



Visualizing and characterizing DNS lookup behaviors via log-mining



Qingnan Lai^a, Changling Zhou^b, Hao Ma^{b,*}, Zhen Wu^b, Shiyang Chen^a

^a School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

^b Computer Center, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:

Received 1 May 2014

Received in revised form

2 September 2014

Accepted 11 September 2014

Available online 16 May 2015

Keywords:

DNS lookup

Visualization

User behavior

ABSTRACT

The Domain Name System (DNS) is a critical Internet service, which translates easily memorized domain names to numerical IP addresses for locating computer resources and services. In this paper, we try to explore the behaviors of DNS lookup by mining DNS logs from three primary DNS servers in a large university campus network in China. Our dataset is made up of two parts, namely DNS query logs and messages received or send by DNS servers. Firstly, through analyzing these DNS query logs, we are able to understand the overall trend of users' surfing. For dealing with huge DNS dataset, we introduce an algorithm we call *DNSReduce*, which can be used to dig out top 10 client IP addresses and top 10 destination domain names efficiently. Moreover, we make comparative analysis of lookup behavior between wired and wireless users. Secondly, with messages received or send by DNS servers we can find these DNS servers' behaviors, i.e., TTLs, equivalent answers and are able to accurately identify domain names with dynamic IP addresses. We provide different and specific visualization techniques for presenting these analysis results and show these techniques are very useful for understanding user behaviors, analyzing security events and characterizing overall tendency in campus network management.

© 2015 Published by Elsevier B.V.

1. Introduction

As an infrastructure of the Internet, the Domain Name System (DNS) service is critical for normal functioning of almost all Internet applications, which translates easily memorized domain names to numerical IP addresses for locating computer resources and services. Although the Internet Protocol (IP) does not need DNS for operation, users need to distinguish machines by their names, so the DNS protocol is needed to resolve names to IP addresses (and vice versa) [1]. DNS's highly redundant design and aggressive caching makes DNS very scalable and fault-tolerant [2].

Data visualization analysis helps us to perceive patterns, trends, structures, and exceptions in even the most complex data sources, and communicating with visuals allows for comprehension of large amount of data. The visualization in DNS traffic allows the network administrators to identify concepts and relationships that are hardly explored and exploited in previous work [3].

In this paper, we use visualization techniques to present meaningful information that we extract from DNS logs. DNS lookup behaviors such as query tendency, top 10 client IP addresses and

destination domain names are valuable and helpful in managing and optimizing campus network. For example, because of the widely deployed Content Distribution Network (CDNs) which make use of DNS to direct users to exactly the content replica, so it will be very useful for network operators to understand IP dynamics of heavy requested domain names in improving performance in a multi-homed campus network. These lookup characteristics could be mined from DNS datasets, an algorithm termed *DNSReduce* is introduced for dealing with these huge data. We use Passive DNS tool to capture live inter-server DNS messages to extract requested domain names and resolved IP addresses [4] for identifying domain names having dynamic IP addresses and tcpdump tool to capture DNS response data to analyze DNS servers' behaviors.

The remainder of this paper is organized as follows. In Section 2, we discuss prior studies that are related to our own. In Section 3, we provide an overview for our dataset and methods. The detailed experiments are described in Section 4. In Section 5, we present our results and discussions. Section 6 is the conclusion of this paper.

2. Related work

Methods for analyzing the characteristics of DNS behavior have been studied in a large amount of prior work. Bojan et al. collected DNS responses at the University of Auckland Internet gateway and

* Corresponding author. Tel.: +86 10 62751986.

E-mail addresses: laqn@pku.edu.cn (Q. Lai), zcl@pku.edu.cn (C. Zhou), mah@pku.edu.cn (H. Ma), wuzhen@pku.edu.cn (Z. Wu), yssy1026@163.com (S. Chen).

analyzed them to detect unusual behavior, they observed that current attempts to reduce spam had greatly increased the number of A records being resolved and locality of DNS requests diminished because of domains advertised in spam [1]. Ren et al. presented Flying Term, a new perceptually motivated visual metaphor for visualizing the dynamic nature of DNS queries [5]. Hadi et al. offered a comprehensive review of network security visualization and provided a taxonomy in the form of five use-case classes encompassing nearly all recent works in this area [3].

Schonewille and Helmond's research was a first glance at the usability of DNS traffic and logs for detection of this malicious network activity. Detection of bots was possible by DNS information gathered from the network by placing counters and triggers on specific events in the data analysis [6]. David and Paul considered three classes of DNS traffic: canonical, overload and unwanted, and showed preliminary results on how DNS analysis could be coupled with general network traffic monitoring to provide a useful perspective for network management and operations [7].

Kirkpatrick et al. introduced a method for clustering misconfigured DNS sources [2]. Using machine learning methods, they analyzed 24 h of DNS requests that were collected on the A-root DNS server. Their research provided preliminary results that were validated via discussion with DNS system operators. Shan et al. proposed an interactive visual analysis system for the DNS log files to intuitively detect the anomalies in DNS query logs [8]. Albrecht-Buehler used motion to visualize trends among text–theme relationships and allowed user interaction of the temporal controls and theme relations [9]. Brandes et al. used animation to illustrate the dynamics of international political and military conflicts [10]. In Pieter's research, a visual analytics approach was used on a large set of DNS packet captures to gain insight into ways that authoritative name servers were abused for denial of service attacks [11]. Several tools were developed to identify patterns in DNS queries and responses. Yu presented a visualization analysis tool for detecting, analyzing and responding to the Distributed Denial of Service attack termed the Domain Name Service (DNS) amplification attack [12]. In Born's study both quantitative analysis and visual aids were provided that allowed the user to make determinations about the legitimacy of the DNS traffic [13].

Jung et al. proposed the Localized Big Data Analysis (LBDA) with which data mining techniques were applied to the DNS log [14]. Ruan et al. proposed a novel periodic trend mining method as well as a periodic trend pattern based traffic prediction method [15]. David Dagon described a tutorial on large-scale DNS data analysis, the tutorial was a pragmatic course on data collection, analysis and techniques [16]. Callahan et al. made an initial empirical understanding of a broad range of characteristics of modern DNS behavior, such as TTLs, equivalent answers and proximity [17].

In contrast to prior studies, our work focuses on DNS query characteristics and visualization in a university campus network, our analyses consider both query logs and DNS messages to reveal a deeper view of DNS lookup behavior, and our methods could be used and verified in similar campus network environment.

