information

Article

DNS Request Log Analysis of Universities in Shanghai: A CDN Service Provider's Perspective

Zhiyang Sun [1], Tiancheng Guo [2], Shiyu Luo [2], Yingqiu Zhuang [2], Yuke Ma [2], Yang Chen [2,*] and Xin Wang [2]

[1] School of Information Science and Technology, Fudan University, Shanghai 200438, China
[2] Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200438, China
* Correspondence: chenyang@fudan.edu.cn

Abstract: Understanding the network usage patterns of university users is very important today. This paper focuses on the research of DNS request behaviors of university users in Shanghai, China. Based on the DNS logs of a large number of university users recorded by CERNET, we conduct a general analysis of the behavior of network browsing from two perspectives: the characteristics of university users' behavior and the market share of CDN service providers. We also undertake experiments on DNS requests patterns for CDN service providers using different prediction models. Firstly, in order to understand the university users' Internet access patterns, we select the top seven universities with the most DNS requests and reveal the characteristics of different university users. Subsequently, to obtain the market share of different CDN service providers, we analyze the overall situation of the traffic distribution among different CDN service providers and its dynamic evolution trend. We find that Tencent Cloud and Alibaba Cloud are leading in both IPv4 and IPv6 traffic. Baidu Cloud has close to 15% in IPv4 traffic, but almost no fraction in IPv6 traffic. Finally, for the characteristics of different CDN service providers, we adopt statistical models, traditional machine learning models, and deep learning models to construct tools that can accurately predict the change in request volume of DNS requests. The conclusions obtained in this paper are beneficial for Internet service providers, CDN service providers, and users.

Keywords: DNS request log; universities; CDN service providers; CERNET; IPv6 protocol; time series prediction

Citation: Sun, Z.; Guo, T.; Luo, S.; Zhuang, Y.; Ma, Y.; Chen, Y.; Wang, X. DNS Request Log Analysis of Universities in Shanghai: A CDN Service Provider's Perspective. *Information* **2022**, *13*, 542. https://doi.org/10.3390/info13110542

Academic Editor: Kurt Maly

Received: 12 October 2022
Accepted: 4 November 2022
Published: 15 November 2022

1. Introduction

In recent years, with the popularity and development of higher education in China, over 34 million students have enrolled in universities and colleges nationwide [1], constituting a representative group of Internet users. Understanding their Internet access behaviors helps Internet service providers better allocate resources and further optimize their services.

Internet users normally need the Domain Name System (DNS) to access Internet services [2]. As a representative Internet service provider (ISP) in China, China Education and Research Network (CERNET) offers DNS services to universities which are members of CERNET. CERNET DNS services support both IPv4 and IPv6 protocols. In particular, the Shanghai branch of CERNET offers Internet services for most universities in Shanghai, such as Shanghai Jiaotong University, Fudan University, and Tongji University. We obtained the DNS request log data provided by CERNET Shanghai from 21 August 2021 to 30 September 2021. Each DNS request included information on the university that sent the request, the domain name, CNAME record, A record, and AAAA record.

Although DNS request log-based analyses have been extensively performed [3–7], works from the perspective of large groups of university users are still lacking. Some researchers [8–11] accomplished great work on using IPv6 by CERNET users, but there

were some limitations on the dataset they used. For example, the time span was relatively short (less than two weeks), and the number of involved universities was only one or two. Moreover, the data sources were mostly traffic volume data, unlike DNS data which provides more accurate access to each request and the associated content delivery network (CDN) requests. There also exist some studies [12–16] on the prediction of CDN request numbers, but such studies did not consider university users as a targeted user group, and the data collection required a lot of time and equipment.

To overcome the main shortcomings of the above works, we leverage a representative DNS request log dataset spanning more than one month, containing requests from university users in Shanghai. We first analyze the dynamic changes in the number of DNS requests for different universities and refine the possible causes in relation to the universities' characteristics. Subsequently, we collect the CNAME records of 13 mainstream Chinese CDN companies and focus on analyzing the access to CDN services in the DNS requests of university users. Based on the analysis results, we compare the prediction performance of statistical models, traditional machine learning models, and deep learning models for the number of CDN requests.

Our main contributions are listed below:

- We select the DNS request log data spanning more than one month, containing requests from university Internet users in Shanghai, and conduct a systematic analysis. We find that DNS requests of different types of universities have distinctive characteristics of fluctuation within a day, as well as significant differences over semesters and holidays.
- We analyze the usage shares of representative CDN service providers in IPv4 and IPv6 protocols. We find that some mainstream companies have large variability in their share of IPv4 and IPv6 protocols, while some are working in tandem.
- To optimize the dynamic resource scheduling capability of CDN service providers, we adopt three types of models, i.e., statistical models, traditional machine learning models, and deep learning models, to predict the changing pattern of the number of DNS requests of CDN service providers. Results show that the deep learning models can achieve the best prediction accuracy with an average absolute percentage error of less than 3%, which has a practical application value.

This research is of great practical significance for the in-depth understanding of Internet users' behavior and the network service ecosystem in universities in Shanghai and a further targeted optimization of CDN services. In the next sections, we first introduce the related work in Section 2. After that, in Section 3, we introduce how the dataset was collected and analyze the dataset from the perspective of the overall DNS requests of different universities. In Section 4, we study and analyze the DNS requests from the perspective of different CDN service providers. In Section 5, we experimentally discuss the prediction accuracy of three different types of models on DNS request numbers. Finally, Section 6 concludes the work and provides several prospective future work.

2. Related Work

There exist numerous pieces of research on the analysis of DNS request logs. A part of the research [3–7] focused on the logs' information itself, analyzing user behavior in terms of domain names and resolution records (Section 2.1). In addition, there has been another research direction [8–11] that has concentrated on a specific group of university users and analyzed user actions of such groups based on DNS logs and network measurements (Section 2.2). Some researchers focused on the request prediction for CDN service providers [12–16] to optimize the performance of CDN and DNS services (Section 2.3).

2.1. DNS Log

DNS access is a necessary step for users to access the Internet on a daily basis. Therefore, DNS logs are often used for network traffic analysis, users' network behavior analysis, and anomalous network behavior detection, among others. Li et al. [3] proposed a multi-scale hierarchical framework for resolving user behavior ambiguity and polymorphism by

analyzing raw DNS requests and profiling users' network access behavior. Their analysis used network traffic data from 159 end users over 10 days. Lai et al. [7] analyzed query logs from three major DNS servers in a large university network in China to understand the overall trends in network usage. Robberechts et al. [4] analyzed DNS query logs from a Belgian ccTLD registry and designed a system architecture for detecting security threats in the logs and storing the logs. Dan et al. [5] extracted features from incoming email logs and DNS query logs for detecting the sender domain names of spam emails. Ghafir et al. [6] built a malicious-domain blacklist detection mechanism by analyzing DNS request traffic.

2.2. User Actions of CERNET

CERNET is one of the major backbone networks in China. In China's exploration of the next-generation Internet development path from IPv4 to IPv6, CERNET has taken the lead in providing large-scale IPv6 deployment support in a special group of universities. The pilot experiment of CERNET provides technical solutions and practical experience for the overall transition to the next-generation Internet in China. Therefore, the IPv6 development of CERNET can also reveal the latest development status of IPv6 in China to a certain extent. Wu et al. [8] described how CERNET provided IPv6 access services, including security, billing, and roaming services, to students and faculty members of several Chinese universities. Wang et al. [11] performed network anomaly detection by analyzing data collected on CERNET by modeling complex relationships between metrics. Wang et al. [9] found that users preferred flat-rate billing over network traffic-based billing on CERNET deployed at Tsinghua University. Zhang et al. [10] studied adult websites and traffic from multiple perspectives using raw IPv6 traffic from CNGI-CERNET2, which is an IPv6-only network deployment in China that has a clear separation from the IPv4 network [8].

2.3. Request Prediction for CDNs

Hu et al. [12] believed that most users in the same community watched content with a high consistency and used geographic and community information to predict CDN requests. Wu et al. [13] used friend similarity to predict the number of requests accepted by a single CDN and offload the model training to the cloud for accelerated training. Liu et al. [14] avoided network congestion and performed load balancing by predicting requests to the CDN. Hours et al. [15] focused on and predicted the impact of DNS service selection on CDN server location and configuration. Calder et al. [16] proposed a scheme to optimize CDNs with DNS redirection by predicting requests from anycast-based CDNs.

