Network Traffic Prediction Based on the Feature of Newly-Generated Network Flows

Shaohe Li[1,2], Junping Song[1], Luyang Xu[1,2], Yahui Hu[3], Wanming Luo[1], Xu Zhou[1]
[1]*Computer Network Information Center, Chinese Academy of Sciences, China*
[2]*University of Chinese Academy of Sciences, China*
[3]*China University of Mining & Technology, Beijing, China*
E-mail: {lishaohe, songjunping, xuluyang, lwm, zhouxu}@cnic.cn[1], huyahui@cumtb.edu.cn[3]

***Abstract*—Network traffic prediction is essential for intelligent network management, such as resource reservation and burst warning. Existing prediction approaches are vulnerable in accurately capturing the sudden surge or plunge, uniformly denoted as the traffic burst. To solve this problem, we extract the time series of the number of newly-generated network flows (NoNGF) from the network flow information, explaining the intrinsic mechanism of network traffic bursts. We use time-lagged cross-correlation analysis to identify directionality between the NoNGF series and traffic series. It proves that we can perceive the future fluctuation and burst of network traffic by NoNGF in advance. The comprehensive prediction experiments of the whole network traffic and three application-level network traffic demonstrate that our proposed approach exhibits a significant performance improvement over the original LSTM and TCN models. Our approach can accurately capture the moment of network burst and the predicted value much more precisely when the burst occurs. In summary, our proposed traffic prediction based on NoNGF can significantly improve the prediction accuracy, especially for network burst traffic.**

***Keywords*—Network traffic prediction, network flow feature, traffic burst prediction, cross-correlation analysis**

I. INTRODUCTION

Network traffic has increased exponentially, and network burst has become more frequent, because of the massive network devices access to the Internet and the further improvement of users' requirements for network service quality [1,2]. Widely employing advanced network technologies makes intelligent network control a reality [3]. In order to improve the quality of service and optimize the allocation of network resources in advance, it is essential to predict the network traffic with high accuracy [4].

Network traffic prediction makes forecasts of future traffic demands by observing the historical time series data, which is a kind of time series prediction. The field of network traffic prediction based on deep learning has made great progress in recent years [5]. Deep learning methods can learn high-dimensional features of the historical series of network traffic for predicting the future traffic. Because network traffic varies drastically and traffic burst is highly random, helpful features of random fluctuations and bursts cannot be learned from historical traffic series [6]. Predicting random network traffic fluctuations and traffic bursts is a difficult problem of network traffic prediction, and it usually causes non-robust prediction, which means the predicted value is very close to the previous true value.

As a novel idea of the network traffic prediction method, we mine the feature series strongly correlated with the network traffic from the network flow data and input it as a covariate into the prediction models. The generation of a network flow indicates that communicating parties have established an information path in a network. Data will be transmitted in the following period, generating network traffic volume. Generating a network flow affects traffic volume for the next few minutes. We count the number of new network flows generated per unit time which we call the number of newly-generated network flows (NoNGF). This feature is simple to obtain but remarkably positively affects the traffic prediction accuracy by inputting into the prediction model as an external variable along with the network traffic. If the prediction model accepts not only the predicted time series, but also other external variables (called covariate), then this prediction method is called covariate prediction [7]. In this paper, we propose a covariate assisted prediction method of network traffic based on NoNGF series as the covariate. Analyzing or applying external features as covariates for network traffic prediction have not been proposed yet. Our research work is the first to use external features of network traffic as covariates for traffic prediction and significantly improves the prediction accuracy.

We theoretically demonstrate that the NoNGF feature variable has a strong correlation with the traffic series by feature engineering and cross-correlation analysis. Moreover, we identify directionality between the time series of NoNGF and network traffic as a leader-follower relationship [8]. It indicates NoNGF as the leader initiates a fluctuation which is repeated by the network traffic as the follower, which proves that we can predict the future fluctuation and burst of network traffic by NoNGF at the current moment.

We experimentally verify that the NoNGF series has a considerable advantage for accurately predicting the highly random network traffic and traffic burst by using various prediction models that support covariate prediction such as LSTM and TCN [9, 10]. The covariate prediction method we proposed that inputs both the NoNGF series and the network traffic series has a vast improvement over the prediction method that only inputs the traffic series. Our proposed method significantly improves not only the accuracy of overall network traffic prediction but also the prediction of network traffic bursts, which is much more precise in predicting traffic burst peaks and troughs.

