Abstract—Cyber attacks are an unfortunate part of society as an increasing amount of critical infrastructure is managed and controlled via the Internet. In order to protect legitimate users, it is critical for us to obtain an accurate and timely understanding of our cyber opponents. However, we lack effective tools to do this at the moment. In this article, we summarize the work on modelling malicious activities from various perspectives, discuss the pros and cons of current models, and present promising directions for possible efforts in the near future for interested readers.

The Internet has become an important part of our society in a number of ways, such as in economics, government, business, and daily personal life. Further, an increasing amount of critical infrastructure (e.g. power grid, air traffic control) is managed and controlled via the Internet, in addition to traditional infrastructure for communication, such as the DNS system for the Internet. However, today’s cyber space is full of attacks, such as distributed denial of service (DDoS), information phishing, financial fraud, email spamming, and so on. We can see cyber space has become a heaven for intelligent criminals, who are motivated by significant financial or political reward. According to an annual report from the FBI’s Internet Crime Complaint Centre, financial loss resulting from cyber attack totaled US$559.7 million in 2009. Symantec identified more than 240 million distinct new malicious programs in 2009, double that of 2008.

At the same time, we are embarrassed to face inquiries from the public, such as who are cyber criminals? and where are they? The reason for this desperate situation is that as defenders, we have few effective tools to identify cyber criminals and their malicious activities. As a result, it is easy for attackers to initiate attacks and fly under the radar, but hard to identify and trace back to attackers for those who wish to defend the Internet.

There are a few important reasons for today’s passive situation for cyber defenders. First of all, the Internet came from the ARPANET, which was designed in the 1960s as a private network with no security component in its original design. In other words, the Internet was born with vulnerability, and this is the root reason for the vulnerability of cyber space, even though many patches have been added to cover the inherent disadvantage. Second, as the largest and most complex man-made system in human history, our understanding of this giant system is limited or even incorrect to date. Due to this fact, the American National Research Council proposed a new research field of network science in 2006, targeting the advancement of our knowledge of networks and networking, and including the Internet as a major object of study [1]. Third, the anarchistic management environment of the Internet is hard to organize a large scale collaboration against cyber criminals.

Botnets have become the dominant malicious networks in today’s cyber space landscape. A malicious network is an overlay network on the Internet, and because of our shallow understanding of the Internet, it is even harder for us to understand malicious networks, such as their structure, size, propagation behavior, and so on.

To date, the majority of current dominant Internet modeling is based on the random graph model proposed in 1959 [2], which is a number of years previous to the birth of the Internet and the Web. Recently, an increasing number of observations indicate there is a great discrepancy between the random graph based models and the reality. Started around the end of last century, new discoveries and models of the Internet and the Web were constantly reported, such as the small world model, the scale free model, and complex networks. Power law was found pervasive in nature, economics and man-made systems, such as individual income among a group of people, or word frequency in a language. The probability distribution of power law is usually expressed as

\[ p(x) = Cx^{-\alpha}, \]

where \( C \) is a constant and \( \alpha \) is called the exponent of the power law.

Scientists have also found many examples of power law phenomenon in cyber space, such as the popularity of web pages and the size of web documents. However, computer scientists are not sure whether power law is dominant in cyber space or not, even doubt about the correctness of network science [3]. One thing is certain that we need invest more energy to this field to find the answer.

In this article, we simply survey the work of malicious activity modeling to the best of our knowledge, and discuss the challenges and opportunities in this research field for interested readers. As botnets are the dominant and typical malicious networks, we will mainly discuss botnet related malicious activities in the context of this article, and use the words botnet and malicious network interchangeably.

Botnet as a malicious network

A botnet is the engine of cyber attacks, and is a typical and dominant malicious network. A botnet is a group of compromised computers (referred to as bots) on the Internet, controlled by botmasters through control and command centers.
(referred as to C&C). There are various kinds of botnets, such as DSNXbot, evilbot, G-Sysbot, sdbot, and Spybot. Botnets are pervasive, existing simultaneously in many commercial, production and control networks. The size of a botnet could be as large as millions. Because of the large number of bots, botnets can be lethal in bringing down targeted networks, such as power grids, air traffic control networks or communication networks.