3. Approach architecture

The architecture of our proposed approach is shown in Fig. 1, our data source is from the DNS system, which is composed of three servers in campus network of Peking University in China. Let us use DNS A, DNS B and DNS C to denote these three servers. DNS A and DNS B are allocated to deal with wired user queries, and DNS C serves for the rest of wired users and all wireless users. So it is possible to analyze wired and wireless users' lookup behaviors separately. DNS servers provide two kinds of data, i.e., query logs and messages received or sent by DNS servers. We stored these data into a MongoDB database, then used different data mining approaches on the dataset to achieve the following two goals:

- (1) Describing the trend of queries and dig out top 10 client IP addresses and destination domain names.
- (2) Identifying domain names which have dynamic IP addresses.

Lastly, we visualized our results. The detailed algorithm is provided in Section 4.

3.1. DNS query logs

We collected DNS query logs within a period of three days from each server as dataset for our experiments. Each log message contained a timestamp, client IP address and port, query class and type. For example

```
13-Apr-2014 09:25:08.955
client 162.105.51.200#64533
query: news.eoe.cn IN A +
```

Table 1

DNS query distribution: other query types are not shown.

Servers	Queries/Day	A (%)	AAAA (%)
DNS A	122,158,559	71.38	19.60
DNS B	141,399,824	74.00	20.70
DNS C	85,527,359	66.15	32.16

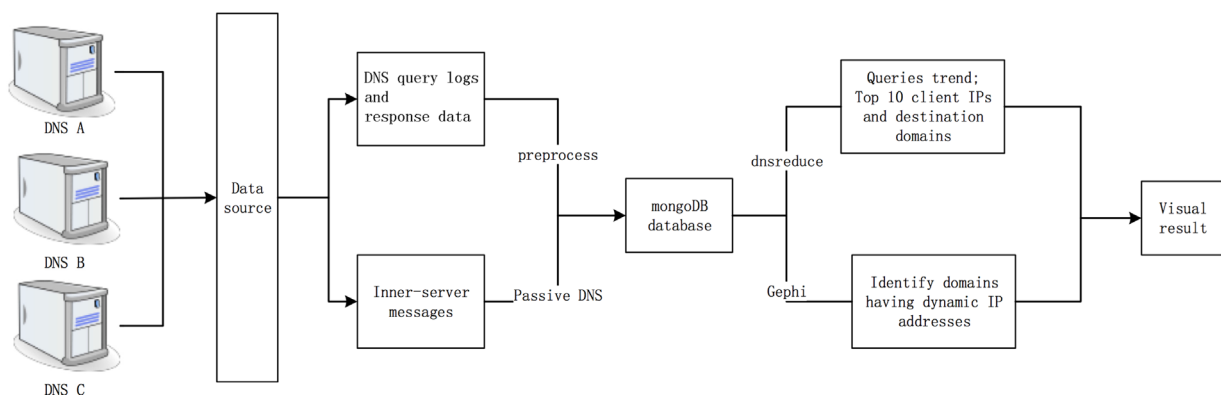


Fig. 1. Approach architecture.

We stored those logs into a MongoDB database. In order to improve retrieval efficiency, we preprocessed data to generate information such as queries every 1 min, 2 min and so on, then stored them into a collection of our database for fast access.

Table 1 shows the DNS query types' percentage of total DNS traffic observed from April 12, 2014 through April 13, 2014. In the table "A" indicates a type of IPv4 address, "AAAA" means IPv6 address. Other common query types include Mail eXchanger (MX), Pointer (PTR) records used to translate IP addresses to names, Service Location (SRV) records used for automatic discovery of services, and Text (TXT) records used for other various purposes [7]. The number of DNS queries reached 100 million per day.

3.2. Messages received or send by DNS servers

We used Passive DNS software to capture inter-server DNS messages to identify domain names with dynamic IP addresses. Also we stored these messages into a MongoDB database after preprocessing. Each message contained a timestamp, domain name and resolved IP address:

```
2014-4-4 12:00:02
www.vose.me.
128.134.244.74
```

At the same time, we deployed tcpdump tool to capture the flow of DNS servers to get DNS responses. Each response contained a timestamp, DNS server's IP address, client IP addresses, query, query type, and look up result, for example

```
09:54:18.001719
162.105.129.27.53 to 222.29.145.12.49340
q: A? fundin.douban.com.
2/2/0 fundin.douban.com. [3m31s] A 211.147.4.32, fundin.douban.com. [3m31s] A 211.147.4.31 ns: douban.com. [1m14s] NS ns1.douban.com., douban.com. [1m14s] NS ns2.douban.com.
```

In the above example, "q" stands for query, "2/2/0" indicates "2" answer resource records, "2" authority resource records and "0" additional resource records, "[3m31]" is the TTL of the record stored in DNS servers.

4. Analysis approaches

In this section, we elaborated how to achieve our goals.

4.1. DNS lookup behaviors in query logs

To implement our analysis of DNS lookup behaviors in query logs, we retrieved queries as an input to plot a time series. Based on this, we could watch one-day or several-day queries which varies from time. Moreover, we tried to dig out top 10 client IP addresses and destination domain names. The number of queries per day was about 100 million, which made it infeasible to finish data processing job at once. Nonetheless, we had some prior knowledge about IP address range, i.e., in our campus network, a class B IP address block and 160 class C IP address block were allocated to wired connections while another class B IP address block was allocated to wireless connections. Hence, we proposed to solve this problem by separating the logs according to IP address blocks. We called this method as *DNSReduce*. Fig. 2 is the flow diagram for *DNSReduce*, below is the processing steps:

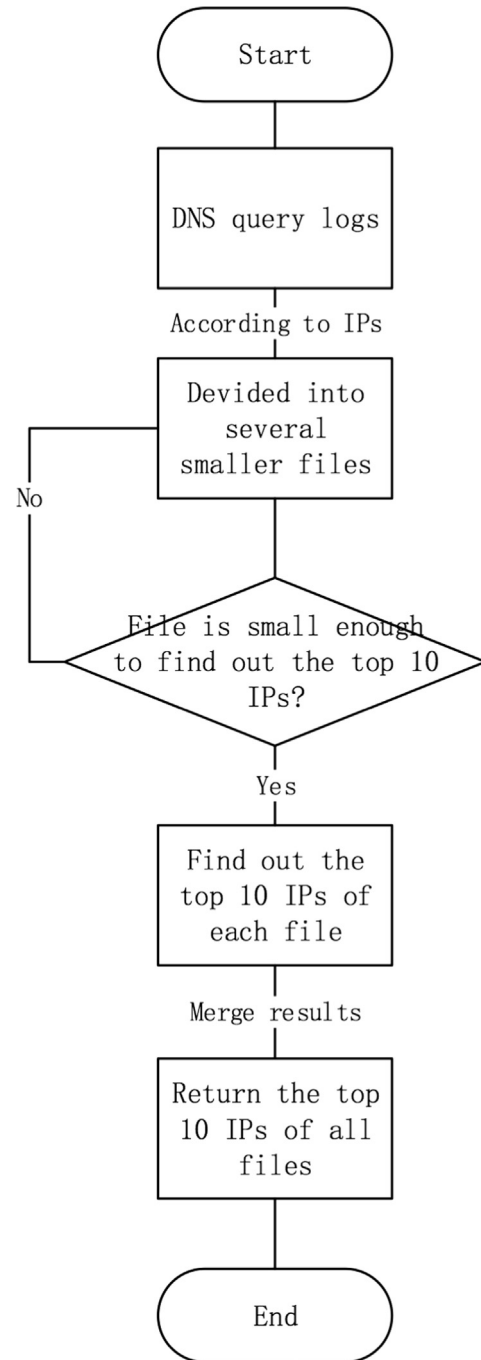


Fig. 2. *DNSReduce* algorithm.

Step 1: Collect DNS query logs as input.

Step 2: Using IP addresses ranges as a criterion to partition the logs into several smaller files, because we cannot finish this job at once.

Step 3: If one file is small enough to find out top 10 client IP addresses using Hash algorithm, then store it as an intermediate results. Else, we should partition it again until finding out top 10 client IP addresses.

Step 4: Merge all the intermediate results and return top 10 client IP addresses.

Fig. 3 is an example of *DNSReduce*'s procedure. Similarly, using top domain names, such as "com", "net", "com.cn" as a criterion to

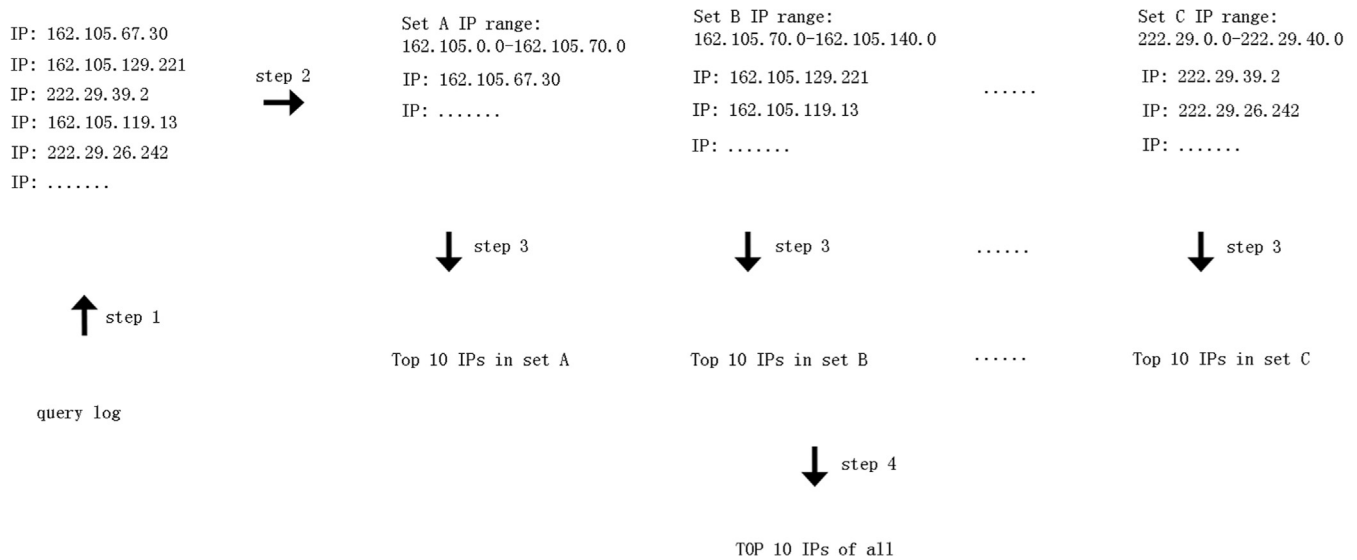


Fig. 3. Example of DNSReduce procedure.

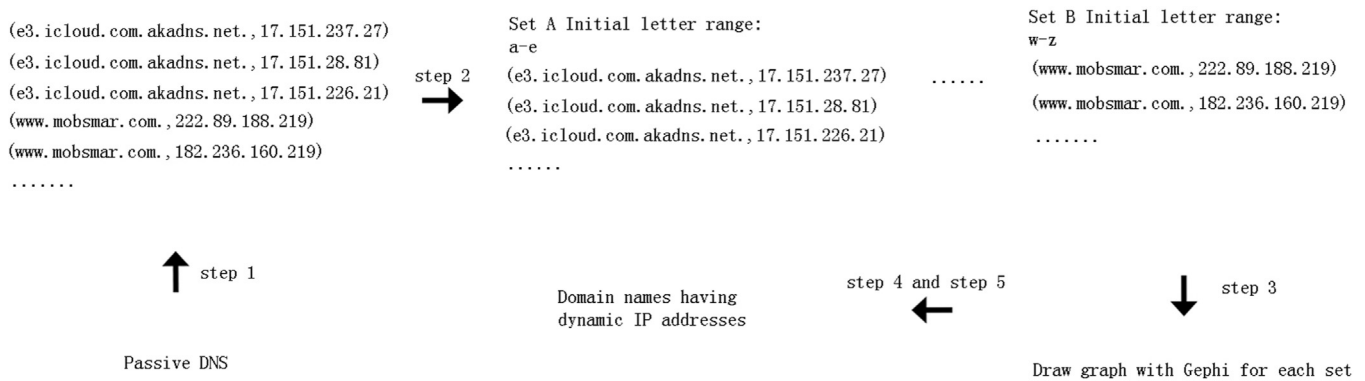


Fig. 4. Example of identifying domain names having dynamic IP addresses.