Section II of this paper describes the related work of network traffic prediction methods and introduces some time series prediction models used in the experiments. Section III introduces our work on network flow features, including the detail of the used network flow dataset, the definition of the number of newly-generated network flows, and the analysis of interrelationships between the NoNGF and traffic series. Section IV presents our prediction experiment design and

ISBN 978-3-903176-48-5

comparative experiment results, including the prediction error analysis and image comparison. Our proposed method has better prediction accuracy and performs excellently during network bursts. Section V concludes the whole paper and delivers the outlook of the following work.

II. Related Work

Network traffic prediction models can override or optimize classical time series prediction models by considering network traffic data as a time series. In recent years, many network traffic prediction models and methods have been proposed, mainly including machine learning and deep learning [11]. For predicting traffic bursts, some feature extraction methods and prediction models have been proposed. These will be introduced in this section.

Traditional methods mainly include linear regression methods such as autoregressive integrated moving average (ARIMA) models and nonlinear regression methods such as support vector machine (SVM), which can predict the network traffic in the following time [12, 13].

Network traffic prediction by deep learning is a current research hotspot. Recurrent neural network (RNN) is a deep learning model consisting of recurrent neurons, which recursively loop through time steps to learn the features of the previous series and pass them to the next neuron [14]. Long short-term memory (LSTM) neural network is a variation of RNN widely used in time series prediction. LSTM constructs input gate, output gate, and forget gate inside neurons to remember effective features and selectively forget ineffective features, which can solve the gradient vanishing and explosion problems during training of long series [15]. LSTM model can effectively learn the periodic features and stable trends of network traffic time series, and performs well in predicting stationary traffic series. Convolutional neural network (CNN) is a classical deep learning model, which effectively extracts features and reduces computation through convolution and pooling operations, which is advantageous in dealing with temporal-spatial composite scenarios [16]. Temporal Convolutional Network (TCN) obtains historical information through causal convolution, makes the receptive field more flexible through inflation convolution, and solves gradient vanishing problem through residual connections, that can achieve or even surpass the effect of RNN models in time series prediction [17].

For network traffic prediction, peculiar network traffic features need to be mined and extracted by feature engineering to improve the prediction accuracy. [18] decompose the network traffic time series into several feature series by wavelet transform to separate the burstiness, periodicity and non-stationary, then put them together in an LSTM model for prediction. [19] divide the network traffic burst into several scenarios to describe the network traffic burst process and achieve better performance and higher accuracy.

Covariate assisted prediction is a method of time series prediction. When predict a series with poor autocorrelation, low stationarity, and high randomness, it is easy to have no predictive effect. In this situation, it is necessary to find suitable external feature variables as covariates to explain the internal mechanism of time series variation [7]. In the field of network traffic prediction, researchers mainly focus on the network traffic time series itself to do feature extraction and model optimization to improve the prediction accuracy. [20] use network traffic series from several regions as covariates to learn the spatial features of traffic but still do not involve external variables. Barely analysis or research to find external characteristic variables of network traffic for covariate prediction has emerged.

III. Correlation Analysis between Network Traffic and Flow Features

Network flow and network traffic are two sides of the same coin. The establishment of network flow indicates that there will be data packets transmitted between communicating parties, which will inevitably generate network traffic in the following period. Therefore, feature variables of network flows can be used as covariates for network traffic prediction to explain the intrinsic mechanism and generation cause of network traffic changes or bursts. It is necessary to do some feature extraction and analysis on network flows. For this reason, we extracted an effective network flow feature, NoNGF. In addition, we proved that it can predict the fluctuation and variation of network traffic in advance, and can significantly help us predict network traffic burst.

A. Network Traffic and Flow Data

The dataset we use is network traffic data and network flow data for a city of over 3 million people in China. This dataset records data for a full month (31 days) of July 2020. This dataset is collected by deploying optical splitter and deep packet inspection (DPI) server on the metro core network And of course, the data has been desensitized.

Network traffic data is a time series about the size of network traffic volume (in KB). Network flow data is a detailed description of the network flows established in the city during the month, including desensitized IP address, port number, flow start time, flow end time, and network traffic volume. The network flow dataset is huge with 3740GB for just one month's duration because it records log information of each flow. This dataset also labels the traffic with application categories so that we can do more detailed analysis and prediction on the application-level network traffic.