Although related works include the analysis of IPv6 development based on IPv6 traffic and DNS request records [17,18], most of them cover a short period of time of fewer than two weeks and have a limited coverage. In our work, we use a dataset with a time span of up to one month and a half and we cover several universities in Shanghai. Therefore, the conclusions of our analysis are much more convincing.

3. Dataset

Our dataset was derived from the DNS request logs of users from universities in Shanghai. University users access Internet resources through CERNET, which provides DNS recursive resolution services for users and logs the information of all requests. For universities using DNS services provided by CERNET, there are one or more fixed exits from which DNS requests within the campus network are sent to CERNET servers. These exits have relatively fixed IPv4 addresses. Combined with the IPv4 address segments of major universities in Shanghai provided by CERNET, we can deduce which university a request belongs to. Meanwhile, CERNET has its own independent DNS resolution servers for IPv4 and IPv6 DNS requests separately.

In this paper, we used the DNS request logs of CERNET users in universities in Shanghai from 15 August 2021 to 30 September 2021.

3.1. Dataset Overview

The dataset was the DNS request logs collected by CERNET and the corresponding resolution logs, stored by date. The daily log size was approximately 1 GB to 2 GB, with a total dataset size of 57.9 GB.

The daily DNS logs contained several user requests, each on a separate row. Each row consisted of nine fields: source IP address, requested domain name, resolution time, A record resolution address, resolution result code, requested DNS record type, CNAME record, AAAA record resolution address, and server IP address.

3.2. Dataset Description

To preliminarily explore the dataset characteristics, we analyzed the DNS requests for different universities. Among the users of CERNET, we selected the top seven universities in Shanghai in terms of DNS requests for dynamic analysis. The seven universities were the University of Shanghai for Science and Technology, Shanghai University of International Business and Economics, Tongji University, Shanghai University of Sport, Shanghai Publishing and Printing College, ShanghaiTech University, and Shanghai Xingjian College. In order to better understand the web browsing behaviors of university users, we first analyzed the dynamic variation in the number of visits of the seven universities for 24 h a day in September. It could be seen that the web browsing behavior fluctuated greatly at different times of the day. We chose Tongji University and Shanghai University of Sports as two different types of universities for the analysis. Tongji University is a comprehensive university with strong strengths in various disciplines and belongs to the "Double First-Class" initiative universities [19,20]. Shanghai University of Sports is a single-discipline university with strengths in sports and it has a "Double First-Class" initiative discipline in sports [19,20]. "Double First-Class" is the abbreviation of world first-class university and first-class discipline, which is a major strategic decision jointly promoted by China's Ministry of Education, Ministry of Finance, and the National Development and Reform Commission to lay a solid foundation for building a strong nation of higher education [19]. Therefore, these two selected universities were representative.

Figure 1a illustrates that the number of DNS requests at Tongji University rose to an average level during the day around 9:00 a.m., while there was a significant drop at night before 4:00 a.m. As shown in Figure 1b, DNS requests of Shanghai University of Sport were generally less, and the peak of daily DNS requests basically occurred around 8:00 a.m. This suggested that students of Shanghai University of Sport tended to use network resources in the early morning before 9:00 a.m., and they used relatively fewer network resources during the rest of the time. The comparison between Figure 1a,b can reflect the difference between the active time periods of students' Internet access in the two universities. Students at Tongji University wake up and sleep later than students at Shanghai University of Sport on average, and this difference in work and rest may be related to the different focuses of the two universities. Tongji University has more course projects and assignments, while Shanghai University of Sport will have more early morning exercise requirements.

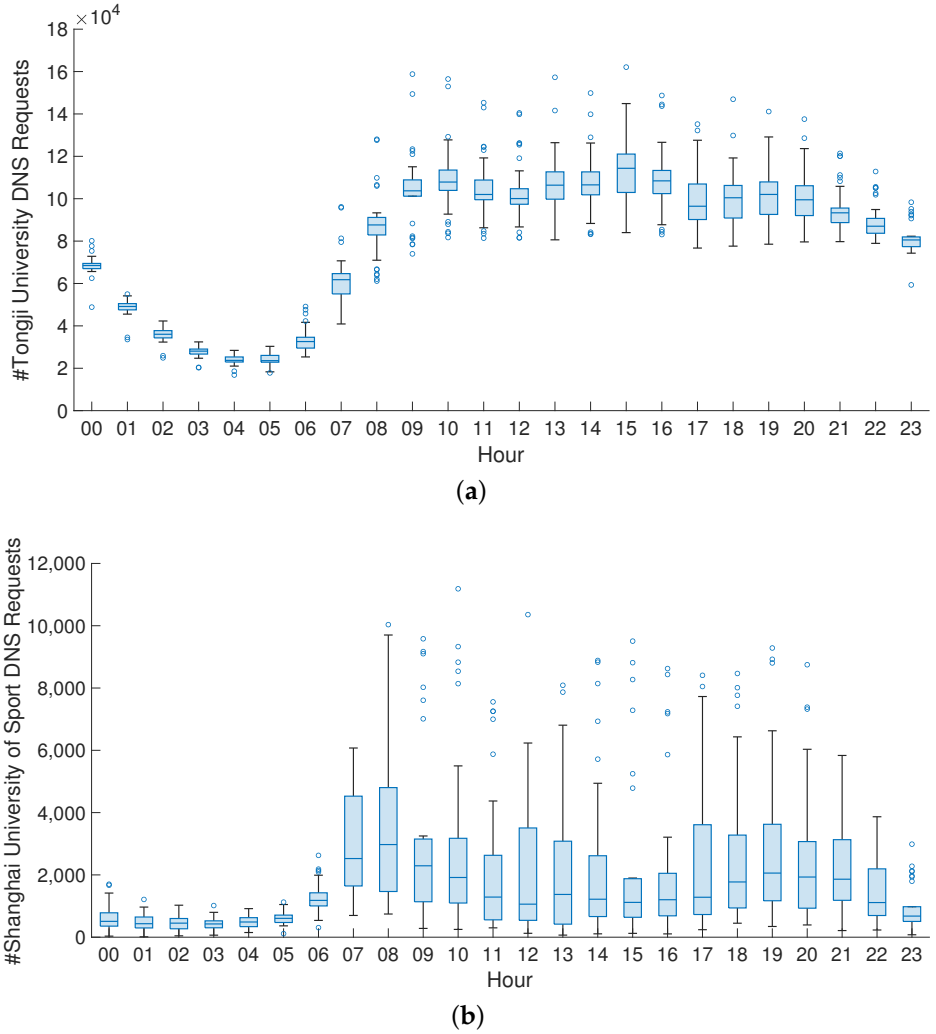


Figure 1. (**a**) Twenty-four-hour DNS requests of Tongji University. (**b**) Twenty-four-hour DNS requests of Shanghai University of Sport.

We conducted a trend dynamic analysis of the data on the number of visits from 21 August 2021 to 30 September 2021, with a daily granularity, for the seven universities. Figure 2 displays that seven universities generally showed a trend of rising DNS requests near the start of the fall semester in early September, which coincided with the pattern of students returning to university after the start of the university year. In addition, except for Shanghai Xingjian College, there was only a small drop in DNS requests during the summer vacation. We inferred that this was due to the high percentage of local students at Shanghai Xingjian College, so the rise in DNS requests due to students returning to school near the end of the summer vacation was more noticeable. In addition, we observed that Tongji University and University of Shanghai for Science and Technology did not experience a significant drop in the number of daily DNS requests during the summer vacation; one possible explanation was that their students would choose to stay on campus more often during the vacation for study, internship, or research.