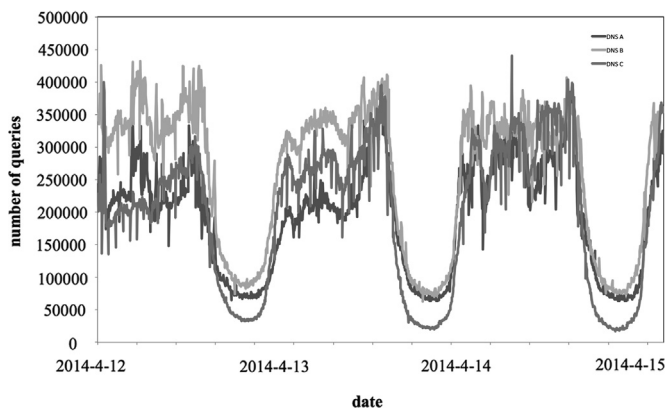


Fig. 5. Query trend of three DNS servers from 2014.4.12 to 2014.4.15.

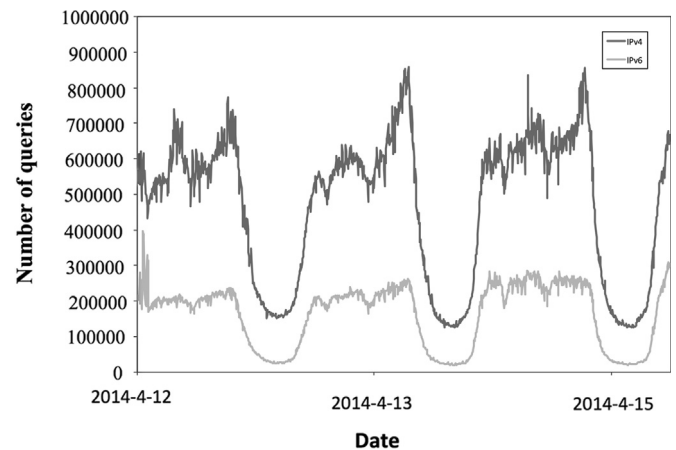


Fig. 6. Three DNS servers total IPv4 and IPv6 query trend from 2014.4.12 to 2014.4.15.

partition, we digged out top 10 destination domain names in the same way.

4.2. Identifying domain names having dynamic IP addresses

The input data set was collected from Passive DNS. In this section we used a visualization tool named Gephi [18], an open-source software for visualizing and analyzing large network

graphs, to identify domain names with dynamic IP addresses in the following steps:

- Step 1 Extract a relation from one message, like domain name and resolved IP address, then store it into a new file.
- Step 2 Partition the file into several smaller files by the domain name's initial letter, because we cannot generate the graph once.

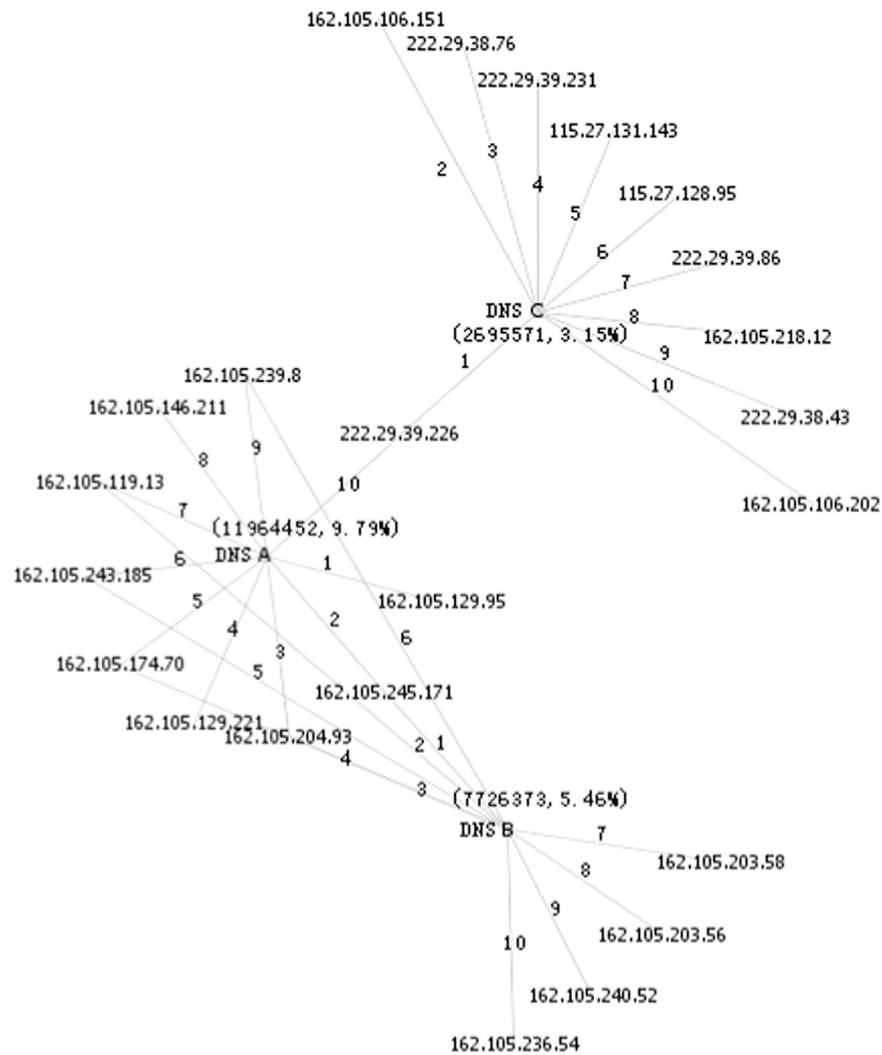


Fig. 7. Top 10 client IP addresses of three DNS servers.

Table 2

DNS A top 10 client IP addresses: m.n. and r.s. are the IP address prefix, which belongs to the IP address range in this campus network.

No.	IP address	Queries/Day
1	m.n.129.95	2,877,991
2	m.n.245.171	1,584,458
3	m.n.204.93	1,297,298
4	m.n.129.221	1,231,298
5	m.n.174.70	1,040,615
6	m.n.243.185	920,525
7	m.n.119.13	890,750
8	m.n.146.211	844,316
9	m.n.239.8	672,147
10	r.s.39.226	605,054

Table 3

DNS A top 10 destination domain names.

No.	IP address	Queries/Day
1	baidu.com	6,234,065
2	in-addr.arpa	5,054,671
3	xunlei.com	4,392,942
4	qq.com	3,844,502
5	sina.com.cn	2,446,784
6	taobao.com	2,297,344
7	weibo.com	2,134,947
8	pku.edu.cn	2,114,030
9	h3q.com	2,055,242
10	google.com	1,779,111

Step 3 Use each file as input, draw a graph with Gephi.