The dataset is divided into more than 20 application categories. Each application category is subdivided into specific applications and protocols. Since some application categories were created early and the protocols used are old, there is almost no traffic for these applications. We filtered out three representative applications: WEB video application, chat tool application, and cloud drive application.

The network traffic of WEB video applications has the largest traffic volume, about half of the whole network traffic. The network traffic of the chat tools has high-frequency bursts that can increase several times in minutes, which is the main research object on traffic burst prediction of this work. The traffic of cloud drive applications has the lowest correlation between the NoNGF and the traffic series. We use the network traffic of these three applications and the whole network traffic as the experiment data for analysis and comparison. In fact, we did experiments for all applications' traffic and in general the experimental results are consistent with the results obtained from the filtered applications' traffic.

Based on the metadata of each network flow, we find an effective feature of network flows by counting the number of new network flows generated per unit time. The start time of a flow falls in the unit time, and the number of newly-generated network flows of this unit time is added by one, thus obtaining a time series.

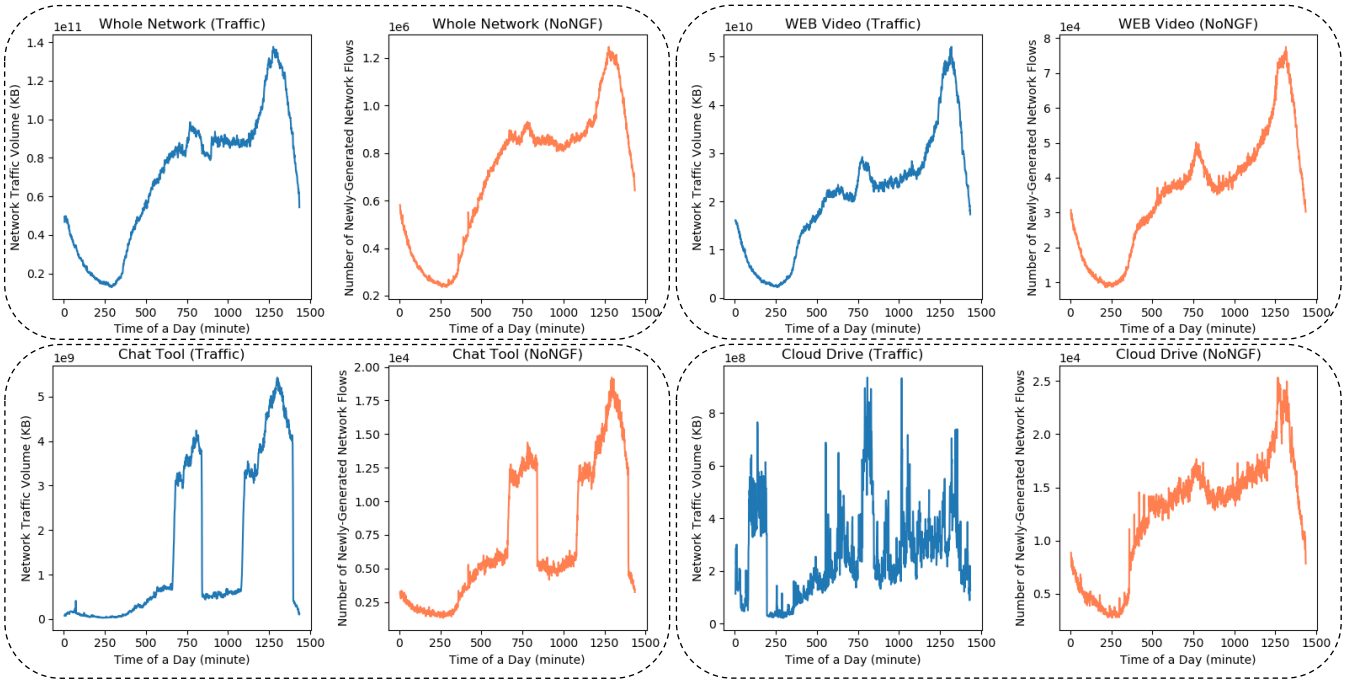

Fig. 1. One-day time series of network traffic volume and the number of newly-generated network flows for the whole network, WEB video, chat tool and cloud drive application

Fig. 1 shows the time series of network traffic volume and the time series of the NoNGF for a particular day (1440 minutes). The blue curve is the network traffic time series, and the corresponding vertical axis indicates the traffic volume (size in KB). The orange curve is the time series of the NoNGF, and the corresponding vertical axis presents the number of newly-generated network flows. We can intuitively conclude that there is some correlation between the network traffic volume and NoNGF. Especially for the whole network, web video and chat tool network scenarios, the fluctuations, troughs, peaks, and bursts of both time series are almost simultaneously. However, the traffic of cloud drive does not lead to the similar conclusion. As shown in Fig. 1, in the whole network, WEB video, Chat Tools network scenarios, the images of the traffic series and NoNGF series are highly similar and have some correlation, implying interrelationships between these two time series, so it is necessary to conduct further analysis and research.

