “homogenous mixing” assumption, they cannot analyze hybrid worm propagation.

B. Defense Strategies Against Worm Propagation

There are some countermeasures (e.g., analytical equation analysis) to find some conditions for preventing worm propagation. The paper [40] focused on finding threshold conditions for worm extinction. However, the method above cannot comprehensively understand social worm propagation to defend against them. Papers [9], [21], [31] analyzed the impacts of user social behaviors on worm propagation to reduce the spreading speed from the perspective of the social engineering.

Some others such as detection technologies and exposing the influential nodes have been presented to defend against worms. The former prevents users' privacy disclosure. Recently, Xu et al. [41] developed a novel detection system of social worms using worm propagation characteristics and OSN topological properties. This paper [42] presented a client-side detection scheme for the characteristics of reconstructing and propagating XSS worm payload, and prevented worm propagation by monitoring outgoing self-replicating payloads. However, if new worms produce some unknown patterns, it is difficult for the existing detection technologies to detect them. The latter finds some most influential nodes to block the worm propagation. Researchers constrained topological worm spreading in social networks by immunizing the highly-connected nodes [9], [43] or the nodes with most active neighbors [22]. However, this view was argued by other works [14], [15] as authors found the best defense positions may not come from the large-degree nodes. However, none of the above countermeasures can be applied in physical networks to patch vulnerable hosts automatically.

In addition, benign worms have been presented in [44] and [45] to explore an active immunization method. This method can quickly immunize nodes and effectively prevent worm propagation. However, they cannot solve the immunization deployment in hierarchical networks effectively and fast.

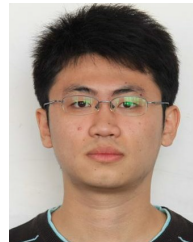
VII. CONCLUSION

Motivated by the problem statement and analysis, we study the spread of hybrid worms from the perspectives of propagation and defense. The HSP model is proposed for characterizing two critical processes of malware spreading in the hierarchical network: the important scanning and enhanced reinfection. The mathematical derivation illuminates the model rationality, and then the experimental results validate that our model has better superiority on characterizing hybrid propagation behavior. On the other hand, we develop a novel HMI defense strategy. Compared with other immunization strategies, our strategy has remarkable advantages in restraining the scale and speed of worm propagation.

REFERENCES

- [1] K. Haley, N. Johnson, and J. Fulton, “Symantec internet security threat report 2015,” Symantec Corp., Mountain View, CA, USA, Tech. Rep. 21, Apr. 2016.
- [2] M. Abu Rajab, F. Monrose, and A. Terzis, “On the impact of dynamic addressing on malware propagation,” in *Proc. 4th ACM Workshop Recurring Malcode*, Nov. 2006, pp. 51–56.
- [3] G. Serazzi and S. Zanero, “Computer virus propagation models,” in *Performance Tools and Applications to Networked Systems*. Cham, Switzerland: Springer, 2004, pp. 26–50.
- [4] C. C. Zou, D. Towsley, and W. Gong, “On the performance of internet worm scanning strategies,” *Perform. Eval.*, vol. 63, no. 7, pp. 700–723, 2003.
- [5] W. Yu, X. Wang, P. Cui, D. Xuan, and W. Zhao, “Modeling and detection of camouflaging worm,” *IEEE Trans. Depend. Sec. Comput.*, vol. 8, no. 3, pp. 377–390, Mar. 2011.
- [6] W. Yu, N. Zhang, X. Fu, and W. Zhao, “Self-disciplinary worms and countermeasures: Modeling and analysis,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 10, pp. 1501–1514, Oct. 2010.
- [7] P. K. Manna, S. Chen, and S. Ranka, “Inside the permutation-scanning worms: Propagation modeling and analysis,” *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 858–870, Mar. 2010.
- [8] Z. Chen and C. Ji, “Spatial-temporal modeling of malware propagation in networks,” *IEEE Trans. Neural Netw.*, vol. 16, no. 5, pp. 1291–1303, Sep. 2005.
- [9] C. C. Zou, D. Towsley, and W. Gong, “Modeling and simulation study of the propagation and defense of internet e-mail worms,” *IEEE Trans. Dependable Secure Comput.*, vol. 4, no. 2, pp. 105–118, Apr./Jun. 2007.
- [10] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, “Modeling propagation dynamics of social network worms,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 8, pp. 1633–1643, Aug. 2013.
- [11] T. Wang, C. Xia, and Q. Jia, “The temporal characteristic of human mobility: Modeling and analysis of social worm propagation,” *IEEE Commun. Lett.*, vol. 19, no. 7, pp. 1169–1172, Jul. 2015.
- [12] B. J. Tillett, “Symantec internet security threat report 2011,” Symantec Corp., Mountain View, CA, USA, Tech. Rep. 17, Apr. 2012.
- [13] M. Fossi and J. Blackbird, “Symantec internet security threat report 2013,” Symantec Corp., Mountain View, CA, USA, Tech. Rep. 19, Apr. 2014.
- [14] S. Wen, W. Zhou, Y. Wang, W. Zhou, and Y. Xiang, “Locating defense positions for thwarting the propagation of topological worms,” *IEEE Commun. Lett.*, vol. 16, no. 4, pp. 560–563, Apr. 2012.
- [15] M. Kitsak et al., “Identification of influential spreaders in complex networks,” *Nature Phys.*, vol. 6, pp. 888–893, Aug. 2010.
- [16] E. Estrada, D. J. Higham, and N. Hatano, “Communicability betweenness in complex networks,” *Phys. A, Stat. Mech. Appl.*, vol. 388, no. 5, pp. 764–774, Mar. 2009.
- [17] U. Brandes and D. Fleischer, “Centrality measures based on current flow,” in *Proc. Annu. Symp. Theor. Aspects Comput. Sci.* Berlin, Germany: Springer, 2005, pp. 533–544.
- [18] P. Bonacich, “Some unique properties of eigenvector centrality,” *Social Netw.*, vol. 29, no. 4, pp. 555–564, Oct. 2007.
- [19] U. Brandes, “On variants of shortest-path betweenness centrality and their generic computation,” *Social Netw.*, vol. 30, no. 2, pp. 136–145, May 2008.

