MP-CA: A Malware Propagation Modeling Methodology Based on Cellular Automata

ZAHRA BAKHSHI1, MINA ZOLFY LIGHVAN2 and REZA MOSTAFAVI3

1, 2, 3 Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

E-mail: 1z.bakhshi91@ms.tabrizu.ac.ir, 2mzolfy@tabrizu.ac.ir, 3r.mostafavi91@ms.tabrizu.ac.ir

ABSTRACT

The variety of security threats caused by malwares has turned their dispersion into a potential danger. Malware propagation modeling is a facility that allows the researchers to predict the side effects of a new threat and understand the behavior of the modeled malware. On the other hand, due to the high cost and diversity of existing networks and the capability of those networks to be infected by such malwares, behavioral modeling of malware becomes a challengeable issue in recent works. In the last few years, the growing popularity of smart phones has made them an attractive target to hackers and malware writers. One of possible communication channels for the penetration of mobile malware is the Bluetooth interface. In this paper, a new analytical modeling methodology for malware propagation using three-dimensional cellular automata and based on the epidemic theory has been presented and as a case study the propagation of Bluetooth worm has been discussed.

Keywords: Malware, Propagation, Modeling, Cellular Automata, Bluetooth.

1 INTRODUCTION

A Malware is a broad term for different kinds of malicious programs including worms, spyware, viruses, and adware [1]. A program is known as malware if it installs itself without awareness and user satisfaction. The goal and infection type of malwares identifies their type [2]. Spyware is a program that gathers user’s information without his authorization and sends them to other places. Adware is another type of malware which displays uninvited advertise and other undesirable marketing ads. A virus replicates itself and constantly places new copies in different files and programs. After a few decades from the spreading of the first computer virus, malware propagation takes significant contributions in various fields of security challenges [3]. With the development of information technology in all aspects of life, the threat of malwares have turned into a major concern. While email is a basic service for computer users, email malware is a crucial security danger. Moreover, according to capabilities and applications smartphone, it can be exposed to various attack vectors such as SMS, MMS, Bluetooth, Wi-Fi, etc. On the other hand, in wireless sensor networks each sensor node can be attacked by different types of malwares such as worms, virus and Trojan. Due to the potential damages caused by malware, researchers have proposed numerous models to describe the propagation process of malicious software in which modeling objectives can be summarized as follows [4]:

1-Understanding the behavior of malicious software including: attributes and spreading prerequisites and its influencing factors.
2-Anticipate propagation of malware before they happen.
3-Assess the system accessibility for spread of malware and evaluate the impacts of malware spreading on the Network.
4-Identify the potential ability of malware in subversive activities.
5-Detecting the malware propagation speed and the time needed to contaminate the whole network.
6-Adopting the suitable preventive measures and appropriate defensive actions based on behavior of the given malware.

7-Describing the required efficiency of countermeasures in order to control the propagation.

8-Facilitating design a reliable network that be resilient against all types of malware attacks.

9-Forseeing the failures of the universal network infrastructure.

To this purpose, based on the available mathematical modeling and epidemic theories, mathematical epidemiology has been introduced. Epidemic modeling is utilized to mimic the dissemination of infectious illness for a given crowd, such as influenza, H1N1, and SARS. Contaminated persons propagate the infection to healthy individuals that they contact with. Since computer worms are similar to such biological viruses in their self-replicating and diffusion behaviors, epidemiological models for examining the propagation of malware, especially worms is not a new criteria [12]. Studying computer worms overall, and Internet worms specifically, is a popular subject for analysts. Numerous endeavors have been made to model the spread behaviors of malwares in different networks [5],[6],[7],[8],[9].The epidemic models can be categorized into two primary groups. The first is the deterministic model, which is represented by the ordinary differential equation [9].The second is the stochastic model which contains two types: one is based on Markov chain [8],[10] and the other is based on cellular automata. Most models have focused on the technology of differential equations and the Markov chain [8].Models based on differential equations fail to catch the local features of propagation processes. They also neglect to interaction behaviors among individuals. On the other hand, the models based on the Markov chain are complex to explain the spatial temporal process of worm propagation. Cellular automata [13] is the answer for this problems. Because Cellular automata (CA)can dominate these issues, it has been used as an effective alternative method to describe epidemic spreading and malware propagation[12],[14],[13],[15],[16],[17].In fact, cellular automata can model the physical computation capabilities, biological, or environmental complex phenomena, such as growth processes, reaction–diffusion systems, epidemic models, and the spread of forest fire.