B. Correlation Analysis

In order to determine the relationship between the network traffic and NoNGF, it is necessary to quantitatively analyze the cross-correlation to investigate whether the NoNGF can reveal the fluctuation and burst of network traffic in advance. Thus, we use cross-correlation analysis to lay a theoretical foundation that NoNGF series can assist network traffic prediction.

Cross-correlation is a similarity measurement for two time series, widely used in signal processing. It can determine the shape similarity of two time series well by considering their amplitude and phase fluctuations [21]. Cross-correlation tracks the movements of two or more sets of time series data relative to one another [22]. It compares multiple time series and objectively determines how closely they match and when the best match occurs.

This work adopts three methods to analyze the cross-correlation of two time series: scatter plot, cross-correlation coefficient, and time-lagged cross-correlation. The Scatter plot is used to determine whether the two series have a linear relationship [23]; Cross-correlation coefficient is used to quantify the magnitude of the correlation between the two time series; Time-lagged cross-correlation method can derive whether one time series fluctuates due to the fluctuation of the other time series and can determine the lag relationship between the two time series [24].

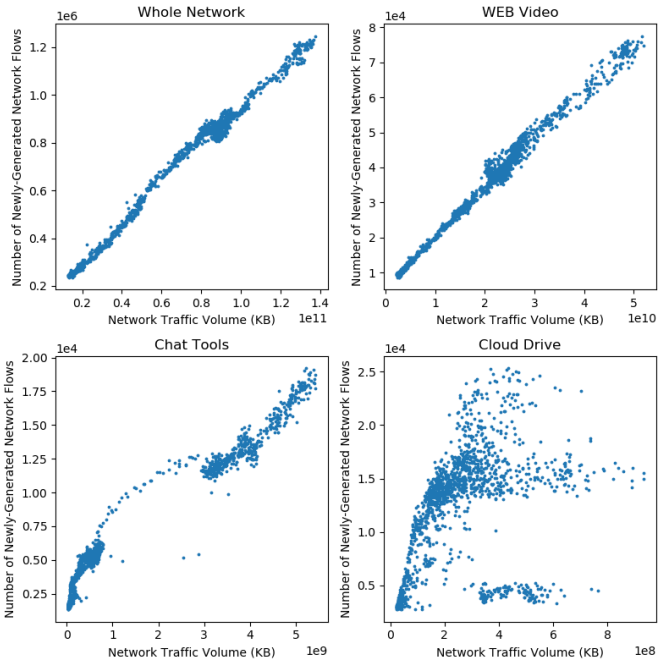

Fig. 2. The scatter plots of the network traffic volume series and the number of newly-generated network flows series of the whole network, WEB video, chat tool and cloud drive application

As shown in Fig.2, we use the scatter plot method to verify whether there is a linear correlation between the network traffic series and the NoNGF series. The scatter plots of the whole network, WEB video, and chat tool application can be fitted as a straight line, proving that the traffic volume and the

NoNGF series have a linear correlation. The scatter plot of cloud drive does not show a linear relationship.

The time series of network traffic for a day is defined as a vector $T_d = [t_1, t_2, \cdots, t_{1440}]$, where t_i denotes the network traffic volume at minute i of each day. The time series of the NoNGF for a day is defined as a vector $N_d = [n_1, n_2, \cdots, n_{1440}]$, where n_i denotes the number of newly-generated flows at minute i of each day.

We obtain the vectorial inner product of T_d and N_d:

$$R(T_d, N_d) = \sum_{i=1}^{1440} T_i \times N_i \quad (1)$$

Then we calculate their cross-correlation coefficient (CC):

$$CC(T_d, N_d) = \frac{R(T_d, N_d)}{\sqrt{R(T_d, T_d) \times R(N_d, N_d)}} \quad (2)$$

The cross-correlation coefficient ranges from 0 to 1, and the closer it is to 1, the stronger correlation is. The calculated cross-correlation coefficients for the particular day are shown in Table I. We compare Fig. 1, Fig. 2 with Table I and come to the expected conclusion. For the whole network, web video and chat tool application, the cross-correlation coefficient of the traffic and the NoNGF is very high, matching what the figures show. While the cross-correlation coefficient for cloud drive application is low.