- [20] S. Wen, J. Jiang, B. Liu, Y. Xiang, and W. Zhou, "Using epidemic betweenness to measure the influence of users in complex networks," *J. Netw. Comput. Appl.*, vol. 78, pp. 288–299, Jan. 2016.
- [21] S. Wen et al., "Modeling and analysis on the propagation dynamics of modern email malware," *IEEE Trans. Dependable Secure Comput.*, vol. 11, no. 4, pp. 361–374, Jul. 2014.
- [22] G. Yan, G. Chen, S. Eidenbenz, and N. Li, "Malware propagation in online social networks: Nature, dynamics, and defense implications," in *Proc. 6th ACM Symp. Inf., Comput. Commun. Secur.*, Mar. 2011, pp. 196–206.
- [23] Z. Chen and C. Ji, "Importance-scanning worm using vulnerable-host distribution," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, vol. 3, Nov. 2005, p. 6.
- [24] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proc. 2nd ACM Workshop Online Social Netw.*, Aug. 2009, pp. 37–42.
- [25] C. Zhang, S. Zhou, and B. M. Chain, "Hybrid epidemics—A case study on computer worm conficker," *PLoS ONE*, vol. 10, no. 5, pp. 1–17, May 2015, doi: [10.1371/journal.pone.0127478](https://doi.org/10.1371/journal.pone.0127478).
- [26] K. Eichman. (Jul. 20, 2001). *Mailist: Re: Possible coded connection attempts*. [Online]. Available: <http://lists.jamned.com/incidents/2001/07/0159.html>
- [27] *University of Michigan Internet Motion Sensor*. Accessed: 2005. [Online]. Available: <http://ims.eecs.umich.edu/>
- [28] X. Yun, S. Li, and Y. Zhang, "SMS worm propagation over contact social networks: Modeling and validation," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2365–2380, Nov. 2015.
- [29] *Caida: Dynamic: Analysis: Security: Nimda*. Accessed: Mar. 1, 2006. [Online]. Available: <https://www.caida.org/dynamic/analysis/security/nimda/>
- [30] W. Kandek, "The laws of vulnerabilities," in *Proc. BlackHat Conf. Japan: Qualys*, 2009, pp. 1–10.
- [31] T. Wang, C. Xia, S. Wen, H. Xue, Y. Xiang, and S. Tu, "SADI: A novel model to study the propagation of social worms in hierarchical networks," *IEEE Trans. Dependable Secure Comput.*, vol. 16, no. 1, pp. 142–155, Jan. 2019.
- [32] D. Zhang and Y. Wang, "SIRS: Internet worm propagation model and application," in *Proc. Int. Conf. Electr. Control Eng.*, Jun. 2010, pp. 3029–3032.
- [33] Y. Wang, S. Wen, Y. Xiang, and W. Zhou, "Modeling the propagation of worms in networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 942–960, 2nd Quart., 2014.
- [34] Q. Liu, J. Jing, and Y. Wang, "An improved method of hybrid worm simulation," in *Proc. 9th Int. Conf. Web-Age Inf. Manage.*, Jul. 2008, pp. 612–618.
- [35] M. López, A. Peinado, and A. Ortiz, "An extensive validation of a SIR epidemic model to study the propagation of jamming attacks against IoT wireless networks," *Comput. Netw.*, vol. 165, Dec. 2019, Art. no. 106945.
- [36] M. J. Farooq and Q. Zhu, "Modeling, analysis, and mitigation of dynamic botnet formation in wireless IoT networks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2412–2426, Sep. 2019.
- [37] D. Acarali, M. Rajarajan, N. Komninos, and B. B. Zarpelão, "Modelling the spread of botnet malware in IoT-based wireless sensor networks," *Secur. Commun. Netw.*, vol. 2019, pp. 1–13, Feb. 2019.