In this paper, an analytical model based on cellular automata for malware propagation has been presented which as a case study the propagation dynamics of Bluetooth worms has been described. The rest of this paper is organized as follows: Section 2 gives an outline of related work. In Section 3 short overview of Bluetooth technology and cellular automata as background knowledge has been provided. We have discussed about the MP-CA in Section 4. In Section 5, the proposed modeling approach for characterizing the epidemic spreading is described explicitly. Model validation and results are presented in Section 6 and the paper is concluded in Section 7.

2 RELATED WORK

3 BACKGROUND

In this section, a brief introduction on the required background is reviewed. First, the cellular automata is described as a basic modeling methodology then the Bluetooth premier is illustrated which has been used as the case study for presented MP-CA methodology.

3.1 Cellular Automata

In the early 1950s von Neumann and Stan Ulam presented the Cellular automata[27][13] as a simple model of self-replicating biological systems. A Cellular automaton is a dynamical system whose behavior is completely based on local communications of individual cells. In CA, the space is characterized as a network of cells. Finite set of states defines that at any moment every cell can be in one of these states. Cellular automata are discrete in time and their rules have been described globally. Impact of neighboring cells on a cell, characterized by cellular automata rules. It means that at any state, each cell acquires its new state with regard to the state of adjacent neighbors .The fundamental features of cellular automata is the following: discrete space, discrete time, limitation the number of possible states for every cell, all cells are identical, certainty of the rules, dependence of rules to limited number values of previous steps each cell and neighbors of this cell. Different types of CAs have been presented over the years. The most of them have common characteristics and overall. In general, they are defined as a one-dimensional cellular networks, two-dimensional, three-dimensional or multi-dimensional. According to the above description, the mathematical definition of cellular automata is a tuple as follows:

\[ \text{CA} = (N, Q, V, F) \]

Where:
- \( N \): Includes an array of cells and identifies dimensions of cellular networks.
- \( Q \): Represents a finite number of discrete states that a cell can take.
- \( V \): Represents the number of neighbors that a cell has.
- \( F \): Represents the transition function that a cell follows.

3.2 Bluetooth Primer

In this section, It is presented a short review of Bluetooth technology [28],[29],[30],[31]. Bluetooth is a standard for short-range communication, low power consumption, low cost and Wireless, which uses radio technology. The current technology, IEEE 802.15 WPAN is entitled. Bluetooth or blue tooth, brand of wireless connectivity with a close spacing to send messages, photos or any other information that is inspired from the name of a king. Bluetooth technology has several key features that have been broadly utilized. Bluetooth wireless technology is the most successful short-range communication technology «Short Range Wireless Communication» that the billions of devices such as mobile phones, headsets, headphones, medical devices, game consoles, music players and portable video «Portable Media Player », etc. have been used. One of the strengths of Bluetooth is facilitate the communication with other devices in its vicinity. Dissimilar to Wi-Fi «wireless networking standard, 802.11b» that most users make is to manually find a radio signal and then prove their identity, In Bluetooth, the user's task is low. Just two Bluetooth-enabled devices placed inside range of each other and the rest will be done automatically. Big and small, old and young, most people are aware of the Bluetooth and how to work with it. Many mobile phones, digital cameras and printers are equipped with this technology. Bluetooth capabilities, such as wireless and short-range allow to the peripherals for communicate with each other by an air interface. Bluetooth supports both voice and data accordingly, it is an ideal technology in light of the fact that numerous devices are able to communicate together. Bluetooth uses irregular frequency and it is accessible anywhere in the world.

