Received May 10, 2015, accepted June 13, 2015, date of publication July 20, 2015, date of current version August 3, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2458581

Detecting APT Malware Infections Based on Malicious DNS and Traffic Analysis

GUODONG ZHAO[1], KE XU[1,2], (Senior Member, IEEE), LEI XU[1], AND BO WU[1]

[1]Tsinghua University, Beijing 100084, China

[2]Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China

Corresponding author: G. Zhao (zgd12@mails.tsinghua.edu.cn)

ABSTRACT Advanced persistent threat (APT) is a serious threat to the Internet. With the aid of APT malware, attackers can remotely control infected machines and steal sensitive information. DNS is popular for malware to locate command and control (C&C) servers. In this paper, we propose a novel system placed at the network egress point that aims to efficiently and effectively detect APT malware infections based on malicious DNS and traffic analysis. The system uses malicious DNS analysis techniques to detect suspicious APT malware C&C domains, and then analyzes the traffic of the corresponding suspicious IP using the signature-based and anomaly based detection technology. We extracted 14 features based on big data to characterize different properties of malware-related DNS and the ways that they are queried, and we also defined network traffic features that can identify the traffic of compromised clients that have remotely been controlled. We built a reputation engine to compute a reputation score for an IP address using these features vector together. Our experiment was performed at a large local institute network for two months, and all the features were studied with big data, which includes ∼400 million DNS queries. Our security approach cannot only substantially reduce the volume of network traffic that needs to be recorded and analyzed but also improve the sustainability of the system.

INDEX TERMS APT, malware infections, DNS, intrusion detection.

I. INTRODUCTION

Advanced Persistent Threat (APT) attacks are increasing on the internet nowadays; Unfortunately, they are hard to detect. It is a set of stealthy and continuous hacking processes targeting a specific entity with high-value information, such as government, military and the financial industry. The intention of an APT attack is to steal data rather than to cause damage to the network or organization. Once hacking into the network has been achieved, the attacker would install APT malware on the infected machine. APT malware, for instance, trojan horse or backdoor, is tailored for anti-virus software and firewalls of the target network. It is not only used for remotely controlling the compromised machines in the APT attack, but also for stealing sensitive information from infected host over an extended period of time. APT malware can evade anti-virus software using polymorphic code, and bypass firewall using protocol on allowed ports.

DNS is an important component of the Internet, and it is the protocol that is responsible for resolving a domain name to the corresponding IP address. Unfortunately, besides being popular for benign use, such as helping to locate web servers and mailing hosts, domain names are also susceptible to malicious use. To remotely control the infected machine, attackers need to build a command and control channel. The command and control channel between the infected machine and the attacker is responsible for sending commands and transferring data. Most malware, such as trojan, backdoor and other remote access tools, makes use of domain names to locate their command and control (C&C) servers and communicate with attackers. For example, famous malware such as Gh0st, PCShare and Poison Ivy, all instructs the attackers to create domains and ports for locating command and control servers firstly.

In an APT attack, the malware needs to maintain a persistent connection to a C&C server. DNS is widely used by the attacker to locate command and control server of the malware. Because if the attacker hardcode the IP of the C&C server into the malware binary, it would cause some kind of failure that can not be recovered. Once the C&C server goes down or the IP address is detected, the compromised machine would be out of attacker's control. Another reason is that, to hide the real attack source, attackers often use the servers they have controlled or managed in different countries and regions as proxies. Since using domain names is flexible to change the IP addresses of the malware C&C servers and migrate

the C&C servers, it helps the attacker to hide the true attack source behind proxy server more easily.

By analyzing lots of malware samples in virtual machines, we found that malware such as trojan and other Remote Access Tool (RAT) often uses DNS especially Dynamic DNS to locate command and control server. Dynamic DNS is a method that can update a name server in real time. The dynamic DNS providers own lots of existing 2LD domains. The user only needs to register a new 3LD sub-domain, and maps an IP address to the new dynamic domain name that is registered. A new 3LD sub-domain that is not registered before can be easily registered. For the malware such as trojan, DDNS is a natural fit. The primary convenience of dynamic DNS is that, the user can change the domain to point to a new IP address at any time. There are plenty of dynamic DNS providers such as NO-IP and DynDNS, and most of them are free.

In this paper, we aim to detect APT malware which relies on DNS to locate command and control servers. Previous researches have studied how to detect botnets through the analysis of DNS traffic (see [1]–[3]). These researches focused on detecting malicious flux services or bots that make use of domain generation algorithm (DGA). Malicious flux service works similarly with content-delivery networks (CDN) service. It makes use of the same theory as CDN. CDN now is a common method to accelerate delivery of content of websites and reduce web server lag. It is a network that consists of large numbers of machines resided in different countries and regions. Whenever a user sending a request to the web server that is part of CDN network, the nearest server is going to respond the website visitor. CDN is an effective method to accelerate content delivery of web servers. Malicious flux service is a DNS technique used by botnets. The difference between content-delivery networks and malicious flux network is that, the CDN consists of large numbers of legitimate servers, and the malicious flux network consists of large numbers of infected machines. Conficker [4] and Kraken are the recent botnets that make use of malicious flux, for being more resistant to detection and discovery. The chance to discover and take down the botnets can be reduced by using malicious flux service.

Domain Generation Algorithm (DGA) can be used to generate a large number of domain names [5]. It is popularized by many malware and botnets, such as Srizbi bots [6] and the Conficker worm [4]. The infected machines will generate large numbers of domain names everyday, for example, Conficker worm generate 50,000 domain names every day for communicating with the C&C servers. The domain name for C&C server is chosen randomly from the domain name list.

APT malware is very different from the bots and worms mentioned above. The primary purpose of APT malware is to remotely control the machines and to steal confidential data, rather than to launch denial-of-service attacks, send spam emails or cause damage. It requires a high degree of stealth over a prolonged duration of operation. For example, in the case of those bots and worms, the attackers need to use the command and control servers to remotely control thousands of infected hosts. But APT attackers do not use the same C&C server to remotely control so many infected end-user machines, because it would increase the risk of exposure. The crafted malware is only used for the end-user machines which are valuable to them.