On top of the complexity of structure and the dynamics of cyber space, botnet owners exhaust their energy in disguising botnet activities and traces against detection and elimination. Attackers have at their disposal state-of-the-art techniques, such as stepping stones, reflectors, IP spoofing, code obfuscation, memory encryption, and peer-to-peer implementation technology to cover and sustain their bots. One critical issue for botnet writers is making sure that all bots contact their C&C center while the physical server and IP of C&C centers frequently change in order to avoid detection or elimination. In order to achieve this, botnet writers, such as Conficker, Kraken and Torpig, have recently developed a new method: DNS “domain fluxing”. As shown in Figure 1, each bot algorithmically generates a large set of domain names and queries each of them until one is resolved. The bot then contacts the corresponding IP address obtained that is typically used to host the C&C server [4]. The current method against DNS domain fluxing is to catch bots using honeypots, and use reverse engineering to obtain the URL generation algorithm. However, this is time consuming and has a low rate of accuracy. It is also ineffective to fight against quick changing botnets. Besides the techniques hackers are using, the duration of botnet activities is short and random, making it harder for defenders to collect botnet related data.

![Fig. 1. A bot connects to its frequently changing C&C server via domain fluxing technique.](image)

Botnets have been investigated using various perspective for around a decade. Some researchers focused on analyzing botnet characteristics, such as IP address distribution, whois records and lexical features of phishing and non-phishing URLs. Statistical learning techniques were employed to study lexical features of URLs (length of domain names, host names, number of dots in the URL, etc.) to automatically determine if a URL is malicious, i.e., used for phishing or advertising spam. Traffic analysis and signatures are also powerful methods for attack detection. Network telescopes have been employed to observe malicious traffic at various vantage points in networks. It is expected that infiltrated or subverted machines (acting as bots) will contact the botmaster at regular time intervals, which can yield an opportunity for their detection.

Based on the literature, it is not difficult to see that our understanding of botnets and their malicious activities is still at an early stage. There are many questions needing to be answered in an accurate and timely manner. We therefore discuss the challenges of establishing accurate and effective models for malicious networks from different perspectives. Moreover, promising directions and tools are also presented for interested readers.

**Observation on malicious activities**

The first challenge to establish effective models is to measure the object of study correctly and collect sufficient data of malicious activities. For business, privacy and security reasons, it is hard to collect attack data from ISPs and related companies. Available data sets are usually collected by honeypots at a limited number of locations on the Internet. There are a number of problems in this field. For example, the collected data is usually not the data we expect. For instance, the observation range is not what we desire, or some information is missing. A global monitoring system for Internet measurement is definitely needed. We have seen some examples of this kind of system in place, such as, the planet lab (planet-lab.org), which is an open global research network with 1137 experimental nodes all over the Internet currently. Moreover, we do not have sufficient storage space to save the large amount of network related data, such as traffic traces. There have been many researches on data compression trying to address the problem, such as the principle component analysis (PCA in short). A promising mathematical tool, compressed sensing [5], has been recently invented. This new methodology significantly outperforms the existing tools in terms of space efficiency, and has attracted a great attention. Furthermore, we desire some mathematical tools to infer a relatively complete picture of networks with limited and partial observation of a studied object.

The next challenge in this category is data processing. The raw data set we collected is a mixture of multiple malicious networks with high probability. For example, there are many botnets that coexist at the same time, and the DNS request failure data includes requests from bots of different botnets. In order to extract the features of one botnet, we need to separate the mixed data set into clusters that respond to each individual botnet.

The challenge here is that we do not know how many botnets there are in a collected data set. As different botnets are written by different authors using different communication strategies, different botnets behave differently. In other words, bots of the same botnet behave similarly. The similarity could
be identified in temporal, spatial or other features, and similarity can be used to differentiate botnets.

Unsupervised machine learning is an existing and promising tool for the clustering challenge. The approaches of unsupervised learning include two categories: clustering and blind signal separation. Researchers have proposed many algorithms for this research field, such as, singular value decomposition, mixture models, k-means, and hierarchical clustering. Besides these traditional methods, we also notice a recently developed technique, graph spectrum [6], which is also promising to address the challenge: A graph of connections among nodes can be established among the nodes in the mixed data set. Based on the graph, we can obtain an adjacency matrix that can be further transformed into the spectra space, where the nodes that belong to the same botnet will establish a straight line in theory. As a result, we can accurately separate different botnets in the spectra space.

**Topology modeling of malicious networks**

It is a significant challenge to perform topology modeling for the Internet and malicious networks. Research from physicists shows that structure determines functions. Therefore, it is especially important for us to understand the topology of botnets or other malicious networks. If we know the topology of a given botnet, then we can figure out the key nodes of the network. As a result, we can work with limited ISPs or organizations to defeat botnets, e.g. terminate possible attacks or block a communication path of bots.