TABLE I. THE CROSS-CORRELATION COEFFICIENTS OF THE NETWORK TRAFFIC AND THE NONGF

Category of traffics	Cross-correlation coefficient of the traffic and the NoNGF
The whole network	0.9880
WEB video	0.9842
Chat tool	0.9693
Cloud drive	0.8859

C. *Time-lagged cross-correlation (TLCC)*

We used the time-lagged cross-correlation (TLCC) method to verify the lagged relationship between NoNGF series and the traffic series, which can determine whether one series affects the other series movement and direction [25]. TLCC is measured by incrementally shifting one time series vector and repeatedly calculating the correlation between two signals, which can identify directionality between two time series, such as a leader-follower relationship in which the leader initiates a response that is repeated by the follower [26]. The peak correlation value indicates that the two time series are most synchronized at that time. If one time series vector leads the other vector, the peak correlation will not be at the center (offset 0). The offset coefficient where the peak correlation is located indicates how much time ago one time series influenced the other time series.

We mentioned earlier that the generation of the network flows affects network traffic volume in the next few minutes with the TLCC verification. We extract the time series of network traffic and the time series of NoNGF with the length of one day (1440 minutes).

We capture 1380 minutes in vector N_d and denote it as $N_j = [n_j, n_{j+1}, \cdots, n_{j+1379}]$ $j \in [1,60]$, where j takes an integer in the range of 0 to 60. This represents the NoNGF vector by forth and back of total 60-minute shifting.

We capture the 31st minute to the 1410th minute of the network traffic series T_d, a total of 1380 minutes, as a fixed traffic vector, denoted as $T_{30} = [t_{30}, t_{31}, \cdots, n_{1409}]$.

With $j = 30$ as the central origin, the time-lagged cross-correlation coefficient is calculated after each translation moving of the NoNGF series, defined as $TLCC_k$, and the value of k is an integer in the range of -30 to 30:

$$TLCC_k = CC(N_{k+30}, T_{30}) \quad k \in [-30,30] \quad (3)$$

The maximum value of TLCC in shifting is the TLCC Peak, noted as TP, and the offset at the peak point is called Peak Offset, denoted as PO:

$$TP = \max(TLCC_k) \quad (4)$$

$$PO = arg\max(TLCC_k) \quad (5)$$

According to TLCC, we fix the network traffic time series and calculate the cross-correlation coefficient after the forth and back translation shifting of NoNGF time series. Fig. 3 shows the TLCC values of each offset for each network category traffic with the peak offset highlighting. The vertical axis indicates the cross-correlation coefficient; The horizontal axis indicates the forth and back time-shift; The black dashed line indicates the original cross-correlation coefficient where the offset is 0; The red dashed line indicates the offset with the peak cross-correlation coefficient.

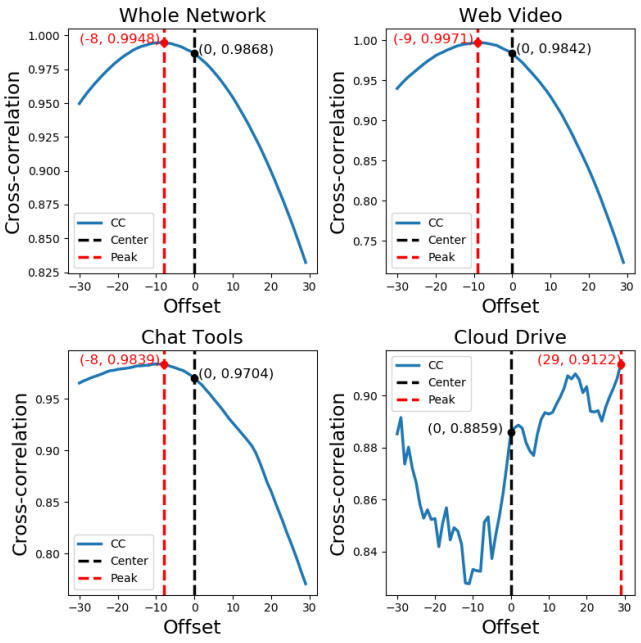

Fig. 3. TLCC of network traffic volume and the number of newly-generated network flows for each network application traffic