The DNS behavioral features of APT malware are very different from malicious flux service and DGA. Flux service and DGA domains have some obvious features. For example, ''short life'' feature is extracted from domains that are generated by a domain generation algorithm. Because the DGA domains are used only for a short duration [7]. ''Alphanumeric distribution'' is also is feature that is extracted from the DGA domains, because a DGA domain do not include ''meaningful'' words [7], [8]. For example, malicious flux DNS traffic also has a obvious common feature, that is the IP addresses resolved to the domain name are varied and changed rapidly.

To identify malicious domains that are involved in APT malware activity is a challenge. The crafted malware in APT attack do not use malicious flux service or DGA domains. The domains for APT malware were registered by the attackers. Compared with these bots and worms, crafted malware requires high degree of stealth. For this reason, the DNS behavioral features of APT malware are unconspicuous. It is too hard to analyze large volumes of inbound and outbound traffic in a large network, such as a large enterprise and an ISP. To detect APT malware infections in a large network is another challenging problem.

In this paper, we propose a novel system ''IDnS'' placed at the network egress points to detect APT malware infection which relies on DNS to locate command and control servers. The main contributions of this paper are as follows:

- We present a novel system placed at the network edge using a combination of malicious DNS detection technology and intrusion detection technology to detect malware infections inside the network. This approach can not only largely reduce the volume of network traffic which needs to be recorded and analyzed, but also improve the sustainability of the system.
- We define 14 APT malware C&C server domain features including dynamic DNS features by studying large volumes of DNS traffic which can be called big data. 7 Of them have not been proposed before in previous works. And abnormal network traffic features are also defined to help identify the traffic of compromised clients that have been remotely controlled.
- We build a reputation engine to decide whether an IP address is infected or not by using these feature vectors together.

II. RELATED WORK

A. DNS MALWARE STUDIES

Researchers have recently proposed the method of identifying malicious domains through DNS traffic analysis. Notos [9]

build a reputation engine for dynamically assigning a reputation score for a new unknown domain to judge whether it is malicious or not. EXPOSURE [7] studied DNS lookup behavior within a local domain below the DNS resolvers to detect domains for malicious use, such as domains used for malicious flux, adult website, spam mails, phising and malware. In paper [10], it gives a summary of the system EXPOSURE [7] which is using passive DNS analysis to automatically detect malicious domains. Compared to previous researches Notos [9] and Exposure [7] , which are based on monitoring DNS traffic from local recursive DNS servers, Kopis [11] offers a new vantage point and introduces new traffic features specifically chosen to leverage the global visibility obtained by monitoring network traffic at the upper DNS hierarchy. It can accurately detect malware domains by analyzing global DNS queries. [12], [13] analyzed DNS lookup behavior at a DNS root server. Castro *et al.* [14] and Brownlee *et al.* [13] attempted to characterize how much DNS traffic at the DNS root server was illegitimate. Gao *et al.* [15] propose a novel approach that looks at the co-occurrence and sequence in domain names. It isolates malicious domain groups from temporal correlation in DNS queries, but it needs known malicious domains as anchors.

Approaches for detecting malware activity by monitoring and analyzing DNS traffic were also studied too. Some approaches focused on detecting botnets which make use of malicious flux service. Perdisci *et al.* [16] aimed to detect malicious flux services by analyzing recursive DNS (RDNS) traffic from multiple large networks. [8] developed a system placed at the network edge to detect and mitigate botnet infections on a network through detecting malicious flux domains from DNS traffic. Unfortunately, the crafted malware in APT attack do not use malicious flux service or DGA domains. [17] propose an anomaly-based mechanism to detect botnet through monitoring and analyzing DNS traffic. The mechanism rely on detecting group activities in DNS queries simultaneously sent by distributed bots. The authors proposed features to distinguish DNS traffic generated by botnets and benign clients. But they only focus on the group activity property of botnet, and the features they identified are not fit to detect APT malware.

No previous work has tried to identify malicious domain names involved in APT malware activity. In this paper, we focus on detecting C&C server domain names for crafted malware in APT activity. We extracted 14 APT malware C&C domain features including features of malicious DDNS, and 7 of them have not been proposed before. We place the system which is called ''IDns'' on the edge of the network and also do the network traffic analysis to detect infected machines inside the network.

B. INTRUSION DETECTION STUDIES

In general, the main studies of network intrusion detection include signature-based detection and anomaly-based detection. Signature-based detection is a technology that relies on a existing signature database to detect known malware infections. By using signature-based detection technology, it can identify malware C&C communication traffic through signature-based pattern matching. So for malware infection detection, it is a typical approach. But signature-based detection technology has a fatal drawback, it can not detect new malware infections if the signature of the new malware is not in the existing signature database.

Snort [18] is a famous signature-based network intrusion detection system. Snort has many rules in the VRT for detecting malicious code and suspicious network activity. And these are excellent sources with many excellent rules for detecting a wide range of threats including malware. Snort is singled out in this paper because of its popularity and its familiarity. Previous researches have focused on the advantages and drawbacks of Snort [19], [20]. There are low false positives as long as the attacks are clearly defined in advance, but it is hard to detect newer or unknown attacks.

Anomaly-based intrusion detection [21], [22] is a technology that detect abnormal behaviors that deviates from ''normal'' behaviors. The ''normal'' behaviors of the network need to be studied and identified at first. The primary advantage of anomaly-based intrusion detection is the capability to detect new or unknown attacks. Because the new or unknown malware whose signature is not available would also generate abnormal behaviors. The primary drawback of anomaly-based intrusion detection is that, it is more prone to generating false positives. Because the behaviors of different networks and applications are so complicated, the ''normal'' behaviors is very hard to accurately identify. Different from signature-based detection, anomaly-based intrusion detection is a broader match which is based on detect abnormal behaviors. Many legitimate applications perform the same abnormal behaviors as malicious ones.

Mcafee [23] showed us the advanced detection technique to identify APT malware command and control (C&C) communications traffic. It also analyzed the traffic that generated when APT malware communicate with the C&C servers, and extracted some network features of serval APT malware including variant Gh0st and Poison Ivy. Kaspersky Lab [24] introduced many APT malware and cyber campaigns including ''Equation'', ''the Mask'', ''Black Energy'' and other famous APT activities. Their reports include the C&C domains and C&C server IP addressed of the APT malware.

III. OVERVIEW OF THE APPROACH

A. EXTRACTING FEATURES FOR DETECTION

IDns is designed to detect malicious domains used for crafted malware in APT attacks and to detect infected machines. For this purpose, we did analysis of large volumes of DNS traffic which can be called big data. And we also analyzed the network traffic of large numbers of suspicious malware C&C servers.