- [38] H. Yang, D. Chen, G. Sun, X. Ding, and Y. Xin, "CC2: Defending hybrid worm on mobile networks with two-dimensional circulation control," *Complexity*, vol. 2019, no. 1, pp. 1–19, Dec. 2019.
- [39] A. Boukerche and Q. Zhang, "Countermeasures against worm spreading: A new challenge for vehicular networks," *ACM Comput. Surveys*, vol. 52, no. 2, pp. 1–25, Mar. 2019.
- [40] D. Chakrabarti, J. Leskovec, C. Faloutsos, S. Madden, C. Guestrin, and M. Faloutsos, "Information survival threshold in sensor and P2P networks," in *Proc. 26th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, May 2007, pp. 1316–1324.
- [41] W. Xu, F. Zhang, and S. Zhu, "Toward worm detection in online social networks," in *Proc. 26th Annu. Comput. Secur. Appl. Conf.*, Dec. 2010, pp. 11–20.
- [42] F. Sun, L. Xu, and Z. Su, "Client-side detection of XSS worms by monitoring payload propagation," in *Proc. Eur. Symp. Res. Comput. Secur.* Berlin Springer, 2009, pp. 539–554.
- [43] X. Fan and Y. Xiang, "Modeling the propagation of peer-to-peer worms," *Future Gener. Comput. Syst.*, vol. 26, no. 8, pp. 1433–1443, Oct. 2010.
- [44] F. Castaneda, E. C. Sezer, and J. Xu, "WORM vs. WORM: Preliminary study of an active counter-attack mechanism," in *Proc. ACM Workshop Rapid Malcode*, Oct. 2004, pp. 83–93.
- [45] X. Zheng, T. Li, and Y. Fang, "Strategy of fast and light-load cloud-based proactive benign worm countermeasure technology to contain worm propagation," *J. Supercomput.*, vol. 62, no. 3, pp. 1451–1479, Dec. 2012.



Tianbo Wang (Member, IEEE) received the Ph.D. degree in computer application from Beihang University, Beijing, China, in 2018. He is currently an Associate Professor with Beihang University. He has participated in several National Natural Science Foundations and other research projects. His research interests include network and information security, intrusion detection technology, and information countermeasure.



Huacheng Li received the bachelor's degree in software engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China, in 2017. He is currently pursuing the Ph.D. degree with Beihang University, Beijing, China, under the supervision of Prof. Chunhe Xia. His research interests include social network analysis and information security.



Chunhe Xia received the Ph.D. degree in computer application from Beihang University, Beijing, China, in 2003. He is currently a Supervisor and a Professor at Beihang University and the Director of the Beijing Key Laboratory of Network Technology. He has participated in different national major research projects and published more than 70 research papers in important international conferences and journals. His current research interests include network and information security, information countermeasure, cloud security, and network measurement.



Han Zhang (Member, IEEE) received the Ph.D. degree from Tsinghua University, Beijing, China. He is currently working with the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include computer network systems and network security. He has published more than 50 papers in his area.



Pei Zhang received the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2012. He is currently working with the School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications. His research interests include computer networks, network security, and AI.