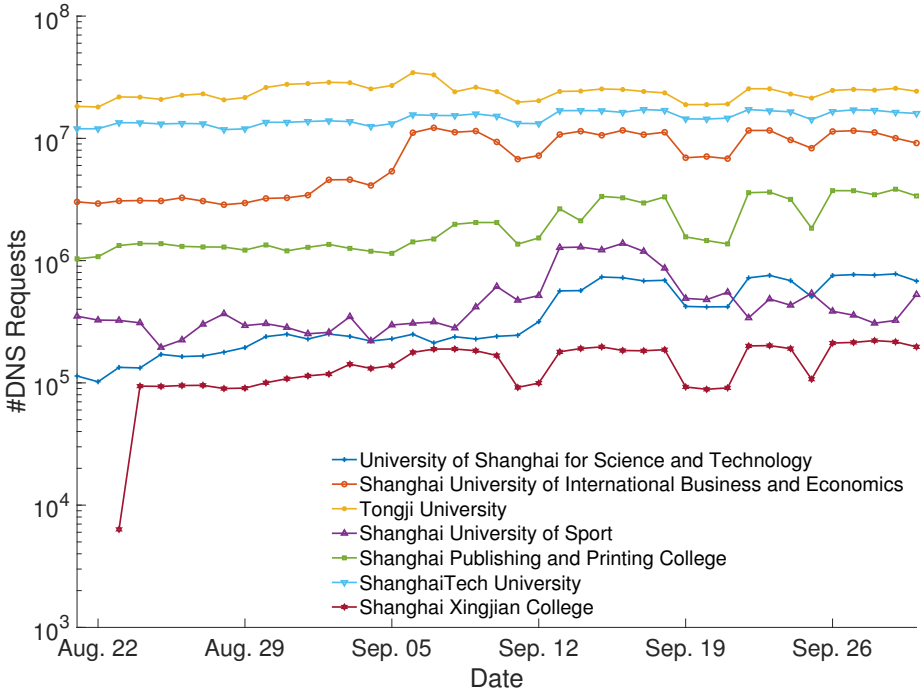
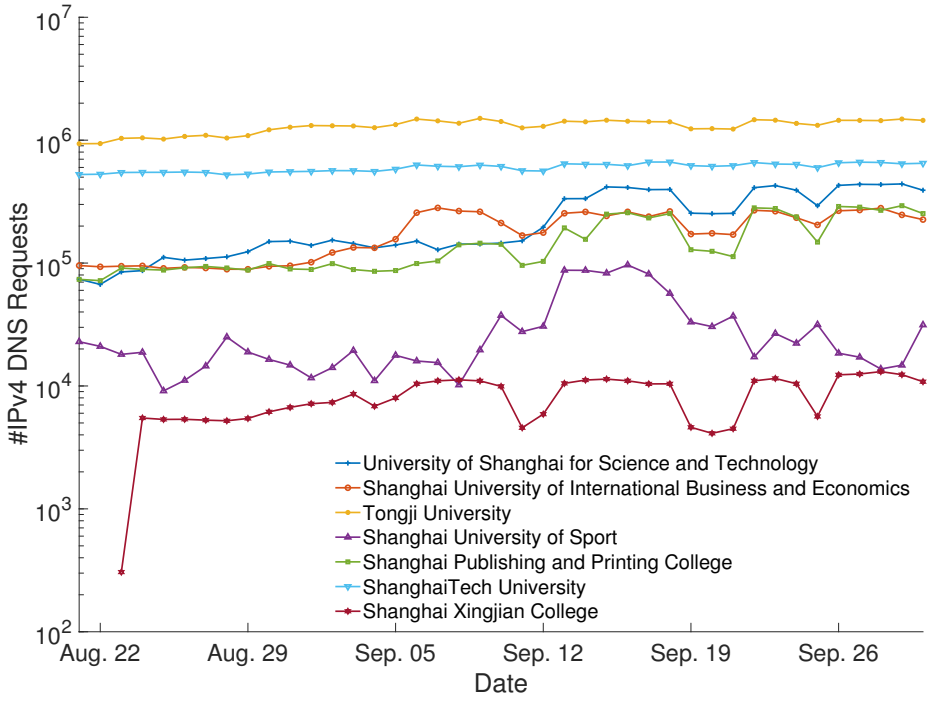


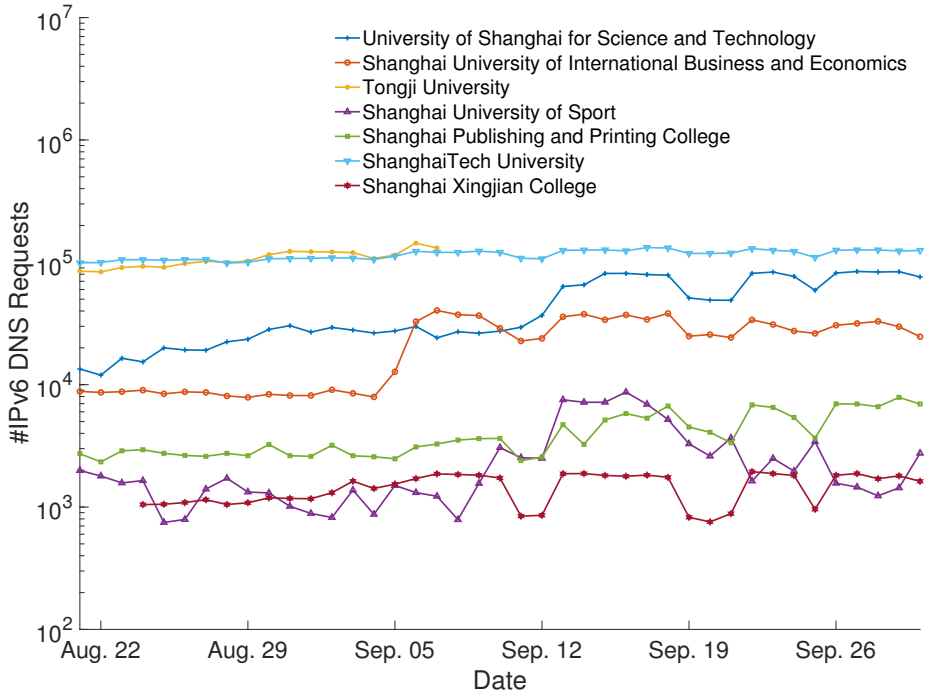
Figure 2. Number of DNS requests from August 21 to September 30 at seven universities.

Figure 3 shows that the trends of daily IPv4 and IPv6 DNS requests were generally consistent across universities. However, for each university, there was an order of magnitude difference in the number of daily DNS requests between IPv4 and IPv6. The results indicated that IPv4 traffic still dominated in the CERNET environment at the time.



(**a**)

Figure 3. *Cont.*



(**b**)

Figure 3. (**a**) DNS request number using IPv4 at seven universities. (**b**) DNS request number using IPv6 at seven universities.

4. Representative CDN Service Providers

As mentioned in Section 3.1, the CNAME record provided in each DNS request log identifies the CDN domain name that served the request. Because CDN domains from the same CDN service provider often have the same prefix or suffix [21], we were able to analyze which CDN service provider was taking over each DNS request based on the CNAME record. CDN servers are deployed in large numbers at the edges of the Internet to provide low-latency access support for users [22], which act as an important infrastructure in the network. With the rapid growth of the Internet, in order to cope with the increasing network traffic and a large number of user requests, many companies choose CDN services to assist in transferring and storing media files [23,24]. It is reported that more than 70% of Internet traffic was carried by CDNs in 2017, and the proportion continues to grow [21]. Understanding the use of CDN services by university users helps CDN service providers better improve user experience.