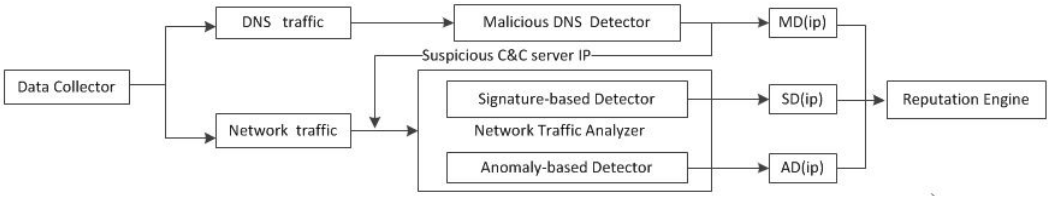


FIGURE 1. Architecture of the System.

TABLE 1. Feature sets(* = new features).

FeatureSet	#.	FeatureName	
Domain Name-Based Features	1	Contain Famous Name	*
	2	Contain Particular Name	*
	3	Contain Phishing Name	*
DNS Answer-Based Features	4	Silent IP	*
	5	Number of distinct IP addresses	
	6	Number of distinct countries	
	7	Number of domains share the same IP with	
	8	IP in the same Class B range of known C&C servers	
Time Value-based Features	9	Daily similarity	
	10	Same query numbers in same time window	*
	11	Very Low frequency query	*
TTL Value-based Features	12	Average TTL	
Active Probe Features	13	Web server or not	*
	14	Whois information	

The features we extracted from big data for detection consist of malicious DNS features and network traffic features. By studying the DNS traffic, we achieved to extract distinguishable DNS features that are able to define the APT malware C&C domains. By studying the behaviors of the crafted malware and benign applications, we achieved to extract distinguishable network traffic features that are able to define the APT malware C&C traffic. Network traffic features, including signature-based detection features and anomaly-based detection features, can help to identify the traffic of compromised clients that have been remotely controlled by crafted malware.

B. ARCHITECTURE OF THE SYSTEM

Figure 1 shows us the architecture of system ''IDnS''. It consists of four main units:

Data Collector: It is placed at the network edge to record the inbound and outbound traffic produced by the network.

Malicious DNS Detector: It is responsible for analyzing the inbound and outbound DNS traffic produced by the network, and detecting suspicious APT malware C&C domains. It would detect the suspicious APT malware-related domains and provide corresponding suspicious C&C server IP addresses for the ''network traffic analyzer'' of the system.

Network Traffic Analyzer: It consists of signature-based detector and anomaly-based detector, for analyzing the network traffic of suspicious C&C server IP addresses. The signature-based detector has defined C&C communication traffic signatures for detecting malware known to the system. The anomaly-based detector detect anomalous behaviors including protocol anomaly, statistical anomaly, application anomaly etc. When the unknown or new malware was identified by anomaly-based detector, new signatures will be defined. All the C&C communication traffic signatures which have been identified will be collected to our TM (Targeted Malware) family.

Reputation Engine: It aims to compute a reputation score for an IP address to judge whether the host or server owning the IP address is infected or not, by using malicious DNS and network traffic feature vectors together.

IV. MALICIOUS DNS FEATURES

In this paper, we identified 14 features to detect APT malware command and control domains (see Table 1) based on big data. 7 Of the features have not been proposed before. And we also give new explanations of some old features that have been proposed before. In this section, we will elaborate on the 14 features that are proposed in this paper and explain the reasons that they can be used for detecting APT malware command and control domains.

A. DOMAIN NAME-BASED FEATURES

Every single domain name is separated to serval parts by period. The last part is called the top-level domain (TLD). The second-level domain (2LD) is the last two parts. The third-level domain (3LD) is the last three parts, and so on. For example, given the domain name ''a.b.c.com'', TLD of the domain name is ''com'', 2LD of the domain is ''c.com'', and 3LD of the domain is ''b.c.com''. For a dynamic DNS, 2LD ''c.com'' is existing part owned by the DDNS provider. The third level sub-domain ''b'' in ''b.c.com'' is created by the users. We extracted three domain name-based features, the third level sub-domain name of

DDNS (dynamic domain name) contains famous name, particular name or phishing name. In previous researches, these 3 features for malware C&C dynamic domain names were not been proposed ever.

Contain Famous Name: We find it interesting that many dynamic domain names registered for C&C servers can tell us they are highly suspicious themselves. We can tell them from the legitimate ones just by the name. Just like we can tell that a long haired man wearing a police uniform is a fake police. Many registered suspicious dynamic domain names contain famous domain names such as windows, yahoo and taobao. And we know that there is little chance that these dynamic domains are used for Microsoft, Yahoo or Alibaba.

Contain Particular Name: We also find that many dynamic domain names registered for C&C servers contain some particular name, such as ''web'', ''mail'', ''news'' and ''update''. These particular names not only make these domain names easy to remember, but also make these domains more like normal ones. And as observed, the particular name and the famous domain name are usually used together, such as ''yahoomail.xxx.com'', ''yahoonews.xxx.com'' and ''windowsupdate.xxx.com''.

Contain Phishing Name: Phishing is a technology that is usually used in social engineering attacks. The attacker tricks the victim to access a crafted fake website which is malicious. When the victim accesses the phishing website, which will try to install malware on the victim. For tricking the users, we all know the phishing domain has a similar name to a legitimate one. Such as ''youtuhe.com'' compared to ''youtube.com'', ''yah00.com'' compared to ''yahoo.com'', etc. we observed that many malicious dynamic domain names used for command and control server of RAT tool not scam server also have a phishing similar name to a legitimate domain one.

B. DNS ANSWER-BASED FEATURES

Silent IP: To hide the C&C server and C&C network traffic, when attackers do not need to send commands to a victim machine, they do not want the domain names to point to the C&C server. For that moment, attackers usually change the domains to point to some specific IPs. Specific IP addresses are usually as follows: 127.0.0.1 (loop back address); 192.168.x.x, 172.16.x.x, 10.x.x.x (private address); x.x.x.255 (broadcast address).