Step 4 Filter the nodes with only one edge, which means the IP address corresponding to the domain name is not changing in that period of time.

Step 5 Summarizing all the results, then we get a graph that shows the domains having dynamic IP addresses.

As an example, we analyzed the variability for a domain name's resolved IP address. We put all IP addresses of one domain name in order in a chart to describe its variability. As we know, the first three bits in the most significant octet of an IP address are defined

as the class of the address, class A is 0.0.0.0-127.255.255.255, class B is 128.0.0.0-191.255.255.255, class C is 192.0.0.0-223.255.255.255. Three classes are marked in the chart as a reference. Fig. 4 shows an example of identifying domain names with dynamic IP addresses.

5. Results and discussion

In this section we discuss the results of our experiments. Section 5.1 shows DNS lookup behaviors in query logs, including

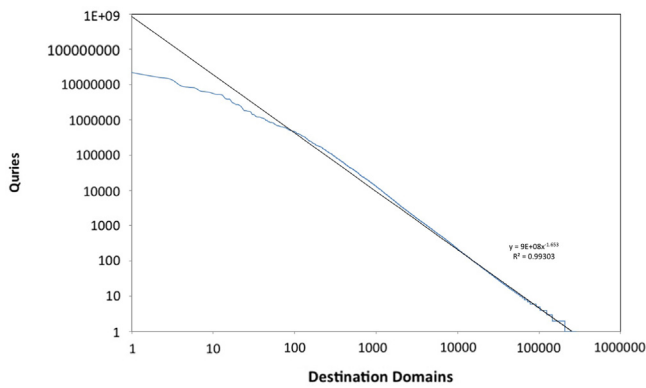


Fig. 8. Distribution of destination domain names in a day.

an overall trend and anomaly behaviors, Section 5.2 shows the dynamics of domain names with dynamic IP addresses.

5.1. DNS lookup behaviors in query logs

5.1.1. Overall tendency

Fig. 5 shows the query trend of three DNS servers during three days, Fig. 6 shows total IPv4 and IPv6 query trend from three DNS servers. The number of IPv4 queries is about three times of IPv6 queries. During daytime, all of three DNS servers' queries maintain at a high level. At mid-night the queries begin to fall down and then climb up at morning. The results are corresponding to users' habit of accessing Internet, i.e., students and faculty become active at 8 am and their activities reach a peak at 11 pm, and at mid-night students and faculty take a rest gradually. With this trend we could split the traffic into different DNS servers to reduce request load and shorten response time. For example, the number of the queries directed to DNS A is smaller than DNS B, we suggested campus network operators to modify some DHCP servers' configuration using DNS B as the first choice of DNS server for balancing network traffic.

Next, we used the algorithm provided in Section 4 to dig out top 10 client IP addresses and destination domain names according to the queries from 2014.4.12 11:00 am through 2014.4.13 11:00 am.

In Tables 2 and 3, we showed top 10 client IP addresses and destination domain names in DNS A and their queries, respectively. On the basis of these data, we presented a plain visualization design in Figs. 7 and 9 separately. There are some illustrations about the two figures. Firstly, three nodes in the center of Figs. 7 and 9 represent three DNS servers separately. Secondly, the nodes around DNS servers represent IP addresses or domain names, every address or name has at least one edge with one DNS server. Thirdly, the number on the edge represents the corresponding rank in that DNS server; the number besides DNS server means the total queries of top 10 IP addresses and the percentage of the whole day queries. If one address or name has more than one edge, which means this address or name ranks in top 10 in more than one DNS server. There are 7 IP addresses having two edges, and 6 domain names having 3 edges, 3 domain names having 2 edges. Top 10 client IP addresses request a great number of queries, we can track them to find if there exist anomalies. Top 10 destination domain names traffic is 26.4%, 31.5% and 41.96% by the whole day in DNS A, DNS B and DNS C, respectively, this imply that it is possible to provide cache service for these domain names to improve response speed. The query distribution of destination domain names is shown in Fig. 8, which follows Zipf's law, a well-known law describing the frequency distribution of different items

in an itemset [19]. So we should not only pay attention to top 10 destination domain names, but also notice the long tail.

On the other hand, we also focused on DNS lookup behaviors of wireless users. As the IP addresses of wireless users are in a specific range, which makes it possible to deal with the queries from wireless users solely. Top 10 wireless client IP addresses and destination

Table 4

Top 10 wireless client IP addresses: p.q. is IP address prefix, which belongs to the IP address range in this campus network.

No.	IP address	Queries/Day
1	p.q.131.143	238,412
2	p.q.128.95	227,478
3	p.q.137.107	113,983
4	p.q.129.213	95,923
5	p.q.106.27	84,350
6	p.q.34.164	74,186
7	p.q.79.136	73,413
8	p.q.141.26	65,044
9	p.q.130.39	62,229
10	p.q.48.59	52,606

Table 5

Top 10 wireless destination domain names.

No.	IP address	Queries/Day
1	baidu.com	3,789,755
2	qq.com	2,661,189
3	akadns.net	2,387,316
4	apple.com	1,790,093
5	taobao.com	1,186,935
6	in-addr.arpa	1,089,644
7	google.com	981,895
8	weibo.com	798,301
9	sina.com.cn	785,648
10	360.cn	755,655

Table 6

Top 10 wired users subdomain names of "qq.com" in DNS B.

No.	Domain name	Queries/Day
1	www.qq.com	515,396
2	ptlogin2.qq.com	128,232
3	monitor.uu.qq.com	113,404
4	3gimg.qq.com	94,696
5	dns.weixin.qq.com	92,946
6	pingma.qq.com	88,844
7	pingfore.qq.com	84,746
8	imgcache.qq.com	83,044
9	isdsped.qq.com	78,897
10	v.gdt.qq.com	75,177

Table 7

Wireless users' top 10 subdomain names of "qq.com" in DNS A.