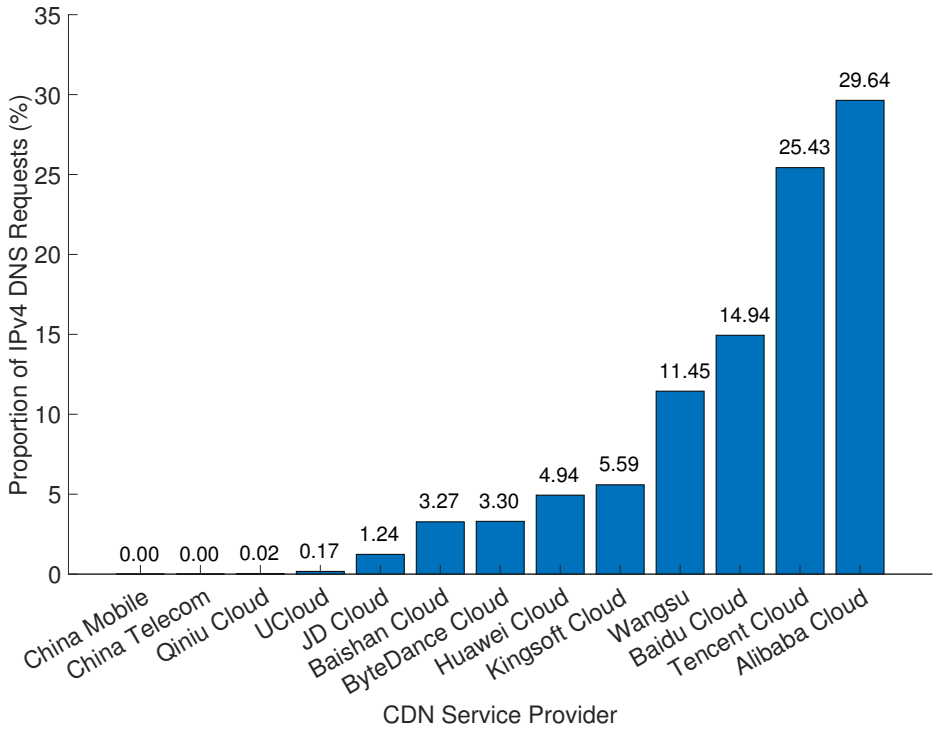
For our analysis, we selected the top 13 CDN service providers ranked by China's National IPv6 Development and Monitoring Platform [25], in order: Huawei Cloud, Tencent Cloud, Alibaba Cloud, Kingsoft Cloud, China Mobile, Baishan Cloud, Baidu Cloud, China Telecom, ByteDance Cloud, JD Cloud, Wangsu, UCloud, and Qiniu Cloud. In order to better match the CNAME record with the corresponding CDN service providers, we obtained the styles of their CDN domain names from these 13 CDN service providers through various ways, such as official website inquiries, telephone consultations, technical report access, blog subscriptions, and measurement validations. Table 1 lists the CDN service providers we used and the main domains covered by the correspondence. We obtain these domain name suffixes from academic papers, blogs, technical reports, and the official websites of various CDN service providers, and put them together. We have done accessibility verification for each domain name suffix. The time span of the table of domain name suffixes acquisition is from May to October 2022.

Table 1. CDN service providers and corresponding domain name suffixes.

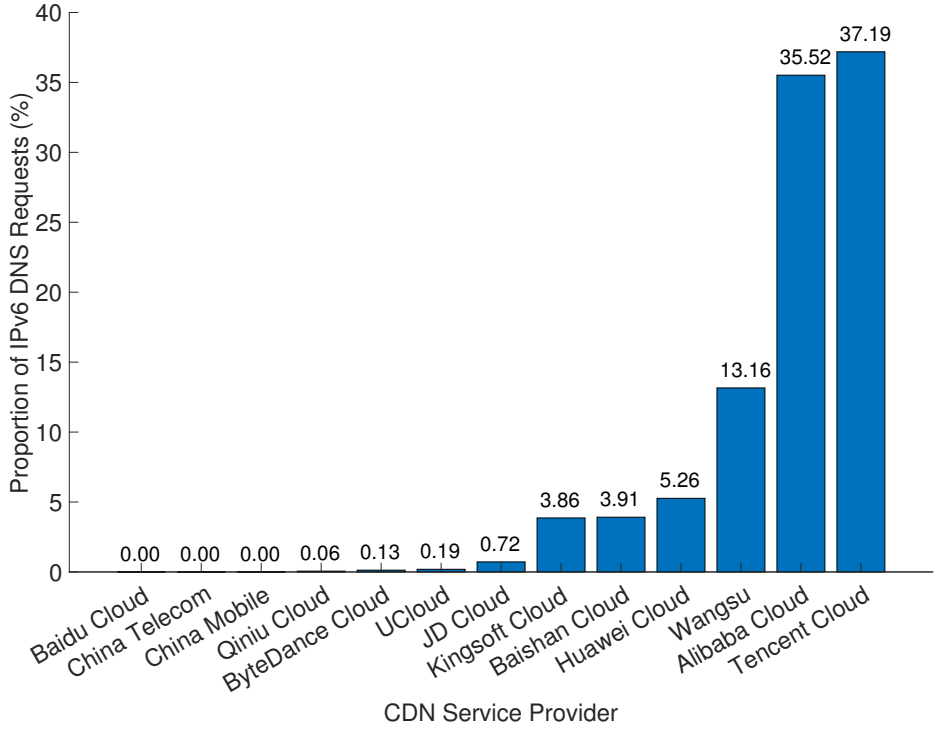
CDN Service Provider	Domain Name Suffixes
Huawei Cloud	c.cdnhwc1.com; c.cdnhwc2.com; c.cdnhwc3.com
Tencent Cloud	dnspod.com; dnsv1.com; *tcdn.qq.com; tcdnlive.com; tdnsv5.com
Alibaba Cloud	aliyundoc.com.cn; aliyundoc.com; w.kunlunsl.com; cdngslb.com; kunluncom; tbcache.com; alicdn.com
Kingsoft Cloud	ks-cdn.com
China Mobile	cdn.10086.cn
Baishan Cloud	qingcdn.com; bsclink.cn; trpcdn.net; bsgslb.cn
Baidu Cloud	bdydns.com; jomodns.com
China Telecom	ctycdn.com
ByteDance Cloud	bytedns.net; cdn.bytedance.com; bytefcdn.com
JD Cloud	jcloud-cdn.com; jdcdn.com
Wangsu	wsdvs.com; wscdns.com; wsglb0.com; cdn20.com; qtlcdn.com; mwcloudcdn.com; lxdns.co
UCloud	ucloud.com.cn
Qiniu Cloud	qiniudns.com

We first compared the number of DNS requests which used IPv4 and IPv6. Figure 4 shows the traffic share of 13 CDN service providers in the number of IPv4 and IPv6 requests, respectively. We found that Tencent Cloud and Alibaba Cloud were at the top of the list in both IPv4 and IPv6 requests, with the sum of the two accounting for more than 50 percent of the total. Alibaba Cloud had the highest share in both IPv4 and IPv6 requests. However, CDN service providers did not necessarily have similar levels of support for IPv4 and IPv6 requests. For example, Baidu Cloud had a share of nearly 15% in IPv4 traffic, while its share was less than 0.01% in IPv6, indicating that Baidu Cloud was a little behind in IPv6 deployment. Wangsu, Huawei Cloud, and Baishan Cloud had a higher share in IPv6 than in IPv4 requests, which indicated that these companies paid more attention to the IPv6 protocol and invested more in it. Since our retention accuracy was two decimal places, the 0.00% of CDN service providers shown in the figure still had a certain number of DNS requests.

Figure 5 shows the changes in the share of different CDN service providers with time series. Overall, the proportion of IPv4 requests for each CDN service provider was relatively stable without obvious fluctuations. The top four CDN service providers were Alibaba Cloud, Tencent Cloud, Baidu Cloud, and Wangsu. The proportions for Alibaba Cloud and Tencent Cloud showed an obvious upward trend in IPv6 requests, which indicated that IPv6 networks were still under development. In addition, the share of the two leading CDN service providers in IPv6 requests was higher than that in IPv4 requests, and the gap between them and other CDN service providers was significantly widened, indicating that the two leading CDN service providers attached great importance to the IPv6 market. Baidu Cloud's share in IPv6 requests was almost zero, meaning that some CDN service providers were still focusing on IPv4 networks.