Predefined IP: Some advanced malware in APT attack improved this method. When the attackers were developing and coding the advanced malware, a predefined IP was hard-coded into the malware binary. The silent mechanism works like this, when the domain is resolved to the predefined IP, the malware would turn to silent-mode and would not initiate a connection until the domain is resolved to another IP address. Predefined IP addresses are usually some invalid IP addresses that have obvious features, such as 5.5.5.5, 2.3.3.2. In this paper, this specific IP addresses and predefined IP addresses are all called Silent IP. The feature of predefined IP has not been proposed before in previous research.

Number of Distinct IP Addresses&Number of Distinct Countries: To hide the true attack source, attackers usually use servers reside in different countries or regions they control or manage as C&C servers. To the attacker, C&C servers better not reside in the same country of the attacker or the victim. Because if C&C servers reside in the same country of the victim, it is easier for the victim country to analyze this attack. If C&C servers reside in the same country of the attacker, it is easier to trace the real source. These 2 features have been used in previous work to detect botnets domains (see [12], [16]).

Number of Domains Share the Same IP With: This is also a feature that EXPOSURE [7] proposed before. And we study and train this same feature to detect malicious dynamic DNS, it works as well. In the APT attack scenario, a single attacker seldom own more than 30 dynamic names to locate command and control server in the same time, because it is not necessary and it is hard to maintain them. So the number of malicious dynamic domains share the same IP with is defined less than 30.

IP in the Same Class B Range of Known C&C Servers: We have performed statistical analysis of numbers of C&C servers that have been detected. The result shows that, there are many C&C servers in the same Class B IP addresses range and even in the same Class A range. There may be two reasons for this. The first is more and more APT attackers rent VPS servers as C&C servers. Because VPS server is stable, hard to trace back and easy to manage. VPS servers rented from the same service provider are mostly in the same Class B IP addresses range and even in the same Class A range. The second reason is, some advanced attackers constructed special network for C&C servers.

C. TIME VALUE-BASED FEATURES

Daily Similarity: This feature is proposed before in EXPOSURE [7]. They check if there are domains that show daily similarities in their request count change over time, an increase or decrease of the request count at the same intervals everyday. In our detection, we check if the domains have daily similarities in changing IP address at the same intervals everyday. For example, organized APT attackers usually change the domains to point to C&C servers at the start time of one-day work hours, and change the domains to point to silent IP at the end time of one-day work hours. Some malware typically connect to C&C servers at same intervals of everyday, monitoring consistent intervals for DNS requests will help.

Same Query Numbers in Same Time Window: This feature means in the same time window, the number of domain queries are about the same. When the infected host is online, but there is a connection failure for some reason. The infected host will mistaken DNS errors and send large amounts of repeated DNS queries.

Very Low Frequency: This is a new feature that has not been proposed ever in any previous work. We found that a few high advanced APT malware query domains to locate command

and control server at very low frequency, at one time for serval days or even serval weeks. We believe behaviors of these domains are well-designed for avoiding the detection by advanced APT attackers. As observed, these domains have other common features as well. Most of the domains are all web servers, and these web servers have common features in website content and design as well. The resolving IP address and Time To Live (TTL) of the domain name are all stable.

D. TTL VALUE-BASED FEATURES

Time To Live (TTL) is set by an authoritative name server for a DNS record. TTL means how long the a resolver may cache the response result for a domain. If a stub resolver queries the caching nameserver for the record before the TTL has expired, the caching server will simply reply with the already cached resource record rather than retrieve it from the authoritative nameserver again.

Average TTL: Setting TTL values of host names to lower values can help the attacker to change the C&C server rapidly. Moreover, based on our measurements, TTL values of Dynamic Domain Name Service, such as DynDNS, NO-IP and ChangeIP, are usually set to 30, 31, 60, 300 seconds. But not all the malware C&C domains set TTL values to lower values. As we mentioned in ''Very low frequency'', there are advanced malware domains setting higher TTL values, such as 86400 seconds as observed. Because they do not change the resolving IP address for weeks.

E. ACTIVE PROBING FEATURES

The malicious DNS features listed above are all based on passive analysis. In this part, we propose active probing methods to assist detecting malicious domain.

Web server or not: We propose this new method to probe domain's 80 port and check it is a true web server or not. If a domain keep TCP port 80 open but not a web server, it is highly suspicious. We can check whether it is a command and control server keep TCP port 80 open listening for the infected host to connect by analyzing response packets. But if it is a web server, we can not confirm it is a command and control server or not. Because attackers can use a web server as a command and control server.

Whois Information: By querying Whois, we can get more information of the domain name, such as the registration date, the registrar, the registrant name, the registrant email and the registered country. Comparing these information with the whois information of previous known malware C&C domains is an effective method. For example, all C&C domains of the famous APT malware ''Equation'' appear to have been registered through the same two major registrars, using ''Domains By Proxy'' to mask the registrants information.

V. NETWORK TRAFFIC FEATURE

Our system IDnS uses signature-based detection and anomaly-based detection together to provide the maximum defense for the monitoring network.

A. SIGNATURE-BASED DETECTION FEATURES

Ruleset plays a crucial role in signature-based IDS, and the number and accuracy of the rules determine how much infections can be detected. To apply publicly open rule sets of well known signature-based IDS, we use rules from VRT Rule sets [18] of snort. Our system focuses on detecting malware infections, so the rules applied to the system are mainly from malware-backdoor rules, malware-cnc rules, malware-other rules and blacklist rules. After a long detecting period, the system has detected and confirmed a lot of malware infections by malicious DNS detection combined with anomaly-based detection.

Signature-based detection features we mentioned in this paper means the features of C&C network traffic generated when malware communicate with C&C servers. By analyzing the network traffic produced when the malware communicate with command and control servers, we extract network traffic communication features of 21 unknown malware or trojans. We attribute the unknown malware to our TM (Targeted Malware) family, So all the malware in our TM (Targeted Malware) family can be habitually detected. We will continue to do this work in the future, because it is an efficient way to detect malware infections.

The network traffic generated when malware communicate with a command and control server is more prone to have consistent features. This is because the command and control channel the attacker build between the infected machine and the control server is steady. The C&C protocol the malware used for communicating with C&C server usually have consistent or regular content [23].

For example, while the malware TM1 in our TM (Targeted Malware) family launch a connections to command and control server through HTTP protocol, URL parameters is always consistent. TM1 has consistent URL parameters *"GET/1/login.php?u=YmFsY2s=&p=cGFzc3dvcmQyMDE1 HTTP1.1"*. Another targeted malware TM2 in our TM family has regular URL parameters *"GET/{6characters}.php?id={12 characters} HTTP1.1"*, the 12 characters string is the encrypted MAC address of the infected machine. When malware communicated with a command and control server over HTTP protocol, analyzing HTTP headers is a useful generic way to detect malware communications. We can extract network traffic features from URL parameters, Content-type, Content-Length and User-Agent in HTTP POST/GET request.