No.	Domain name	Queries/Day
1	monitor.uu.qq.com	93,873
2	3gimg.qq.com	91,402
3	new-event.sng.qq.com	75,014
4	dns.weixin.qq.com	69,722
5	pingma.qq.com	59,389
6	short.weixin.qq.com	49,634
7	msfwifi.3g.qq.com	47,323
8	www.qq.com	45,067
9	long.weixin.qq.com	41,101
10	imgcache.qq.com	39,451

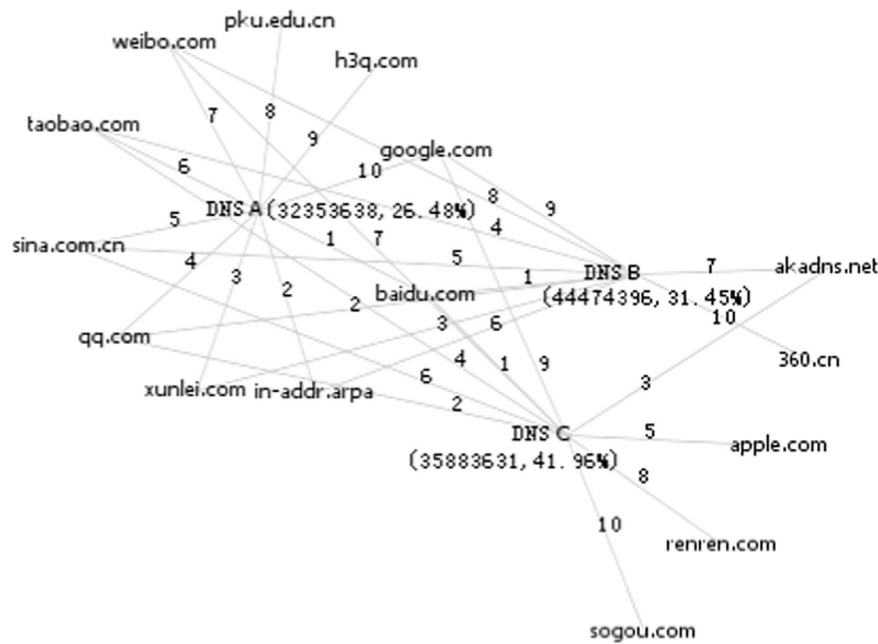


Fig. 9. Top 10 destination domain names of three DNS servers.

domain names are shown in Tables 4 and 5, respectively. The queries of wireless IP addresses are obviously less than the wired ones, which means mobile devices' network traffic is still less than non-mobile ones. Top 10 destination domain names shown in Table 5 are almost the same with that of wired devices except some differences in the rank. In addition to the decrease in queries, the behaviors of wireless users are almost the same with wired users.

Besides, we are also interested in the subdomain names of top 10 destination domain names between wired and wireless users. For example, the domain name of "qq.com", is chosen for this study. Tables 6 and 7 show top 10 subdomain names of "qq.com" for wired and wireless users, respectively. Some of the subdomain names are the same, some like "new-event.sng.qq.com", "short.weixin.qq.com" and "long.weixin.qq.com" are in top 10 rank of wireless users subdomain names, these queries are probably requested by mobile apps, such as, WeChart, a popular app which can be used to chat with friends in a Live Chat session.

As seen in Fig. 9, we can know that "baidu.com" is top 1 destination domain name in all DNS servers, and some domain names, like "google.com", "weibo.com", "taobao.com" all have three edges. In China, "baidu.com" and "google.com" are the most popular search engine, which also provide mail, map and cloud storage services; "taobao.com" is the most popular e-commerce website; and "weibo.com" is a widely used microblog platform; "sina.com.cn" and "qq.com" are the most popular comprehensive website portal including news, blog and mail services. Others, like "renren.com", leading real-name Social Network Service (SNS) in China, "xuelei.com", the first high-definition video portal in China. However, one domain name is unexpected in top 10 destination domain names, "h3q.com", which is resolved from bittorrent tracker – "denis.stalker.h3q.com". This phenomenon stands for large amount of P2P users existed in campus network. All other domain names are popular and common in China, if a strange domain name occurs in top 10, we should pay attention to it.

5.1.2. Anomaly behaviors

We can easily located anomalies with an overall trend. For example, Nearly 4 pm on 26 April 2014, an anomaly occurs in DNS

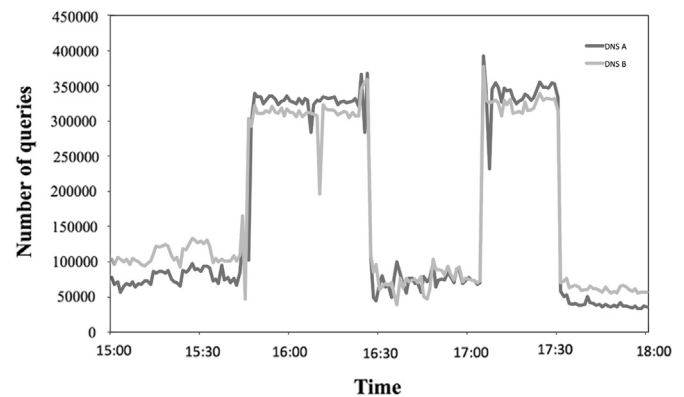


Fig. 10. Query trend of DNS A and DNS B in an anomaly.

Table 8
Overview of three DNS servers' DNS responses in a day.

No.	Total	ServFail	Percent
Total responses	180,774,371	8,126,623	4.50
V4 responses	110,954,171	4,955,398	4.47
V6 responses	38,256,533	1,800,305	4.71
Other	31,563,667	4,955,398	4.34

A and DNS B, which incur network outage for some user communities during that time. Fig. 10 shows this query trend. In two periods, 15:46–16:26 and 17:05–17:30, the queries suddenly increase more than two times than normal level. Furthermore, we digged out top 10 client IP addresses and destination domain names in that two periods. Then we found that one IP address sent queries about 5467 per second, some measures were taken to solve the anomalies by finding this source IP address. So, using the approaches we proposed in Section 4, we could quickly identify anomalies, locate source IP addresses, and solve the problem.

5.2. DNS servers' behaviors

Table 8 is the overview of these servers' DNS responses in a day, about 4.5% requests were failed in name resolving, while 4.47% in IPv4 and 4.71% in IPv6.