However, our understanding in this aspect is extremely limited so far. The possible reason is that the data we have is usually “flat”. For example, when we catch a malicious packet, we can only know its source IP address and destination address, however, the path from the source to the destination is usually hard to obtain.

Graph theory is a traditional and good tool for network topology modeling. Network tomography [7] is another popular method for network topology research. Similar to medical tomography, probation packets are injected into an unknown network, and the output is collected at the other side of the unknown network. We then infer the topology of the network based on the input and the output. However, both of these methods study static graphs or networks. It is easy to notice these tools are not sufficient to model the ever changing Internet or malicious networks.

Moreover, we have noticed that many current network modeling is too loyal to their underneath physical network nodes and links. As shown in Figure 2, a network possesses a source S, a destination D, and three intermediate nodes A, B, and C. It is simple to model the network according to its physical connections, but it increases the complexity of following work, such as interpreting phenomenon and seeking solutions. We can also eliminate intermediate node A and C from the model if they have ignorable impact on the system. A simplified logical topology modeling will dramatically benefit our following work. Of course, it is challenging on how to establish the abstract network.

**Dynamics modeling of malicious networks**

Botnet dynamics include many aspects, with the most important being the number of bots of a given botnet against time. This simply means the size of the botnet. This information is valuable to defenders, as defenders can organize their defense and budget their costs better with this information in place.

There is some work on the size of botnets, however, researchers do not have a solid model for this issue. A direction method to obtain the number of bots is performing botnet infiltration. Stone-Gross et al. [8] registered the URL of the Torpig botnet before the botmaster, therefore hijacking the C&C server for ten days, and collecting about 70G data from the bots of the Torpig botnet. They reported the footprint of the Torpig botnet was 182,800, and the median and average size of the Torpig's live population was 49,272 and 48,532, respectively. They found 49,294 new infections during the ten day takeover. Their research also indicated that the live population fluctuates periodically because of users switching between online and offline.

Another method uses DNS redirection to capture bots by honeypot, which means a C&C server can be identified using source code reverse engineering tools. Researchers reported the footprint of the studied botnet can reach 350,000. There are also negative opinions on the accuracy of botnet size studies. Opponents point out that the extensive employment of the DHCP and NAT techniques can result in inaccuracy of previous methods.

The main challenge in this field is: For a given botnet or a malware and a given range of the network, what is the density...
of bot or malware distribution in the network? There has been plenty of research concerning the recruitment of malicious networks, such as botnets and viruses, based on epidemic theory. However, the research on malware or bot distribution is limited. To date, we only know that the distribution is non-uniform based on information theory, and that network topology has a big impact on the spread of malware.

The dominant model for the size issue is the epidemic model, which is the major theory for biology virus propagation modeling, which is expressed as

\[
\frac{dI_t}{dt} = \beta I_t(N - I_t),
\]

where \( I_t \) is the number of infected hosts at time \( t \), \( N \) is the total vulnerable hosts in the population, and \( \beta \) is the pairwise rate of infection. The solution of equation (2) is

\[
I_t = I_0 \cdot e^{\beta N t},
\]

where \( I_0 \) is the number of initial infected hosts.

Computer scientists have applied this model in studying computer virus propagation. As the member recruitment of botnets is essentially the same as computer viruses, the usage of the epidemic theory looks effective to model the size of the botnet. However, researchers have noticed that the current computer virus propagation model lacks accuracy after the early stage of propagation. Therefore, it is necessary to revisit the accuracy of the epidemic model. It is not possible to collect detailed data of a biological epidemic, and this problem was not found in the past. However, the Internet offers a possible platform for computer scientists to perform large scale experiments and collect sufficient data to work on this issue. Furthermore, the findings from the computer field can be applied to the medical field. In our opinion, this is a very promising field to continue working on.

As botnet dynamics are mainly related to time, time series analysis methods are probably effective to address this problem. Many questions remain unanswered, e.g., periodicity, frequency of various bot recruitment and attacking activities, what is the distribution of a specific botnet or virus, and how many Internet nodes have been compromised since the beginning of a botnet?

**Convert malicious activity modeling**

There are many intrusion detection and virus detection algorithms in place, however, there are only limited malicious activity detection algorithms in the literature. Researchers do not know how many illegal activities go undetected using current detection systems. The false negative rate is an essential challenge for us, since attackers are exhausting their efforts to disguise their malicious traces. In some cases, malicious bots demonstrate decent behaviour most of the time in order to fool our detection systems.