Some APT malware communicates with the command and control servers via HTTPS protocol. In this case, detecting consistent or regular URL parameters does not work, because the URL content is encrypted. There is another way to detect the HTTPS C&C traffic. Because the malware which communicate with the command and control servers via HTTPS protocol have consistent Secure Sockets Layer (SSL) certificates. Detecting consistent default values in SSL certificates is also an efficient way to detect malware infections [23].

B. ANOMALY-BASED DETECTION FEATURES

Anomaly-based intrusion detection is based on detecting anomalous behaviors occurs on the network. The signature-based method needs a database of known signatures. Anomaly-based intrusion detection needs to define anomalous or normal behaviors. We defined APT malware behaviors below including protocol anomaly, statistical anomaly, application anomaly:

Mismatch of Protocol and Port (Protocol Anomaly): For tunneling through the firewall of target network, the malware usually uses C&C communication protocols and ports that are allowed by the firewall. As observed, the most popular ports on which the malware communicate with the C&C servers are 80, 8080, 443, 8000, 1863, etc. The most popular protocols via which the malware communicate with the C&C servers are HTTP and HTTPs. The C&C communication protocol is designed and achieved by developers of the malware at the phase of coding, while the domains and ports are created when the attacker configure the malware for locating C&C servers before using. Since any ports can be configured by the remote access tool users, mismatch of protocol and port is happening sometimes. For example, HTTP protocol traffic occurs not on port 80 or 8080, or non-HTTP protocol traffic occurs on port 80. They are all very likely malicious traffic.

Encrypted Data Transpire on Uncommonly-Used Port (Protocol Anomaly): Not all the malware communicate with the C&C servers on commonly-used protocol ports. Some malware sometimes communicate with the C&C servers on ports which are seldom used by legitimate applications. And most APT malware C&C traffic data is encrypted to evade detections. So encrypted data transpire on uncommonly-used protocol port is also likely malicious traffic.

Mismatch of Uplink and Downlink Traffic (Statistical Anomaly): Normally, the downlink data traffic flow to host is larger than the uplink traffic to server. But the C&C communicating traffic is diametrically opposite. The data traffic that infected host upload to the control server is always larger than the data traffic received from the control server. For example, traffic of HTTP request much more than the HTTP response is very likely malicious traffic.

A number of Small Packets in Long TCP Connection (Statistical Anomaly): When the attacker send sets of command to the infected machine, commands such as file resource search command, file download command would require a lot of waiting time, coupled with the human thinking time, make the connection session a longer duration. And sets of commands are all small packets sent from C&C server to the infected host.

Heartbeat Packet Traffic (Application Anomaly): After the infected host client connected the command and control server, the server would send packets to the client, making sure the other end is on line. This kind of packet is called heartbeat packet. As heartbeat packets have similar size, we cluster all packets by packet size and check whether packets in the same cluster are sent periodically.

Malware Domain Traffic: As we observed in the experiment, the traffic of C&C server is smooth and small at most of the time, but has peak values when attacker were uploading data from infected hosts to the C&C server.

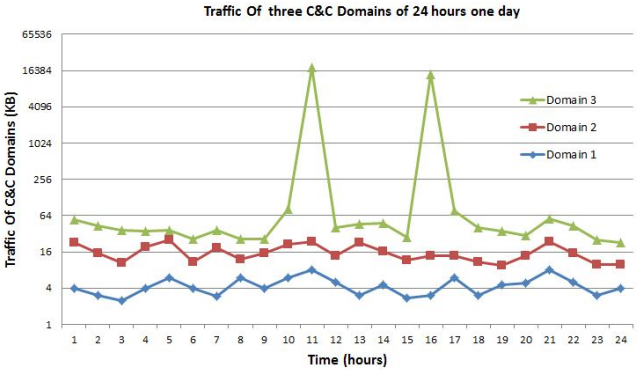


FIGURE 2. Traffic statistic of 3 Malware Domains .

Fig. 2 shows the traffic of 3 malware domains (C&C servers) in 24 hours one day. The C&C server of domain3 controlled 2 infected hosts, and the attacker stole 18.2MB data from the first host at 10:00-11:00AM, stole 14.1MB data from the other infected host at 15:00-16:00PM. The traffic of the other 2 domains is smooth and around 4KB and 16KB per hour all the day because of no data uploading.

According to analysis of the traffic, most hosts in the monitoring network do not know the malware domains and communicate with the C&C server except the infected hosts. The other reason is that heartbeat packet of infected host is tiny, and infected hosts were not always online all the time. Most malware carries the designation of stealing information for specific purpose. There are a number of files in the infected computer, our research from the malicious network traffic shows that attackers prefer to steal the office documents, such as doc, xls and pdf, from the infected computer. The attackers also prefer to upload compression tools such as WinRAR to pack a number of document files.

VI. BUILDING DETECTION MODELS

A. CONSTRUCTING THE TRAINING DATA SET

The training data set plays an important role in machine learning algorithm [25]. We aim to train a classifier that can identify domains used for crafted malware C&C servers, and to train a reputation function that can judge whether an IP address is infected or not by crafted malware.

For this purpose, approximate one thousand domains used for crafted malware C&C servers and one thousand benign domains were collected to construct training data set. These malicious domains in our training set we are talking about are C&C server domains for crafted malware not including malicious flux or DGA (Domain Generation Algorithm) domains. We collected malware C&C server domains from malwaredomains.com [26], VRT rule sets [18],

apt.securelist.com [24] etc. Since our experiment was performed at a large local institute network. Different from Notos [9] and Exposure [7], we took full advantage of the ''Virus Email Detector System'' which is deployed in this network, and extracted hundreds of malicious domains from hundreds of malware samples in virus email attachments. Sending virus emails to specific targets with attached documents that are packed with exploit code and trojan horse programmes has become one of the most important attack vectors in APT attack [27].