Individual DNS responses may contain multiple mappings in response to a single query, we found that only 18.1% IPv4 responses contained only 1 record (including responses resource records, authoritative resource records, extra information records) while the corresponding number in IPv6 is 40.26%. Also 47.44% responses

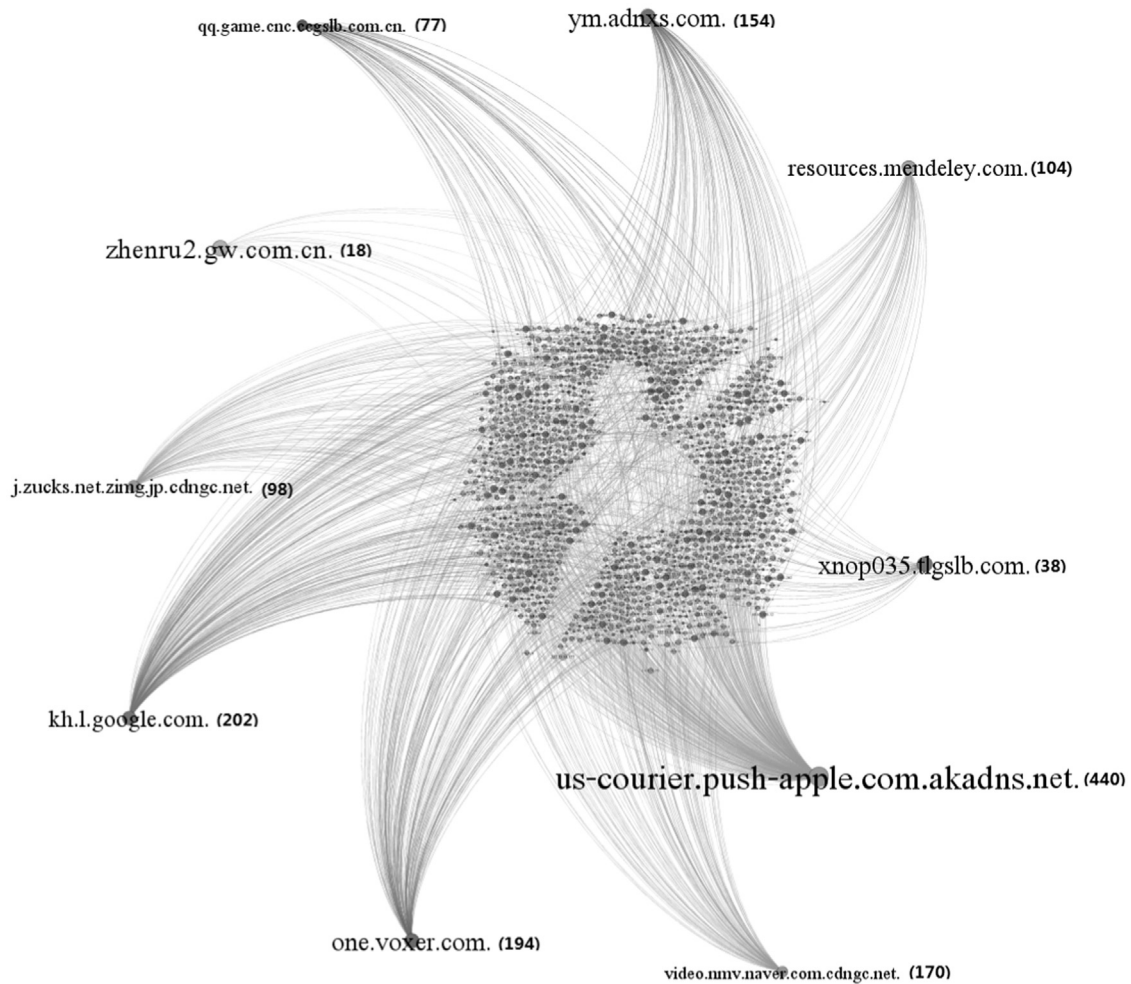


Fig. 11. Domain names having dynamic IP addresses.

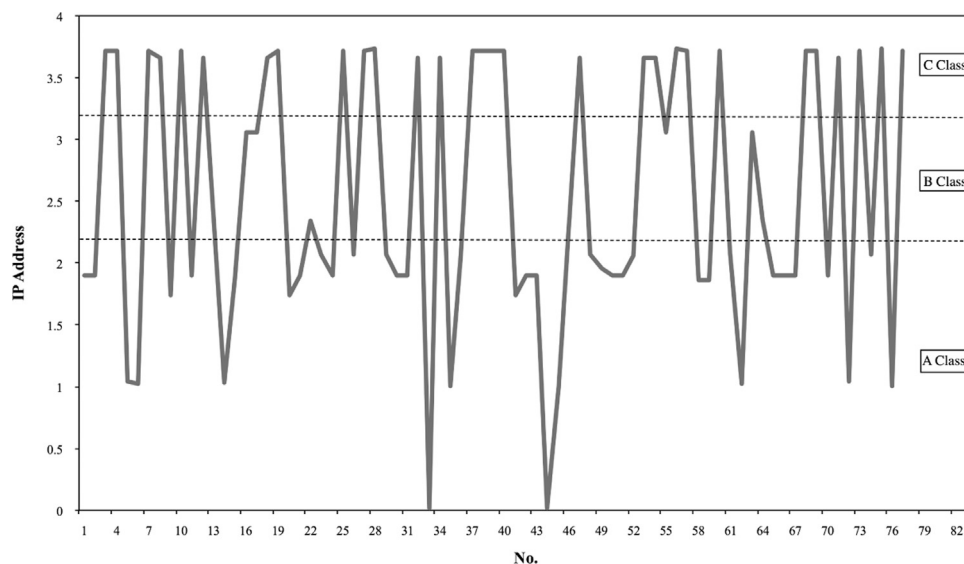


Fig. 12. IP address changing for domain name "qq.game.cnc.ccslb.com.cn".

return less than 10 (including 10) records in IPv4, but 89.96% in IPv6. These indicate that IPv4 resources are richer than IPv6. At last, we studied time-to-live (TTL) associated with hostnames in DNS responses. 9.77% of IPv4 domain names' TTL are less than 60 s, 13.69% are less than 120 s, 23.41% are less than 300 s, 43.26% exceed one day. In IPv6, only 26.90% domain names' TTL exceed one day, but 36.94% less than 300 s, 15.10% less than 120 s. The average TTL of IPv4 domain names is longer than IPv6 domain names.

5.3. Identifying domain names having dynamic IP addresses

To present our results, we used one-day Passive DNS data as an input and selected 10 domain names for example. As shown in Fig. 11, the nodes around are the domain names and total number of dynamic IP addresses is besides the node. The center nodes stand for IP addresses. For example, "qq.game.cnc.ccslb.com.cn.(77)" means the domain name of "qq.game.cnc.ccslb.com.cn", which has 77 corresponding IP addresses. After further study of these 77 addresses, we find that although IP addresses of the domain name change all these IP addresses belong to the same Internet Service Provider (ISP), which can provide some technical reference for optimizing the Internet access route of the campus network.