3.3 Behavior of Bluetooth Worms

A typical Bluetooth worm infection cycle comprises of several steps, as shown in fig 1.
Fig. 1. Infection cycle of a Bluetooth worm

At the point when a Bluetooth worm is actuated, it begins searching for Bluetooth-enabled devices in its neighborhood. At this time, the worm broadcasts Bluetooth enquiry packets and waits for reaction. Once the worm gathers a list of Bluetooth-enabled devices in its communication range, according to the list that has collected, repeating the following steps with each neighbor device. Making a connection to it, starting a connection to a nearby device involves the paging process in the Bluetooth communication (step 1), investigate whether a device is infectable regarding the behavior of the worm (step 2). If the answer is yes, copying worm code on a victim device, the time required for duplicate the worm code onto the victim is depend on both the Bluetooth packet type and the size of the worm code (step 3), and end the connection with it, (step 4). Due to the instability of mobile networks, each of these phases may fail without notice of the other end. Thus, a timer is scheduled at each stage so that the worm can discover a connection failure.

4 MPCA

The Cellular automaton mentioned above is a mathematical representation mechanism for modeling epidemic systems. MP-CA is an extended CA structure with some special properties for modeling malware propagation in different communication systems.

4.1 Formal Definition

In all of the previous studies, the proposed modeling approaches for the malware propagation have been allocated to modelling a specific network. As mentioned before, the primary aim of our study is to evaluate the usability of cellular automata in modeling epidemic spread of infection in communication networks. Unlike the previous works, we do not restrict our model to a specific environment and present a model that is capable for modeling the propagation in any network such as: wireless sensor network, smart phone, Ad hoc network and many other complex networks. The dimensions of cellular network is main differences between our proposed MP-CA and its counterparts in which the third or fourth dimension could be time or motion respectively. In this paper, times is considered as the MP-CA third dimension. In following as a case study, the spreading of infection through Bluetooth worm at smartphone is modeled. Comprehensiveness, simplicity, clarity and flexibility are the main parameters of presented model. This model also has the following features: display a history of the malware propagation including address and location of each device, identify the number of infected nodes, identify the position of nodes which have been infected by each infected node on the network, detect the infection source of every infected node, diagnose the time of infection each machine, identify nodes of effective in the further spread of infection, identify high-risk areas, apply precautionary guidelines and adoption appropriate defensive strategies. This information is necessary to understand the behavior of malware.

4.2 Case study

Bluetooth worm propagation modeling in a smart phone network is used as a case study to evaluate and test the proposed model. Due to the spread feature of Bluetooth worms, seven different epidemic statuses of a cell or node are defined:

1-Health state (H): Nodes that are healthy and are not at danger of infection.

2-Vulnerable state (V): nodes have not been infected by any worm in the network but are prone to infection.

3-Exposed state (E): Nodes that have been infected by the worm, but the worm does not
spread to vulnerable Smartphone while it is possible to transfer data or controlling the messages sent to the phones.

4- Infectious state (I): Nodes that have been infected by worms in the network, and they can contaminate some nodes in the state S.

5- Diagnosed state (D): Nodes that have been identified to have been infected by a specific worm.

6- Recovered state (R): Nodes that have been infected by the worm and then have recovered. These nodes have been secured against this worm. In this state, they will not be able to re-infect or transmit infection to others.

7- Quiet state (Q): At the infection state, infected nodes are searching for devices with Bluetooth turned on. Due to the abundant searches the node energy is decreased and enters the quiet state. In other words, Smartphone battery charge is finished. It should be noted that with recharge the battery, the node goes back to infected state again. Process of transition state is shown in figure 2.

**Table 1: Parameters Description**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Probability with which a node in state H becomes a node in state V</td>
</tr>
<tr>
<td>P2</td>
<td>Probability with which a node in state V becomes a node in state I</td>
</tr>
<tr>
<td>P3</td>
<td>Probability with which a node in state V becomes a node in state E</td>
</tr>
<tr>
<td>P4</td>
<td>Probability with which a node in state E becomes a node in state I</td>
</tr>
<tr>
<td>P5</td>
<td>Probability with which a node in state I becomes a node in state Q</td>
</tr>
<tr>
<td>P6</td>
<td>Probability with which a node in state Q becomes a node in state I</td>
</tr>
<tr>
<td>P7</td>
<td>Probability with which a node in state E becomes a node in state R</td>
</tr>
<tr>
<td>P8</td>
<td>Probability with which a node in state I becomes a node in state D</td>
</tr>
<tr>
<td>P9</td>
<td>Probability with which a node in state D becomes a node in state R</td>
</tr>
<tr>
<td>P10</td>
<td>Probability with which a node in state V becomes a node in state R</td>
</tr>
</tbody>
</table>

Let the number of healthy, vulnerable, exposed, infectious, diagnosed, quiet, and recovered nodes at time t be denoted by \( S(t), E(t), I(t), D(t), Q(t), H(t) \) and \( R(t) \), respectively. Then \( H(t) + V(t) + E(t) + I(t) + D(t) + R(t) + Q(t) = N \).