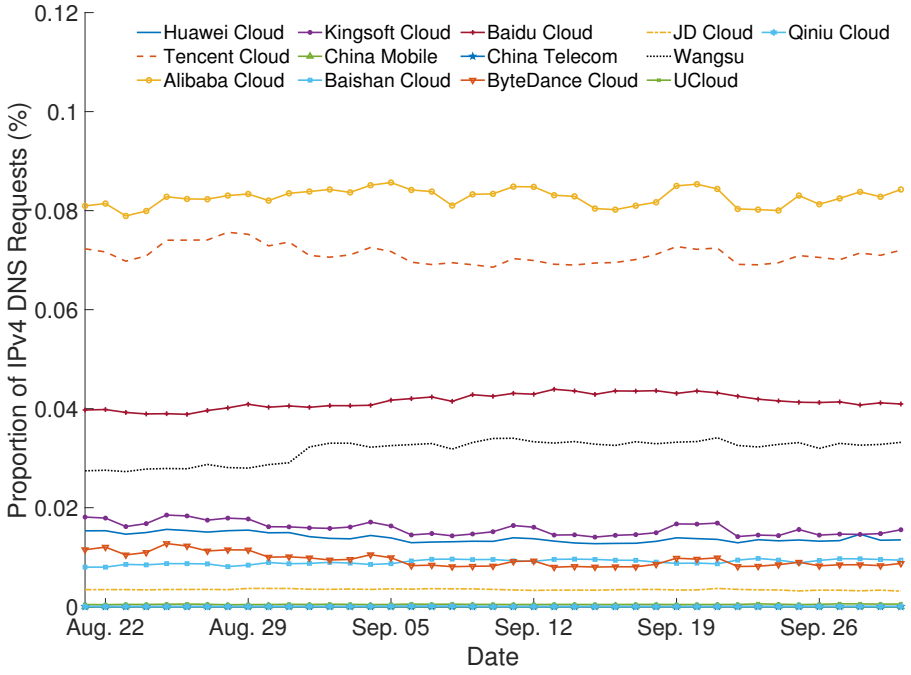


(**a**)

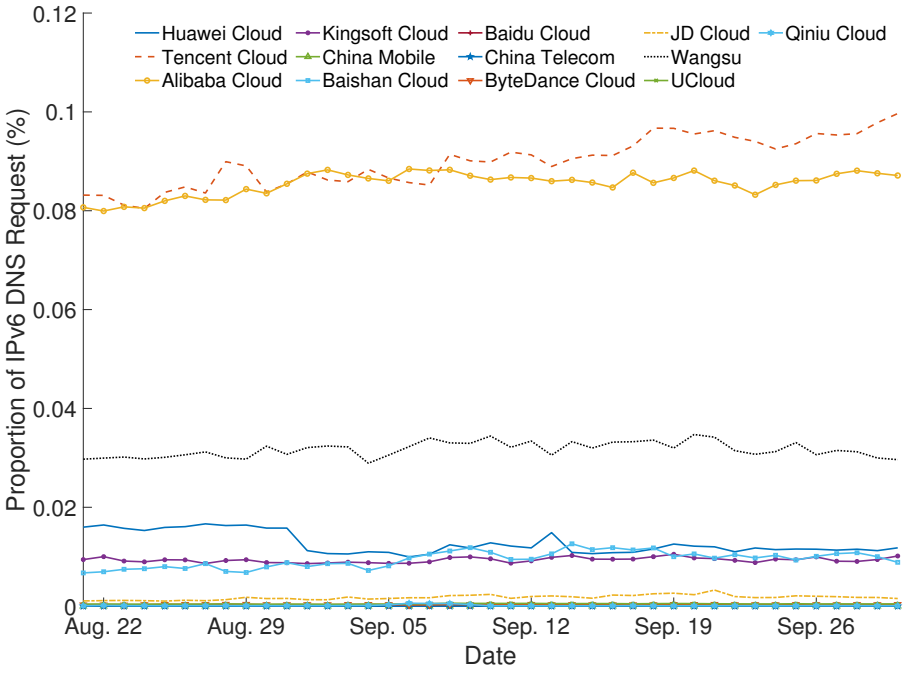


(**b**)

Figure 4. (**a**) Proportion of DNS requests using IPv4 for different CDN service providers. (**b**) Proportion of DNS requests using IPv6 for different CDN service providers.



(**a**)



(**b**)

Figure 5. (**a**) Proportion of DNS requests using IPv4 over time for different CDN service providers. (**b**) Proportion of DNS requests using IPv6 over time for different CDN service providers.

5. Prediction of the Numbers of DNS Requests of CDNs

If CDN service providers could accurately predict the numbers of DNS requests, they would be better able to schedule resources and provide a more stable and efficient service, as in [26]. The relatively stable usage share of different CDN service providers provides better support for the task of predicting the number of DNS requests at a finer granularity. In this section, we use DNS request data from universities and use statistical models, traditional machine learning models, and deep learning models to predict the number of

DNS requests. Our models and experiment results will help CDN service providers better manage the resources dynamically.

5.1. Task Definition

We divided a day into 24 hourly slots, and the number of DNS requests in each slot was the sum of all requests in that slot. For the prediction problem, we wanted to predict the value of the next time interval using a time sequence, which was formulated as follows: assume x_i is the current time interval, the input of the model is continuous data of length 48 $\{x_{i-48}, x_{i-47}, \ldots, x_{i-1}\}$, and the output value of the prediction model is x_i. In this paper, we selected DNS request data from 21 August 2021 to 30 September 2021 and divided them into three prediction datasets based on the CDN information of the data. Based on the analysis in Section 4, we selected Alibaba Cloud and Tencent Cloud as the first two prediction datasets because their market shares far exceeded those of other companies. For the third prediction dataset, we adopted the sum of the DNS request numbers of the top 13 CDN service providers. For each prediction dataset, we selected the sequences whose time periods lay in the first 95% as the training set and the last 5% as the test set, and we used different models to evaluate the prediction results on this problem.

We used three common metrics for evaluating time series prediction models, which are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Their meanings and calculation formulas are shown below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(actual_i - predict_i)^2} \tag{1}$$

$$MAE = \sum_{i=1}^{n}|actual_i - predict_i| \tag{2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{actual_i - predict_i}{actual_i}\right| \tag{3}$$

5.2. Methods and Models

We adopted statistical models, classical machine learning models, and deep learning models to predict the number of incoming DNS requests in each dataset.

5.2.1. Static Models

We adopted the seasonal time series SARIMAX model as a statistical model, which is derived from ARIMA model [27]. An ARIMA model is a classical time series predictive analysis model, which consists of two processes, an autoregressive (AR) process and a moving average (MA). An ARIMA model can be extended to a SARIMA model by adding a difference operator to the seasonal lag when the time series exhibits a seasonal recurring pattern (e.g., daily CDN access counts). The SARIMAX model has been widely used in various time series analysis and forecasting problems, including but not limited to network traffic forecasting [28] and e-commerce products' prices forecasting [29]. In this paper, we considered the effects of exogenous variables on time series and used a SARIMAX model for time series forecasting.

5.2.2. Classical Machine Learning Models

Predicting data values for the next time interval by using some historical data values is a traditional class of supervised machine learning problems. We adopted the classical support vector machine (SVM) method [30] and gradient boosting decision tree (GBDT) algorithm of XGBoost [31] and LightGBM [32] to construct prediction models, and finally compared and analyzed the prediction results of different models. Support vector regression (SVR) is an application of SVMs in regression analysis, which can perform regression analysis and prediction on time series and is widely used in time series prediction prob-

lems such as change and anomaly detection from textural features [33], fake stereo audio identification [34], or drowsiness estimation [35]. XGBoost is an integrated decision tree algorithm in which new trees can correct the results of existing trees in the model so that the model can be made satisfactory by continuously adding decision trees. XGBoost is widely applicable and has been used by researchers for different prediction scenarios, such as image classification [36], intrusion detection [37], malicious account detection [38], and cross-site influential user identification [39]. LightGBM optimizes the temporal and spatial performance based on the traditional GBDT algorithm to speed up the training of GBDT models without compromising the accuracy and has been applied in movie box office prediction [40], credit scoring [41], and network traffic classification [42].

5.2.3. Deep Learning Models

We used a long short-term memory (LSTM) network [43], which is often used in deep learning for temporal data prediction tasks. An LSTM network is based on the ordinary recurrent neural network (RNN) [44], which controls the transmission of information within the network by gating the state and retains the important temporal state in each pass-through step in a long-time memory manner to achieve better prediction results. An LSTM model is generally applicable to various types of continuous temporal prediction models, such as traffic flow prediction [45], network attacks prediction [46], and missing data compensation [47]. Considering the possible discontinuity of DNS data, an additional LSTM variant, time-aware LSTM (T-LSTM) [48], was chosen for comparison in this paper. T-LSTM improves the LSTM for irregular time intervals of time series data, capturing temporal features and burst features more effectively, and is often used for heartbeat prediction [49], disease prediction [50] and malicious account detection [51]. We also propose an LSTM-Attention prediction model, which combines LSTM and the attention mechanism, which can assign the data in different historical periods with different coefficients representing their influence on the prediction results [50,52].