The training period of our system was the first four weeks. During this period of four weeks, ''the time-based behavior'' of malware C&C domains can be observed in a better way. During the first four weeks of experimenting at a large local research institute network with different values, we also labelled about 5 hundred malicious domains and more than 2 hundred infected machines inside the network, by manual analysis of the network connections to each suspicious C&C server domains and manual verification of every infected host inside the network. We are conservative when constructing the malicious domain list and infection host list. We apply a preliminary check before labeling a domain as being malicious, an IP address as being infected and using it in our training set. Every infection is confirmed by on-site examination and manual verification with the cooperation of the network administrator of the research institute. The one thousand benign domains in our training data set were collected from the Alexa top 1000 domains [28].

B. CLASSIFIER OF MALICIOUS DNS DETECTOR

The classifier of Malicious DNS Detector is using J48 decision tree algorithm. J48 decision tree is based on C4.5 algorithm and it has been proved to be efficient in classifying benign domains and malicious domains in EXPOSURE [7]. The J48 decision tree classifier is built in the training period. The condition of some attribute is being examined in every node. Every branch of the tree represents a result of the study.

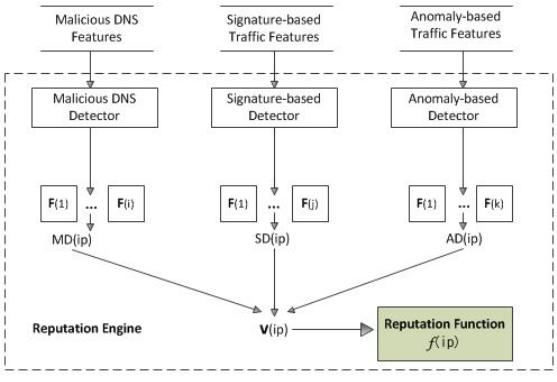


FIGURE 3. Reputation Engine to Assign a Reputation Score.

C. REPUTATION ENGINE

The reputation engine (see Fig. 3) of our system is responsible for detecting whether a host inside the network with IP address i has behaviors that are similar to a infected host or not. It computes a reputation score for an IP address. The reputation score is assigned between 0 and 1. Score 0 represents low reputation (it means malware infected) and score 1 represents high reputation (it means not infected). We implement this reputation function as a statistical classifier.

We make use of three modules which are malicious DNS detector module, signature-based detector module and anomaly-based detector module to compute three output vectors $MD(ip_i)$, $SD(ip_i)$ and $AD(ip_i)$, respectively. After computing the vectors $MD(ip_i)$, $SD(ip_i)$ and $AD(ip_i)$, these three feature vectors will be concatenated into one feature vector $V(ip_i)$. $V(ip_i)$ will be fed into the trained reputation function. The reputation function is responsible for computing a score $S = f(ip_i)$. S varies between 0 and 1. Result 0 represents low reputation, which means malware infected. Result 1 represents high reputation, which means not infected. The lower value represents the lower reputation. The reputation function is trained using data set $L = \{(V(ip_i)), y_i\}_{i=1..n}$. If ip_i is a confirmed infected host, $y_i = 0$, otherwise $y_i = 1$.

VII. EVALUATION

Our experiment was performed at a large local institute network for eight weeks. Note that, during experimental period of 2 months, the first four weeks of experimenting is training period and the last four weeks is for testing. This large local institute network is a type of network with high value information, it tends to be attacked by advanced persist threat attackers. The network has a professional traffic monitoring equipment at the edge to monitor large volumes of inbound and outbound traffic, including the DNS traffic and C&C server traffic. The large local institute network has more than 30,000 users, during experimental period of two months, and we monitored approximately 400 million DNS queries.

Without deploying any filters, it was not feasible to record and analyze this large volume of traffic. Hence, the volume of DNS traffic was reduced by using two filters. The first filter is the most popular domains in a white list. The Alexa Top 1000 Sites [28] and all the ones has the same well-known 2LD or 3LD domains were collected into the white list. By deploying the first filter, 20% of the monitoring traffic can be reduced. The second deploying filter is domains that were queried by more than 1000 hosts in the network we were observing. In the first 3 days of the experiment, we recorded and made a statistics of the domains by numbers of querying-host inside the network. Different from the malicious flux domains and DGA domains, the same C&C server domain for crafted malware in APT attack is seldom used for too many infected machines. By deploying the two filters, 85% of the traffic can be reduced. The second deploying filters made it feasible to record and analyze the traffic.

The professional traffic monitoring equipment can provide us the monitoring traffic by rules of ''Source IP address'', ''Destination IP address'', ''Source Port'' or ''Destination Port''. During experimental period, we can submit the

suspicious C&C server IP as the rule of ''Source IP address'' and ''Destination IP address;'' to the monitoring equipment at any time. Therefore, the C&C server traffic our system should record and analyze is small.

A. EVALUATION OF THE CLASSIFIER AND THE REPUTATION FUNCTION

In order to evaluate the true positive rates and false positive rates of our ''Malicious DNS Classifier'' and ''Reputation function'', we did the evaluating experiment in our training data set using 10-fold cross validation and 66% percentage split. 66% percentage split means 66% of the training data set is used for training and the rest 34% is used for checking. Table 2 shows the results of this evaluating experiment. The 10-fold cross validation and 66% percentage split evaluating experiment shows that the true positive rates of our malicious DNS classifier and reputation function are about 96%, the false positive rates are about 1.5%.

The final purpose of our system is to detect unknown crafted malware C&C domains and crafted malware infections. The evaluation must show that it can detect unknown crafted malware C&C domains and crafted malware infections that are not in the training data set. For this purpose, we used the test data set of the last four weeks of experimenting.

Because the experiment was performed at a large local institute network, it is challenging to determine the real true positive rate with the real-life data set. To build the ground truth, during the last four weeks we collected 900 domains from VRT rule sets [18], malware samples of email attachment from ''Virus Email Detector'' deployed in this network, malwaredomainlist.com [26]. All the collected domains are not ever used in the training data set before. 426 of the 900 collected domains were requested by infected machines in the network during the last four weeks. And the rest 474 domains were not queried. In the experiment, out of the 426 domains, 408 domains were detected as malware C&C domains by IDnS. The true positive rate was 95.8%. By manual analysis of the network traffic of C&C servers pointed by known malicious domains and on-site examination of suspicious infected hosts, 197 machines were identified as being infected. Out of these, 188 machines were detected as being infected by IDnS. The true positive rate of the reputation function is about 95.4%.