**Fig. 3. Random arrangements of the nodes in a two-dimensional grid M×M**

### 4.3 MP-CA Model for Bluetooth Worm Propagation

For describing worm propagation in a Bluetooth network fine definitions can be expressed:

1. **Cells**: All nodes of a specific network are cells. Namely any node is called as a cell.

2. **Cellular Space**: In this paper, we configure a network (see Fig. 3) that is composed of N smartphones which are randomly arranged on a 2-D grid. Hence, the cellular space is formed by a 2-D array of \( M \times M \) cells or \( M \times N \). Each cell has one wireless node that can establish wireless links only with the nodes within a circular space with radius R around it. The value of radius R determines the transmission range. To simplify the investigation, we assume that the horizontal and vertical coordinates of a wireless node are represented by i and j in the 2-D grid (cellular space). That is to say, cell \((i,j)\) denotes a node located in the situation with a coordinate \((i,j)\) in a cellular network.

3. **State set**: Our model is based on cellular automata. The basic unit of cellular automata is a cell. Each cell can be in one of a finite number of mentioned distinct states at every discrete time. Furthermore, according to the transition rules each cell transforms from its current state to a new state (at the next time step) based on its current state and the states of its neighbors.
In our model a cell signifies an individual with a Bluetooth device. Along these lines, each cell can be represented with the state and probability of dangers for exposure and infection by a worm.

State of a wireless node x which is located in cell \((i,j)\) at time \(t\) as follows:

\[
F_{i,j}(t) = \begin{cases} 
0, & \text{cell } (i,j) \text{ is healthy at time } t, \\
1, & \text{cell } (i,j) \text{ is vulnerable at time } t, \\
2, & \text{cell } (i,j) \text{ is exposed at time } t, \\
3, & \text{cell } (i,j) \text{ is infected at time } t, \\
4, & \text{cell } (i,j) \text{ is diagnosed at time } t, \\
5, & \text{cell } (i,j) \text{ is recovered at time } t, \\
6, & \text{cell } (i,j) \text{ is quiet at time } t. 
\end{cases}
\]

(4) Neighborhood: According to the corresponding transmission range \(R\) the neighborhood of each cell is defined as shown in figure 4. In the general case we assume that the length of a cell of grid is 1 unit. If \(R=1\) unit and Von Neumann neighborhood, each node can have 4 nodes as its neighbors. But if Moore neighborhood and \(R=1\) unit each node can has 8 nodes as its neighbors. It is obvious that with expanding the transmission range, the number of neighbors of the node increases.

(5) Transition function: to describe the spread of malware via Bluetooth in smart phone network, it is necessary that the following factors be considered: First is the Spread Rate (denoted by \(SR_{ij}\)) which indicates the degree of spread of infection from node i to node j \((0 \leq SR_{ij} \leq 1)\). If \(SR_{ij} = 0\) it shows that node i has no infection to node j. If \(SR_{ij} = 1\) this means that the node i has potent infection rate to node j. The next parameter is the Resistance Rate (denoted by \(RR_{ij}\)) which determines the resistance rate of each node against infection \((0 < RR_{ij} \leq 1)\).

If \(RR_{ij} = 1\) it implies that the node i has high ability to resist infection from node j. Let \(TR\) indicate the transmission threshold through which a node transforms from state V to other states. Other factor is Distance (denoted by \(D_{ij}\)) which indicates distance between two nodes. By increasing the transmission range \(R\), the number of available neighbors for any node increases. We assume the nodes with less distance are more likely to be infected by Initial infectious node. Therefore, calculate the distance between each node and the initial infected node is necessary.