In order to address related issues, it is necessary to integrate understandings of human criminal behavior with information technology techniques to reduce the false negative rate of detection as much as possible. For a long time, the network security community has focused on technology oriented methodologies, and ignored the human aspect of criminal behavior, which greatly enhances our understanding of criminals. At the moment, we believe game theory and social network technology can be effective tools to address problems in this category. We list two examples here.

1) Identifying the boundary of detection for a given level of security investment using game theory. It is obvious from an attacker's point of view that high frequency of malicious activity results in a high probability of being detected. For example, frequent vulnerability scanning or sensitive data downloading will make the compromised computer stand out from its peers. There is a threshold at which malicious activity is far more likely to be detected. Presently, the network security research community has no conception of where this boundary lies. It is worthwhile to explore this bound between detectable and undetectable using game theory and identify the Nash Equilibrium (if the Nash Equilibrium exists). With the bound information in hand, we can actually estimate the false negative probability in detection. With this information in place, researchers can develop a strict low false negative detection algorithm, which can push the threshold to a minimum, consequently suppressing the frequency of malicious activities.

2) Identifying malicious nodes using social network technologies. In general, we can divide all Internet based nodes into two groups: benign and malicious (e.g., members of one specific botnet). It has been proven that communication among nodes within each group is quite rich, however, there is much less communication among nodes from different groups. Therefore, for a given node, the probability that the node is malicious increases if the node has a certain amount of communication with the known malicious nodes.

**Forensics of malicious activities**

Cyber forensics is an attractive topic, and is extremely important as we have more and more killer applications in cyber space. However, the work on this field is not very substantive. One solid topic is IP traceback, which refers to the capability of identifying the actual source of malicious packets sent across the Internet. Current methods of traceback rely on independent local networks with no global coordination. This means they are incapable of accurately tracing back cyber criminals at the Internet level. We can categorize the methods of IP traceback into three major groups: deterministic packet marking (DPM), probabilistic packet marking (PPM), and the information theoretical based method [9]. The first strategy marks IP packets at the source local area network where the packets are generated, whereas the second strategy marks incoming packets at the edge routers of the local area network where the potential victim resides. Both of these strategies require routers to inject marks into individual packets. Moreover, the PPM strategy can only operate in a local range of the Internet (e.g., ISP networks), where the defender has the authority to manage. However, these kinds of ISP networks are
generally quite small, and we cannot traceback to the attack sources located out of the ISP network. The DPM strategy requires all the Internet routers to be updated for packet marking. However, with only 25 spare bits available in an IPv4 version 4 packet, the scalability of DPM is a huge problem. Moreover, the DPM mechanism poses an extraordinary challenge in storage for packet logging for routers. Therefore, it is not feasible in practice at the present time. Further, both PPM and DPM are vulnerable to packet pollution from hackers. The third method measures the variation of flow entropy at the routers to traceback the attack source. This overcomes the disadvantages of the previous two, however, it needs global collaboration, which is hard to achieve.

Attack source inferring is an applicable method for today's cyber environment because direct traceback is almost impossible. In this case, the Bayesian inference network is probably a good choice. The research community desires effective and efficient tools to carry out cyber forensic tasks.

**Summary and further discussion**

In this article, we highlight that our understanding of malicious networks and their activities is limited because of a number of reasons. We have realized that the random graph based modeling method is inappropriate for cyber space, which is one reason why it is hard to win the hide-and-seek game against cyber criminals. However, researchers have discovered a number of new tools and models to deal with this desperate situation, such as network science, compressed sensing, and graph spectra. We have seen a number of promising directions to explore cyber security.

We have to note that cyber security is a wide concept. In this paper, we confine our discussion on botnet, which is a major engine behind many malicious cyber activities. There are many other cyber security related topics, such as privacy attacks on web browsing, information water marking, and information hiding, we skip them due to space and our knowledge limitation. Moreover, attacks on encrypted content or protocols is a new and challenging part for information safety in cyber space. However, this is out the scope of this paper, we therefore refer interested readers to the work from network defense aspect, such as [10].

The cyber community and its research progressed for the first 20 years. At the first stage of such a complex and giant system, the development was dramatic and unstable from a system's point of view. We have experienced confusion and misunderstandings, but now is a time we see a relatively stable Internet, and invite new tools and theories for the Internet and its cyber security. If we look at the history of electronic communication, Shannon found and developed information theory more than a half century after people started communicating using electronic media. Will a similar thing happen to the Internet and the Web? Time will tell us.

**References**