Fig. 12 shows the dynamics of IP addresses for "qq.game.cnc.ccslb.com.cn.". Three IP address classes, A, B, C, have been marked in the figure, we could see that the IP addresses vary from A Class to C Class, even in the same Class the IP address of this domain name is different. Identifying domain names with dynamic IP addresses and understanding the fluctuation of IP address changing can really help in routing optimization and DNS lookup.

6. Conclusions

In this paper, we used DNS query logs and Passive DNS data to describe the behaviors of DNS lookup. We introduced an algorithm, *DNSReduce*, to dig out top 10 client IP addresses and destination domain names from DNS query logs. This approach showed great accuracy in our experiment. The queries maintained at a high level during daytime, and decreased at mid-night, then climbed up at morning. Wireless user behaviors were almost the same with wired users in addition to a great decrease in queries. Top 10 destination domains were popular and common used in China, as what we expected. At the same time, we took a brief look at behaviors of DNS servers from DNS responses. Our proposed approaches were also helpful in identifying anomalies and troubleshooting. Using Gephi, we identified domain names having dynamic IP addresses and provided some technical reference for routing planning and optimization in a campus network.

In short, these experiment results are very helpful and effective to identify and understand macro- and micro-DNS lookup behaviors, our proposed approaches make a good attempt to visual analysis of DNS logs.

References

- [1] Z. Bojan, B. Nevil, W. Duane, Passive monitoring of dns anomalies, in: Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, Berlin, Heidelberg, 2007, pp. 129–139.
- [2] K. Bonnie, L. Simon, X. Wei, Analyzing Root DNS Traffic (<http://www.eecs.berkeley.edu/bbkirk/papers/cs262a-2004.pdf>), 2004.
- [3] S. Hadi, S. Ali, Ali A. Ghorbani, A survey of visualization systems for network security, *IEEE Trans. Vis. Comput. Graph.* 18 (8) (2012) 1313–1329.
- [4] F. Weimer, Passive DNS replication, in: FIRST Conference on Computer Security Incident, 2005.
- [5] P. Ren, J. Kristo, B. Gooch, Visualizing DNS traffic, in: Proceedings of the 3rd International Workshop on Visualization for Computer Security, ACM, New York, NY, USA, 2006, pp. 23–30.

- [6] A. Schonewille, D.J. van Helmond, The domain name service as an IDS. Research Project for the Master System-and Network Engineering at the University of Amsterdam, 2006.
- [7] P. David, B. Paul, Context-aware clustering of DNS query traffic, in: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement, ACM, New York, NY, USA, 2008, pp. 217–230.
- [8] G.H. Shan, Y. Wang, M.J. Xie, H.P. Lv, X.B. Chi, Visual detection of anomalies in DNS query log data, in: 2014 IEEE Pacific Visualization Symposium (PacificVis), IEEE, New York, NY, USA, 2014, pp. 258–261.
- [9] Conrad Albrecht-buehler, Benjamin Watson, David A Shamma, Visualizing live text streams using motion and temporal pooling, *IEEE Comput. Graph. Appl.* 25 (3) (2005) 52–59.
- [10] Ulrik Brandes, Daniel Fleischer, Jrgen Lerner, Highlighting conflict dynamics in event data, in: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005. IEEE, 2005. New York, NY, USA, pp. 103–110.
- [11] Pieter Lexis, Matthijs Mekking, Identifying Patterns in DNS Traffic, 2013.
- [12] H. Yu, et al., A visualization analysis tool for DNS amplification attack, in: 2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI), IEEE, New York, NY, USA, 2010, pp. 2834–2838.
- [13] Kenton Born, David Gustafson, Ngviz: detecting dns tunnels through n-gram visualization and quantitative analysis, in: Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research, ACM, New York, NY, USA, 2010, p. 47.
- [14] Euihyun Jung, Joonhyung Lim, Juyoung Kim, An Analysis of the Korea National DNS Using Big Data Technology, *Frontier and Innovation in Future Computing and Communications*, Springer, Netherlands (2014) 605–613.
- [15] W.Z. Ruan, Y. Liu, R.L. Zhao, Pattern discovery in DNS query traffic, *Proc. Comput. Sci.* 17 (2013) 80–87.
- [16] David Dagon, Large-scale DNS data analysis, in: Proceedings of the 2012 ACM Conference on Computer and communications Security (CCS '12), ACM, New York, NY, USA, pp. 1054–1055.
- [17] Mark Allman, Thomas Callahan, Michael Rabinovich, On modern DNS behavior and properties, *SIGCOMM Comput. Commun. Rev.* 43 (3) (July 2013) 7–15.
- [18] The Open Graph Viz Platform, (<https://gephi.org/>), May 2014.
- [19] Xuebiao Yuchi, Xin Wang, Xiaodong Lee, Baoping Yan, A New Statistical Approach to DNS Traffic Anomaly Detection, *ADMA 2, Lecture Notes in Computer Science*, vol. 6441, Springer, Heidelberg, Berlin, 2010, pp. 302–313.



Qingnan Lai was born in Jiangxi, China, in 1990. He received the B.S. degree in College of Materials Science and Engineering, University of Science and Technology Beijing, China, in 2012. He is now a graduate student in School of Electronics Engineering and Computer Science, Peking University, China. His research interests include computer network.



Changling Zhou was born in Chongqing, China, in 1977. He received his B.S. and M.S. degrees from Peking University. Currently, he is working at Computer Center of Peking University. He is also a Ph.D candidate of School of Electronics Engineering and Computer Science of Peking University. His current research interests include network traffic analysis, wireless and network management.



Hao Ma was born in Anhui, China, in 1972. He graduated from Peking University, with M.S. in electronics, in 1999. After graduation he has been working as a software developer and network engineer at Computer Center of Peking University. For the past 14 years, he has been engaged in network planning, network management, traffic analysis and information security. Now he is a professor of Engineering in Computer Center, Peking University, and does research on computer networks, next generation Internet and cloud computing.



Zhen Wu was born in Sichuan, China, in 1976. He is now a senior engineer in Computer Center, Peking University, China. His research interests include network management and network security.



Shiyang Chen was born in Inner Mongolia, China, in 1990. She received the B.S. degree in School of Software, Inner Mongolia University, China, in 2013. She is now a graduate student in School of Electronics Engineering and Computer Science, Peking University, China. Her research interests include computer network.