Let \(\beta\) denote an infection index which is calculated as a ratio of the interaction factor between cell \((i,j)\) and its neighbors to its resistance rate. Power is the amount of energy in each node. \(RR_{ij}, SR_{ij}, D_{ij}\) and \(\beta\) described as follows.

\[
SR = \frac{1}{g_1 e^{\gamma_1 IC}} \times \frac{1}{g_2 e^{\gamma_2 IC}} \quad (1)
\]

Where \(IC\) is the number of infected neighbors a particular node at each time step. \(t_{max}\) is total time. \(\gamma 1\) and \(\gamma 2\) are constants, which can be determined according to the practical requirement.

\[
RR = \frac{1 - e^{\gamma}}{1 + e^{\gamma}} \quad (2); \quad a \text{ is adjusted factors for } RR_{ij};
\]

\[
D = \sqrt{(k - i)^2 + (p - j)^2} \quad (3); \quad b = \frac{IR}{RR \sqrt{2}} \quad (4);
\]

4.4 Modelling and Simulation Flow

Modeling and simulation of malware propagation in MP-CA goes through the six steps as below. Figure 5 shows this flow more briefly.

Step 1-1: Determine the dimensions of cellular network.

Step 1-2: Determine the transmission range \(R\) according to the cellular network.

Step 2: Initialize network. All nodes are randomly distributed in a two-dimensional grid, and they communicate with each other through short range radio transmissions.

Step 3: Initialize node state. First the states of all nodes is H (i.e. Bluetooth off). By activating the Bluetooth, each node change its state from H to V with probability of \(p1\) (i.e. Bluetooth on). Then among the vulnerable nodes, node i is randomly selected and its state is set to I. The states of other nodes are set to be stated on V.

Step 4: Collect data. Each node collects the information of its neighbors.
Step 5: Assume node x at time t is accessible.
Step 5-1: As to node x, if its state is I (e.g. $F_x(t) = 3$), its neighbor nodes can be accessed.
If the state of its neighbor node y is V (e.g. $F_y(t) = 1$), and if $\beta$ is not smaller than $TR$, node y changes its state from V to E with probability of $p_3$. Due to the variable transmission range $R$, for each value of $R$, the distance of each node with its neighbors is different. Therefore, for the transmission of infection from an infected node to its neighbors, a factor D will be considered (Infection probability of $P_3$). Otherwise, node y remains in the previous state.
If $SR_{xy} < 0$ or $RR_{xy} > 0.01$, node y changes its state from V to R with probability of $p_{10}$. At the same time, node x transforms its state from I to D with probability of $p_8$. If power=0 (e.g. Drain the infected Smartphone battery) node x changes its state from I to Q with probability of $p_5$. If infected device re-charging (battery charging) node x changes its state from Q to I with probability of $p_6$.
Step 5-2: As to node x, if its state is E (e.g. $F_x(t) = 2$), node x changes its state from E to R with probability of $p_7$, or node x changes its state from E to I with probability of $p_4+p_7$.
Step 5-3: As to node x, if its state is D (e.g. $F_x(t) = 4$), node x changes its state from D to R with probability of $p_9$.
Step 5-4: Repeat the beginning of Step 5 until all nodes in the network are accessed.
Step 6: Increase t: $t = t+1$.

5 SIMULATION

To evaluate the feasibility of the proposed scheme using cellular automata and verify the effectiveness and rationality of the proposed model, we simulate the dynamics of Bluetooth worm propagation in smartphone network by MATLAB. The wireless nodes are organized into a grid, and the length of each grid is 1. The total number of nodes (N) is 1000 and the transmission radius R is 1. The other parameters are set as follows: $p_1=0.5; p_2=0.6; p_3=0; p_4=0.2; p_5=0; p_6=0.15; p_7=0.4; p_8=0.5; p_9=0.4$; (All parameters are assumed in dimensionless units).

Figure 6 shows the number of nodes infected by each particular node. Depending on the neighborhood (Von Neumann or Moore) coordinates of each node and the number of infected nodes that is indicated by the node. It is evident that as the transmission radius $R$ increases, the number of infected nodes increases. In this diagram, the node with the most significant impact on infection can be determined.