As illustrated in Figure 6, in general, the LSTM-Attention model takes the number of historical DNS requests as input features to predict the number of DNS requests. Specifically, we use x_i to represent the number of DNS requests in an hourly slot, which is defined in Section 5.1. There are n historical slots used as the input features of the model, thus the input X of the model is denoted as $\{x_0, x_1, \ldots, x_{n-1}\}$. The input X is first sent to an LSTM neural network module named A, where it is processed and then the output to $\{h_0, h_1, \ldots, h_{n-1}\}$. Following the LSTM layer, the attention layer processes the outputs of the LSTM layer. In this process, there are a linear layer and a softmax layer as the branches for calculating the weight w_i which represents the impact of x_i on the prediction value x_n. After that, the weights w_i and the output h_i of the LSTM layers are weighted and fused to obtain the prediction value x_n, which is the number of DNS requests at the next time slot to be predicted by the model.

5.3. Results

We analyzed the prediction results of different models according to Table 2. Intuitively, the more recent the historical data are, the more effective they are at predicting future results. The results showed that the prediction performance of the deep learning models was better than the traditional machine learning models, and both were better than the statistical models. Among them, the LSTM-Attention model performed better than the LSTM and T-LSTM models in most cases, indicating that the model could fit the temporal information better and obtain more accurate prediction results. There was only one case in Table 2 where the LSTM-Attention model did not perform best when we used the MAPE metric to evaluate the Tencent Cloud dataset. The LSTM-Attention model performed the second best after the LSTM model with a small difference. One possible explanation is that the MAPE metric is asymmetric, which imposes a greater penalty for negative errors (when the predicted value is higher than the actual value) than for positive errors, favoring models that underpredict rather than overpredict. Due to the short duration

of our prediction dataset, the LSTM-Attention model showed more overfitting than the LSTM model, although the overall prediction error was more accurate. Therefore, the LSTM-Attention model performed slightly worse than the LSTM model when the MAPE metric was used to evaluate the Tencent Cloud dataset.

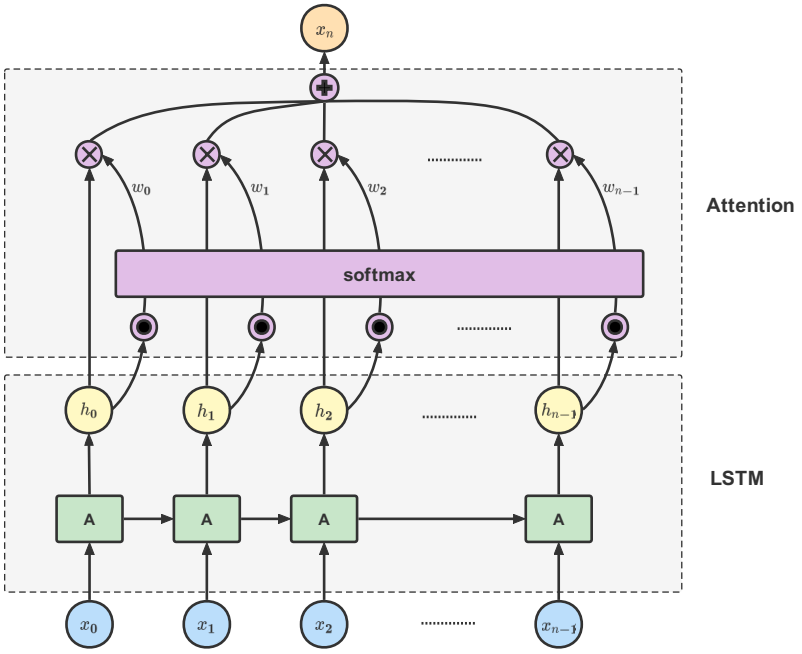
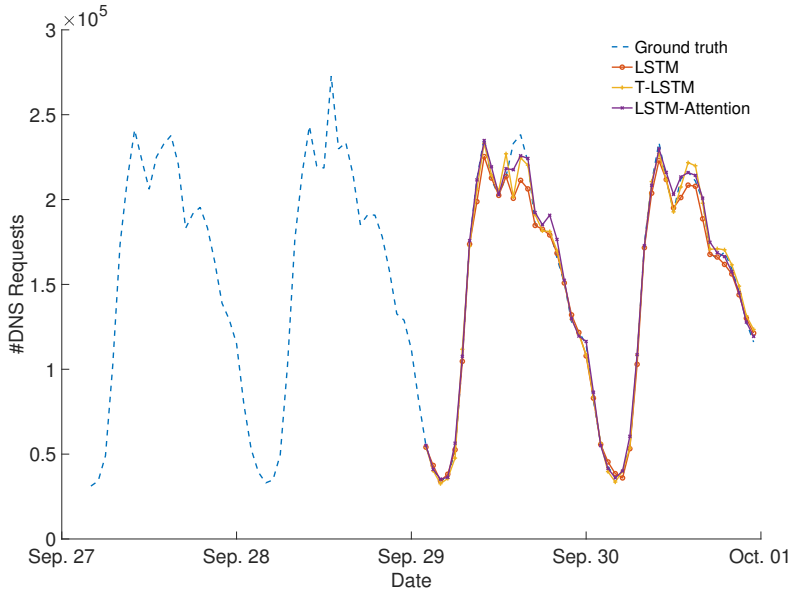


Figure 6. Prediction results using deep learning models.

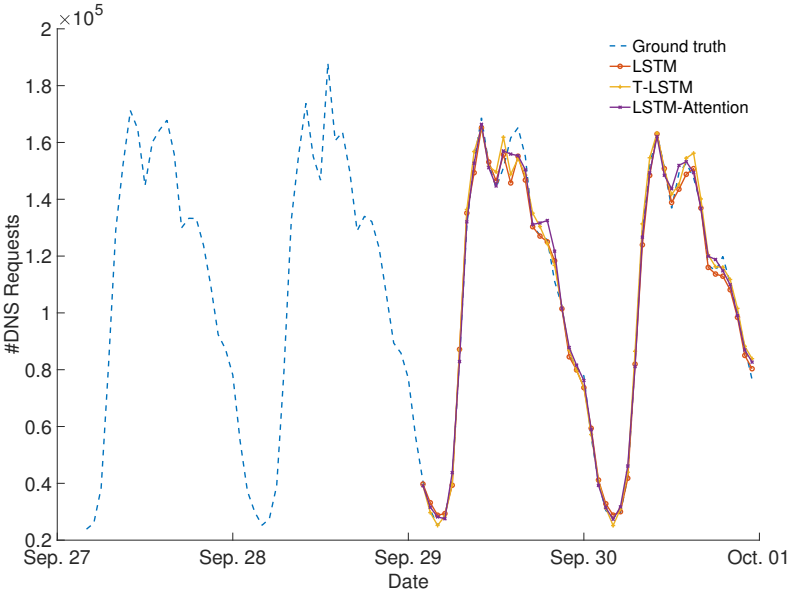
Meanwhile, we compared the three prediction datasets, and the results showed that predicting the total number of DNS accesses for the 13 CDN service providers was more accurate than predicting Tencent Cloud or Alibaba Cloud alone. The total number of predicted DNS accesses for the Alibaba Cloud, Tencent Cloud, and the 13 CDN service providers are illustrated in Figure 7, respectively. The predicted curves almost fit the actual DNS request changes. Therefore, the prediction of DNS requests in CDNs using a deep learning approach is practical and usable. The high-accuracy prediction results of deep learning can effectively help CDN companies predict the number of requests in the next hour in advance, adjust the configuration of CDNs more flexibly, and ensure the stability and efficiency of the service. Overall, the LSTM-Attention model best fitted the ground-truth trace and could be chosen as the default prediction approach.

Table 2. Prediction results of the three prediction datasets with different metrics and models.