During the last four weeks' experimenting period, 459 domains and 227 machines were detected as being malicious and infected by IDnS in total. As explained above, we confirmed every malicious domain and infection by manual network traffic analysis, on-site examination and verification. The results show that the false positive rate is about 2.9% for the malware C&C domains that were identified, about 3.4% for the infected machines that were identified.

B. EVALUATION OF THE SUSTAINABILITY OF THE SYSTEM

We made a statistics of the traffic of 200 C&C servers we have identified, and the result of one day is shown in Table 3. We also found that the traffic of a single C&C server may change a lot everyday, but for a large number of C&C servers, the total traffic volume and the proportion of traffic did not change much. It illustrates that C&C server IP has a small volume of traffic per day under normal circumstances, and the network traffic analyzer of the system only needs to analyze very less traffic.

Without knowing which IP address is suspicious C&C server, most current network-based intrusion detection systems require monitoring and analyzing all the traffic. To signature-based IDS, it is hard to handle large volumes of traffic typical of large enterprise and ISP networks in real time. To anomaly-based IDS, it is not feasible to record and analyze this large volume of traffic. The security approach in this paper is able to substantially reduce the volume of network traffic that needs to be recorded and analyzed. It can improve the sustainability of the system.

VIII. COMPARISON WITH PREVIOUS WORK

A. EXPOSURE: FINDING MALICIOUS DOMAINS USING PASSIVE DNS ANALYSIS

EXPOSURE [7] is a system for detecting domains that are involved in malicious activity using passive DNS analysis. The authors also presented 15 behavioral features that EXPOSURE uses to identify the malicious domains. For detecting botnets, the explanations of the features they extracted shows that EXPOSURE focuses on detecting botnets using malicious flux and domain generation algorithm (DGA).

In ''Time-based features set'', the feature ''daily similarity'' is presented to detect ''an increase or decrease of the request count at the same intervals everyday''. This feature is identified based on that botnets often use malicious flux service [16]. The feature ''short life'' is extracted from domains because they believe domain generation algorithm (DGA) may be used by each bot. The DGA domains are used only for a short duration [7].

In ''TTL value-based feature set'', the features ''Number of distinct TTL values'' and ''Number of TTL change'' were proposed based on tracking the Conficker worms for a week. The Conficker is an example malware that make use of DGA. The features they proposed can not be used for detecting crafted remote access tools (RAT) in APT attacks, because APT attackers seldom use domain flux technique or DGA.

In ''Domain name-based feature set'', the features ''% of numerical characters'', ''% of the length of the LMS (Longest Meaningful Substring)'' that EXPOSURE presented are not fit for detecting APT malware either. The domains used in APT attacks, no matter dynamic domain names (DDNS) or not, were all manually registered by the attackers, they do not have the same features as EXPOSURE defined.

B. BOTNET DETECTION BY MONITORING GROUP ACTIVITIES IN DNS TRAFFIC

[17] proposed a new mechanism to detect botnets by monitoring group activity DNS traffic. They focus on detecting

IEEE Access
The journal for rapid open access publishing

TABLE 2. Evaluation of the classifier and the reputation function.

	TP of DNS classifier	FP of DNS classifier	TP of Reputation function	FP of Reputation function
10-folds Cross-Validation	96.5%	1.4%	95.8%	1.2%
66% Percentage Split	96.3%	1.7%	95.7%	1.3%

TABLE 3. Traffic of C&C servers one day.

Traffic of C&C Server (MB)	% of Servers
<1M	55%
1M-10M	24%
10M-100M	10%
100M-300M	8%
>300M	3%

a group of bots, referred to as a botnet, which are remotely controlled by a C&C server and can be used for sending spam mails, launching DDoS attacks etc. The authors defined feature "group activity" to detect botnet. The feature is identified based on the judgement that the number of bots which queried botnet domain is fixed in general. The group activity is formed by simultaneous DNS queries sent by a number of distributed bots. Most legitimate domain names are queried continuously but not simultaneously.

This anomaly-based detection mechanism can detect botnet which is unknown or new to us. But this approach have intrinsic limits, it can only detect botnet consist of large numbers of bots. To reduce the risk of being detected, advanced attackers seldom use the same C&C server and domain to remotely control large numbers of compromised end-user machines.

IX. DISCUSSION

This section discusses the advantages and limits of IDnS. Malicious DNS analysis is first performed to find out suspicious IP addresses of command and control servers in our approach. Only by network traffic features analysis, IDS sometimes can not accurately judge whether a host is infected or not. But combined with malicious DNS traffic analysis, IDS can significantly increase detecting accuracy. Malicious traffic features occurs in the traffic to a suspicious malicious IP do have very low reputation of normal traffic. Another advantage of this approach is that it can improve the sustainability of the system. For a large and high-speed network, it is too hard to record and post-process all the inbound and outbound traffic. This method can greatly reduce the volume of network traffic which needs to be recorded and analyzed.

The main limitation is the fact IDnS is not good at detecting malware infections that do not rely on domains, such as the trojan use the IP address directly to locate the command and control server. After a long detecting period, we collect a lot of IP addresses of command and control servers. This can also help to detect parts of malware infections that do not rely on domains by analyzing the traffic of C&C servers. Besides, the administrator of this system can also provide the IP addresses they want to concentrate on for the traffic analyzer of the system.

X. CONCLUSION

In this paper, we propose a novel system IDnS placed at the network egress points to detect malware infections inside the network combined with DNS traffic analysis. We extracted new features and built a reputation engine based on big data, which includes approximately 400 million DNS queries. The experimental results show that our security approach is good at detecting APT malware infections and is feasible for improving the sustainability of the system. The system processes advantages of high efficiency and accuracy. We believe that IDnS is a useful intrusion system that can help to fight against cyber-crime especially theft of information from infected host.

XI. ACKNOWLEDGMENT

This work has been supported in part by NSFC Project (61170292, 61472212), National Science and Technology Major Project (2015ZX03003004), 973 Project of China (2012CB315803), 863 Project of China (2013AA013302, 2015AA015601), EU MARIE CURIE ACTIONS EVANS (PIRSESGA2013610524), and multidisciplinary fund of Tsinghua National Laboratory for Information Science and Technology.

REFERENCES

[1] F. C. Freiling, T. Holz, and G. Wicherski, "Botnet tracking: Exploring a root-cause methodology to prevent distributed denial-of-service attacks," *Lect. Notes Comput. Sci.*, vol. 10, no. 2, pp. 319–335, 2005.