Figure 7 shows the history of Bluetooth worm propagation in the smartphone network. In this figure, the history of a node is shown. This information includes: source of infection, time of infection, the number of infected nodes by the node, location coordinates of infected nodes in the cellular network. The history is stored in an output file. This information helps us to understand the behavior of malware, preventive strategies and finally apply an appropriate defensive strategy.
Figure 8 shows the evolutions on the number of Healthy, Vulnerable, Exposed, Infected, Diagnosed, Quiet and Recovered nodes. We found that the number of infected nodes increases from \( t = 1 \) to \( t = 38 \) with Von Neumann neighborhoods (\( R = 1 \)), from \( t = 1 \) to \( t = 27 \) with Moore neighborhoods (\( R = 1 \)), from \( t = 1 \) to \( t = 15 \) with Von Neumann neighborhoods (\( r = 2 \)), and from \( t = 1 \) to \( t = 10 \) with Moore neighborhoods (\( r = 2 \)). On the other, number of vulnerable nodes increases initially, so that from \( t = 1 \) to \( t = 10 \) with Von Neumann neighborhoods (\( r = 1 \)), from \( t = 1 \) to \( t = 7 \) with Moore neighborhoods (\( r = 1 \)), from \( t = 1 \) to \( t = 5 \) with Von Neumann neighborhoods (\( R = 2 \)), and from \( t = 1 \) to \( t = 3 \) with Moore neighborhoods (\( R = 2 \)) reach their maximum. It is evident that the number of health nodes and vulnerable nodes decrease as the number of recovered nodes increases. Furthermore, it can be found that the outbreak point is achieved earlier when \( R \) increases.

Figure 9 shows the effects of the transmission range \( R \) on the worm propagation. The maximum value of \( I(t) \) changes proportionally with the node’s transmission range. Namely, a greater transmission range \( R \) yields every node to be infected sooner. It can be observed that the outbreak point is attained earlier when \( R \) is increased. The reason is that a larger transmission radius outcomes in more neighbors for a single node. Accordingly the likelihood of potential infections for the nodes increases as the number of transmission links related with infected nodes increases.

Figure 10 shows the transient response on the number of vulnerable nodes. As time passes, the number of vulnerable nodes first increases gradually and after reaching the maximum point, it decreases slowly to zero. It can be seen that as the probability of \( p_8 \) increases, \( V(t) \) decreases abruptly and hence more vulnerable nodes will be infected. We also found that \( V(t) \) remains same as probability of \( p_3 \) changes.

Finally, figure 11 shows one trend of malware outbreak in which the infected nodes increases gradually until reaches a peak point then drops down slowly. We found that the probability of \( p_8 \) has also a direct relationship with the number of infected nodes \( I(t) \), and the outbreak point can be achieved quickly. We also observe as the infection probability of \( p_3 \) increases, the results change inversely.
Fig. 7. History of the Bluetooth worm propagation

Fig. 8. The number of Health, Vulnerable, Exposed, Infected, Diagnosed, Quiet and Recovered nodes for Von Neumann neighborhoods and Moore neighborhoods where \( R=1, R=2 \).

Fig. 9. The number of infected nodes with different transmission range \( R \) for Von Neumann neighborhoods, Moore neighborhoods.
6 CONCLUSION

In this paper, MP-CA as a theoretical model to investigate and analyze the process of malware propagation in a network is proposed. MP-CA is based on cellular automata. As a case study we have simulated the Bluetooth worm propagation in a smartphone network and achieved comprehensive results. Various parameters have been used in this process including: Spread Rate, Resistance Rate and Distance factor. The simulation results are obtained through artificially chosen parameters which proves the effectiveness of the proposed model. Moreover the results demonstrate that the presented model is a general model which can be applied to any network (of course taking into account the conditions and assumptions of the network).

For the future works, we will focus on:

1-Using MP-CA for different networks, including wireless sensor networks, Ad hoc and complex networks.

2-Since the proposed model does not characterize the impact of node mobility on worm propagation, applying the mobility patterns of the nodes in the network will be our next target.

3-Using a real dataset to test the proposed model.

7 REFERENCES