Dataset	Metric	SARIMAX	SVR	XGBoost	LightGBM	LSTM	T-LSTM	LSTM-Attention
Alibaba Cloud	RMSE ($\times 10^3$)	17.5	203	8.79	7.83	7.07	6.63	**5.25**
	MAE ($\times 10^3$)	13.8	48.3	6.00	4.91	4.32	4.27	**3.83**
	MAPE (%)	11.6	56.4	4.50	3.98	3.04	2.91	**2.87**
Tencent Cloud	RMSE ($\times 10^3$)	21.1	42.1	10.0	8.46	5.08	4.14	**4.06**
	MAE ($\times 10^3$)	17.7	35.9	6.96	6.03	3.22	2.92	**2.89**
	MAPE (%)	25.0	50.7	6.83	5.41	**2.84**	3.02	2.96
13 CDN service providers	RMSE ($\times 10^3$)	63.9	57.1	31.1	21.6	22.0	23.0	**17.2**
	MAE ($\times 10^3$)	56.0	173	33.2	22.1	14.5	13.4	**13.0**
	MAPE (%)	10.6	48.2	4.81	3.96	2.68	2.72	**2.66**

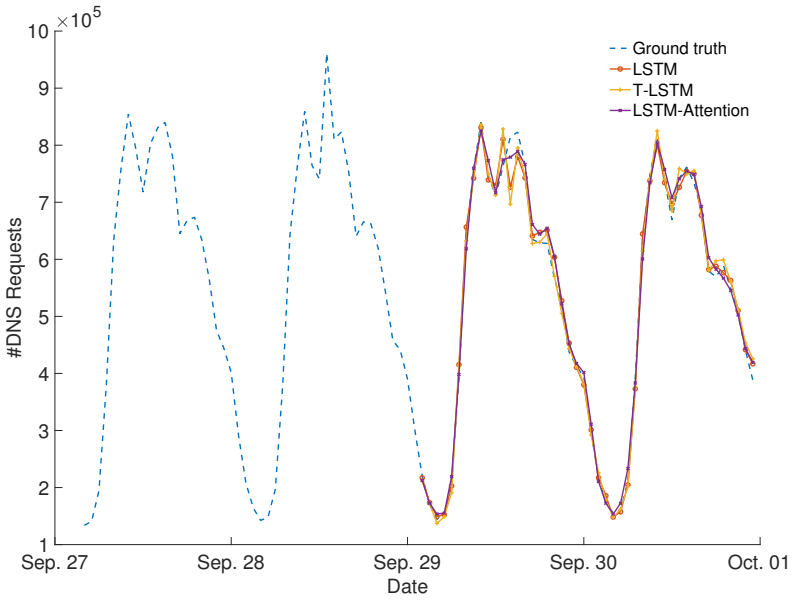


(**a**)



(**b**)

Figure 7. *Cont.*



(**c**)

Figure 7. (**a**) Prediction results using deep learning models on the Alibaba Cloud dataset. (**b**) Prediction results using deep learning models on the Tencent Cloud dataset. (**c**) Prediction results using deep learning models on the 13 CDN service providers dataset.

6. Conclusions and Future Work

In this paper, we focused on the DNS request patterns of university Internet users in Shanghai, China, and conducted a multidimensional analysis of these users' DNS request logs from a CDN provider's perspective. The dataset of DNS request logs collected from CERNET contained several universities and covered a long time span. The results of our analysis provided insights on how universities can improve their own Internet services, how CERNET can better optimize network resources, and how CDN service providers can better serve university users. Our deep learning-based prediction model obtained quite accurate prediction results in predicting the number of incoming DNS requests. In particular, the combination of LSTM and the attention mechanism achieved the best performance, which is of practical application to the traffic scheduling and resource allocation of CDN service providers.

In the future, we will explore more datasets of DNS requests from universities in different cities. Moreover, we will extend the time span of the dataset and refine the time granularity to better validate the scalability and generalizability of our proposed model. In addition, we will consider combining IPv4 and IPv6 traffic data to better explore the development and evolution of IPv6 in China.

Author Contributions: Conceptualization, Z.S. and T.G.; methodology, T.G. and Y.C.; software, S.L., Y.Z. and Y.M.; validation, Z.S., T.G., S.L. and Y.C.; formal analysis, Z.S.; investigation, Z.S.; resources, Z.S., Y.C. and X.W.; data curation, Z.S.; writing—original draft preparation, T.G, S.L., Y.Z. and Y.M.; writing—review and editing, T.G., Y.C., Z.S. and X.W.; visualization, S.L., Y.Z. and Y.M.; supervision, Y.C. and X.W.; project administration, Z.S. and X.W.; funding acquisition, Z.S., Y.C. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (no. 61971145).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DNS	Domain Name System
CDN	Content delivery network
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
ISP	Internet service provider
CERNET	China Education and Research Network
RMSE	Root-mean-square error
MAE	Mean absolute error
SVM	Support vector machine
SVR	Support vector regression
GBDT	Gradient boosting decision tree
LSTM	Long short-term memory
T-LSTM	Time-aware LSTM
RNN	Recurrent neural network

References

1. The Main Results of the National Education Statistics in 2021. Available online: http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202203/t20220301_603262.html (accessed on 1 March 2022).
2. Mockapetris, P.V. Domain names—Concepts and facilities. *RFC* **1987**, *1034*, 1–55. [CrossRef]
3. Li, J.; Ma, X.; Li, G.; Luo, X.; Zhang, J.; Li, W.; Guan, X. Can We Learn what People are Doing from Raw DNS Queries? In Proceedings of the 2018 IEEE Conference on Computer Communications, INFOCOM 2018, Honolulu, HI, USA, 16–19 April 2018; pp. 2240–2248. [CrossRef]
4. Robberechts, P.; Bosteels, M.; Davis, J.; Meert, W. Query Log Analysis: Detecting Anomalies in DNS Traffic at a TLD Resolver. In Proceedings of the ECML PKDD 2018 Workshops-DMLE 2018 and IoTStream 2018, Dublin, Ireland, 10–14 September 2018; Springer: Berlin, Germany, 2018; Volume 967, pp. 55–67. [CrossRef]
5. Dan, K.; Kitagawa, N.; Sakuraba, S.; Yamai, N. Spam Domain Detection Method Using Active DNS Data and E-Mail Reception Log. In Proceedings of the 43rd IEEE Annual Computer Software and Applications Conference, COMPSAC 2019, Milwaukee, WI, USA, 15–19 July 2019; pp. 896–899. [CrossRef]
6. Ghafir, I.; Prenosil, V. DNS traffic analysis for malicious domains detection. In Proceedings of the 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, Delhi, 19–20 February 2015; pp. 613–918.
7. Lai, Q.; Zhou, C.; Ma, H.; Wu, Z.; Chen, S. Visualizing and characterizing DNS lookup behaviors via log-mining. *Neurocomputing* **2015**, *169*, 100–109. [CrossRef]
8. Wu, J.; Wang, J.H.; Yang, J. CNGI-CERNET2: An IPv6 deployment in China. *Comput. Commun. Rev.* **2011**, *41*, 48–52. [CrossRef]
9. Wang, J.H.; An, C.; Yang, J. A study of traffic, user behavior and pricing policies in a large campus network. *Comput. Commun.* **2011**, *34*, 1922–1931. [CrossRef]
10. Zhang, S.; Zhang, H.; Yang, J.; Song, G.; Wu, J. Measurement and Analysis of Adult Websites in IPv6 Networks. In Proceedings of the 20th Asia-Pacific Network Operations and Management Symposium, APNOMS 2019, Matsue, Japan, 18–20 September 2019; pp. 1–6. [CrossRef]
11. Wang, Z.; Yang, J.; Zhang, S.; Li, C.; Zhang, H. Automatic Model Selection for Anomaly Detection. In Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, China, 23–26 August 2016; pp. 276–283. [CrossRef]
12. Hu, H.; Wen, Y.; Chua, T.; Wang, Z.; Huang, J.; Zhu, W.; Wu, D. Community based effective social video contents placement in cloud centric CDN network. In Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2014, Chengdu, China, 14–18 July 2014; pp. 1–6. [CrossRef]
13. Wu, C.; Chen, X.; Zhu, W.; Zhang, Y. Socially-Driven Learning-Based Prefetching in Mobile Online Social Networks. *IEEE/ACM Trans. Netw.* **2017**, *25*, 2320–2333. [CrossRef]
14. Liu, J.; Yang, Q.; Simon, G. Congestion Avoidance and Load Balancing in Content Placement and Request Redirection for Mobile CDN. *IEEE/ACM Trans. Netw.* **2018**, *26*, 851–863. [CrossRef]
15. Hours, H.; Biersack, E.W.; Loiseau, P.; Finamore, A.; Mellia, M. A study of the impact of DNS resolvers on CDN performance using a causal approach. *Comput. Netw.* **2016**, *109*, 200–210. [CrossRef]
16. Calder, M.; Flavel, A.; Katz-Bassett, E.; Mahajan, R.; Padhye, J. Analyzing the Performance of an Anycast CDN. In Proceedings of the 2015 ACM Internet Measurement Conference, IMC 2015, Tokyo, Japan, 28–30 October 2015; ACM: 2015; pp. 531–537. [CrossRef]