[2] A. Karasaridis, B. Rexroad, and D. Hoeflin, "Wide-scale botnet detection and characterization," in *Proc. 1st Conf. 1st Workshop Hot Topics Understand. Botnets*, 2007, p. 7.

[3] J. Jung, M. Konte, and N. Feamster, "Dynamics of online scam hosting infrastructure," in *Proc. 10th Int. Conf. Passive Active Netw. Meas.*, 2009, pp. 219–228.

[4] H. Porras, H. Saïdi, and V. Yegneswaran, "A foray into Conficker's logic and rendezvous points," in *Proc. USENIX Conf. Large-Scale Exploits Emergent Threats, Botnets, Spyware, Worms, More*, 2009, p. 7.

[5] S. Yadav, A. K. K. Reddy, A. L. N. Reddy, and S. Ranjan, "Detecting algorithmically generated malicious domain names," in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 48–61.

[6] J. Wolf. (2008). *Technical Details of Srizbi's Domain Generation Algorithm*. [Online]. Available: http://tinyurl.com/6mdasc

[7] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "EXPOSURE: Finding malicious domains using passive DNS analysis," in *Proc. NDSS*, 2011.

[8] E. Stalmans and B. Irwin, "A framework for DNS based detection and mitigation of malware infections on a network," in *Proc. Inf. Secur. South Africa (ISSA)*, Aug. 2011, pp. 1–8.

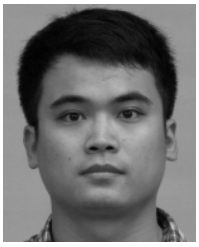
[9] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster, "Building a dynamic reputation system for DNS," in *Proc. 19th USENIX Secur. Symp.*, 2010, pp. 273–290.

[10] LASTLINE. (2015). *Using Passive DNS Analysis to Automatically Detect Malicious Domains*. [Online]. Available: https://www.lastline.com/papers/dns.pdf

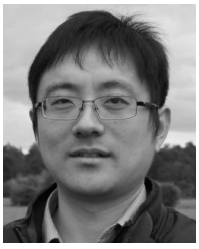
[11] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, II, and D. Dagon, "Detecting malware domains at the upper DNS hierarchy," in *Proc. USENIX Secur. Symp.*, 2011, p. 27.

[12] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling, "Measuring and detecting fast-flux service networks," in *Proc. NDSS*, 2008.

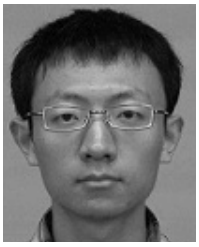
[13] N. Brownlee, K. Claffy, and E. Nemeth, "DNS measurements at a root server," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, vol. 3. 2001, pp. 1672–1676.
[14] S. Castro, D. Wessels, M. Fomenkov, and K. Claffy, "A day at the root of the Internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 5, pp. 41–46, 2008.
[15] H. Gao *et al.*, "An empirical reexamination of global DNS behavior," in *Proc. ACM SIGCOMM Conf. SIGCOMM*, 2013, pp. 267–278.
[16] R. Perdisci, I. Corona, D. Dagon, and W. Lee, "Detecting malicious flux service networks through passive analysis of recursive DNS traces," in *Proc. Annu. Comput. Secur. Appl. Conf. (ACSAC)*, Dec. 2009, pp. 311–320.
[17] H. Choi, H. Lee, H. Kim, and H. Lee, "Botnet detection by monitoring group activities in DNS traffic," in *Proc. 7th IEEE Int. Conf. Comput. Inf. Technol. (CIT)*, Oct. 2007, pp. 715–720.
[18] Sourcefire. (2015). *Snort Network Intrusion Detection System Web Site*. [Online]. Available: https://www.snort.org/
[19] M. Roesch, "Snort—Lightweight intrusion detection for networks," in *Proc. 13th LISA*, 1999, pp. 229–238.
[20] V. Kumar and D. O. P. Sangwan, "Signature based intrusion detection system using SNORT," *Int. J. Comput. Appl. Inf. Technol.*, vol. 1, no. 3, pp. 35–41, 2012.
[21] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur.*, vol. 28, nos. 1–2, pp. 18–28, 2009.
[22] F. Gong, "Deciphering detection techniques: Part II anomaly-based intrusion detection," McAfee Security, White Paper, 2003.
[23] N. Villeneuve and J. Bennett. (2012). Detecting apt activity with network traffic analysis. Trend Micro Inc. [Online]. Available: http://www.trendmicro.com/cloud-content/us/pdfs/securityintelligence/white-papers/wp-detecting-apt-activity-with-network-traffic-analysis.pdf, accessed Oct. 31, 2013.
[24] Kaspersky Lab. (2015). *Targeted Cyberattacks*. [Online]. Available: https://apt.securelist.com/
[25] K. Koutroumbas and S. Theodoridis, "Pattern recognition," in *Encyclopedia of Information Systems*. 2003, pp. 459–479.
[26] (2015). *Malware Domains List*. [Online]. Available: http://www.malwaredomains.com//
[27] V. Prenosil and I. Ghafir, "Advanced persistent threat attack detection: An overview," Tech. Rep., 2014.
[28] Alexa Web Information Company. (2015). [Online]. Available: http://www.alexa.com/topsites



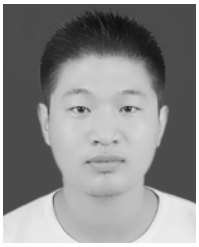
GUODONG ZHAO received the B.S. degree in computer science from Information Engineering University, China, in 2007. He is currently pursuing the master's degree in computer science and technology with Tsinghua University, Beijing, China. His research interests include network security and software-defined networking.



KE XU (M'02–SM'09) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2001. He is currently a Full Professor with Tsinghua University. He is also a Visiting Professor with the University of Essex. He has authored or co-authored over 100 technical papers. He holds 20 patents in the study of next-generation Internet, P2P systems, Internet of Things, network virtualization, and optimization. He is a member of the Association for Computing Machinery. He has been a Guest Editor of several special issues of the IEEE and Springer journals.



LEI XU received the B.S. degree in computer science from the Beijing Institute of Technology, China, in 2006. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University, under the supervision of Prof. Y. Jiang. His research interests include datacenter networking and software-defined networking.



BO WU received the B.S. degree in computer science from Shandong University, China, in 2014. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include next generation Internet and network security.