17. Han, C.; Li, Z.; Xie, G.; Uhlig, S.; Wu, Y.; Li, L.; Ge, J.; Liu, Y. Insights into the issue in IPv6 adoption: A view from the Chinese IPv6 Application mix. *Concurr. Comput. Pract. Exp.* **2016**, *28*, 616–630. [CrossRef]
18. Gao, H.; Yegneswaran, V.; Chen, Y.; Porras, P.A.; Ghosh, S.; Jiang, J.; Duan, H. An empirical reexamination of global DNS behavior. In Proceedings of the ACM SIGCOMM 2013 Conference, SIGCOMM 2013, Hong Kong, China, 12–16 August 2013; pp. 267–278. [CrossRef]
19. Notice on the Announcement of the Second Round of "Double First-class" Initiative Construction Universities and Construction Disciplines. Available online: http://www.gov.cn/zhengce/zhengceku/2022-02/14/content_5673496.htm (accessed on 12 March 2022).
20. Sun, J.; Li, Y.; Zhao, X.; Zhang, N. An Evaluation on Investment of Research Funds with a Neural Network Algorithm in "Double First-Class" Universities. *Complex* **2020**, *2020*, 7496126:1–7496126:8. [CrossRef]
21. Yang, J.; Sabnis, A.; Berger, D.S.; Rashmi, K.V.; Sitaraman, R.K. C2DN: How to Harness Erasure Codes at the Edge for Efficient Content Delivery. In Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2022, Renton, WA, USA, 4–6 April 2022; pp. 1159–1177.
22. Zhou, M.; Guo, T.; Chen, Y.; Wan, J.; Wang, X. Polygon: A QUIC-based CDN server selection system supporting multiple resource demands. In Proceedings of the the 22nd International Middleware Conference: Industrial Track, Québec City, QC, Canada, 6–10 December 2021; ACM: Windsor, ON, Canada, 2021; pp. 16–22. [CrossRef]
23. Wang, J. A survey of web caching schemes for the Internet. *Comput. Commun. Rev.* **1999**, *29*, 36–46. [CrossRef]
24. Wang, K.; Zhang, J.; Bai, G.; Ko, R.K.L.; Dong, J.S. It's Not Just the Site, It's the Contents: Intra-domain Fingerprinting Social Media Websites Through CDN Bursts. In Proceedings of the WWW '21: The Web Conference 2021, Virtual Event, Ljubljana, Slovenia, 19–23 April 2021; pp. 2142–2153. [CrossRef]
25. National IPv6 Development and Monitoring Platform. Available online: https://www.china-ipv6.cn/#/client/simpleInfo (accessed on 19 April 2022).
26. Li, X.; Chen, Y.; Zhou, M.; Guo, T.; Wang, C.; Xiao, Y.; Wan, J.; Wang, X. Artemis: A Latency-Oriented Naming and Routing System. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 4874–4890. [CrossRef]
27. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
28. Jiang, W. Internet traffic prediction with deep neural networks. *Internet Technol. Lett.* **2022**, *5*, e314. [CrossRef]
29. Carta, S.; Medda, A.; Pili, A.; Recupero, D.R.; Saia, R. Forecasting E-Commerce Products Prices by Combining an Autoregressive Integrated Moving Average (ARIMA) Model and Google Trends Data. *Future Internet* **2019**, *11*, 5. [CrossRef]
30. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
31. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
32. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3154.
33. Avola, D.; Cinque, L.; Mambro, A.D.; Diko, A.; Fagioli, A.; Foresti, G.L.; Marini, M.R.; Mecca, A.; Pannone, D. Low-Altitude Aerial Video Surveillance via One-Class SVM Anomaly Detection from Textural Features in UAV Images. *Information* **2022**, *13*, 2. [CrossRef]
34. Liu, T.; Yan, D.; Wang, R.; Yan, N.; Chen, G. Identification of Fake Stereo Audio Using SVM and CNN. *Information* **2021**, *12*, 263. [CrossRef]
35. Akbar, I.A.; Igasaki, T. Drowsiness Estimation Using Electroencephalogram and Recurrent Support Vector Regression. *Information* **2019**, *10*, 217. [CrossRef]
36. Jiao, W.; Hao, X.; Qin, C. The Image Classification Method with CNN-XGBoost Model Based on Adaptive Particle Swarm Optimization. *Information* **2021**, *12*, 156. [CrossRef]
37. Dhaliwal, S.S.; Nahid, A.A.; Abbas, R. Effective Intrusion Detection System Using XGBoost. *Information* **2018**, *9*, 149. [CrossRef]
38. Gong, Q.; Chen, Y.; He, X.; Zhuang, Z.; Wang, T.; Huang, H.; Wang, X.; Fu, X. DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks. *IEEE Commun. Mag.* **2018**, *56*, 21–27. [CrossRef]
39. Gong, Q.; Chen, Y.; He, X.; Xiao, Y.; Hui, P.; Wang, X.; Fu, X. Cross-site Prediction on Social Influence for Cold-start Users in Online Social Networks. *ACM Trans. Web.* **2021**, *15*, 6:1–6:23. [CrossRef]
40. Ni, Y.; Dong, F.; Zou, M.; Li, W. Movie Box Office Prediction Based on Multi-Model Ensembles. *Information* **2022**, *13*, 299. [CrossRef]
41. Niu, B.; Ren, J.; Li, X. Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. *Information* **2019**, *10*, 397. [CrossRef]
42. Hua, Y. An efficient traffic classification scheme using embedded feature selection and lightgbm. In Proceedings of the 2020 Information Communication Technologies Conference (ICTC), Jeju Island, 21–23 October 2020; pp. 125–130.
43. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
44. Bengio, Y.; Simard, P.Y.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]

45. Xie, Q.; Guo, T.; Chen, Y.; Xiao, Y.; Wang, X.; Zhao, B.Y. Deep Graph Convolutional Networks for Incident-Driven Traffic Speed Prediction. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management, Virtual Event, 19–23 October 2020; pp. 1665–1674. [CrossRef]
46. Muhuri, P.S.; Chatterjee, P.; Yuan, X.; Roy, K.; Esterline, A.C. Using a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to Classify Network Attacks. *Information* **2020**, *11*, 243. [CrossRef]
47. Kwon, H.; Kim, P. A Missing Data Compensation Method Using LSTM Estimates and Weights in AMI System. *Information* **2021**, *12*, 341. [CrossRef]
48. Baytas, I.M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A.K.; Zhou, J. Patient Subtyping via Time-Aware LSTM Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 65–74. [CrossRef]
49. Ma, F.; Gao, J.; Suo, Q.; You, Q.; Zhou, J.; Zhang, A. Risk Prediction on Electronic Health Records with Prior Medical Knowledge. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, 19–23 August 2018; pp. 1910–1919. [CrossRef]
50. Zhang, Y.; Yang, X.; Ivy, J.S.; Chi, M. ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 4369–4375. [CrossRef]
51. Ye, Q.; Gao, Y.; Zhang, Z.; Chen, Y.; Li, Y.; Gao, M.; Chen, S.; Wang, X.; Chen, Y. Modeling Access Environment and Behavior Sequence for Financial Identity Theft Detection in E-Commerce Services. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–22 July 2022; pp. 1–8. [CrossRef]
52. Brauwers, G.; Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Trans. Knowl. Data Eng.* **2021**. [CrossRef]