The curves of the whole network, web video, and chat tool are similar with smooth curve, reaching the cross-correlation peak after a few minutes of lag, and the peak CC is extremely high (TLCC Peak > 0.98). For example, for web video applications, the cross-correlation coefficient between NoNGF and the network traffic peaks when the time series of NoNGF is shifted forward by 9 minutes, which means the change in NoNGF affects the fluctuation of the network traffic volume after 9 minutes. It proves that for the whole network,

WEB videos and chat tool application, the NoNGF dominates the trend of network traffic fluctuations after a few minutes.

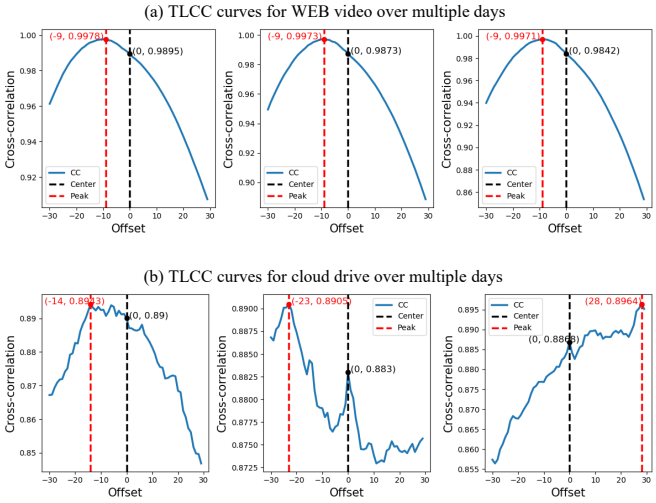

Fig. 4. TLCC of network traffic volume and the number of newly-generated network flows for web video and cloud drive over multiple days

Fig. 4 (a) shows the TLCC of WEB video over multiple days, and it is easy to find that they are all smooth and similar to each other, with peak offset around -9 minutes and extremely high peak cross-correlation coefficients. It turns out that the TLCC image with the smooth curve and a fixed peak offset is not a coincidence. When the TLCC peak is high enough, the same result always occurs. Therefore, for some network scenarios, NoNGF does affect the fluctuation of network traffic volume after a few minutes and is universally applicable.

The TLCC curve of cloud drive is the opposite of the other network application traffic mentioned above. Fig. 4 (b) shows the TLCC of cloud drive data for multiple days, which shows no pattern with low cross-correlation coefficient, and none of them are similar. There is little correlation between the NoNGF and network traffic for cloud drive. Our analysis yields that, in the download application, the generation of network flows has less influence on the network traffic volume. The network bandwidth or the downloaded file's size has more impact on the network traffic volume.

TABLE II. AVERAGE TLCC OF THE NONGF AND TRAFFIC SERIES RESULTS FOR THE WHOLE NETWORK AND THREE APPLICATION-LEVEL NETWORK TRAFFIC

	Whole Network	**WEB Video**	**Chat Tool**	**Cloud Drive**
Avg CC	0.988	0.984	0.969	0.886
Avg TLCC Peak	0.993	0.998	0.987	unstable
Avg TLCC Offset	-8	-9	-8	unstable

Table II shows the TLCC results of NoNGF and traffic series for the whole network traffic and three application-level network traffic.

The formula for defining each data item in the table is as follows:

$$Avg\_CC = \frac{1}{31}\sum_{\text{day}=1}^{31}\left(CC(T_{day}, N_{day})\right) \quad (6)$$

$$Avg\_TLCC\_Peak = \frac{1}{31}\sum_{\text{day}=1}^{31}\left(TP_{day}\right) \quad (7)$$

$$Avg\_TLCC\_Offset = \frac{1}{31}\sum_{\text{day}=1}^{31}\left(PO_{day}\right) \quad (8)$$

The subscript day indicates the data of date day. Since the data set has a total of 31 days, the maximum value of day is 31.

Consistent with the results shown in Fig. 3, Table II shows that these two time series of the whole network, WEB video, and chat tools have a high correlation with each other, whose TLCC peak and offset values maintain stable performance. The TLCC performance of these two series for cloud drive applications is not stable, and the correlation between NoNGF and traffic series in these networks is not apparent.

Through TLCC analysis, we conclude that the feature series of NoNGF has a few minutes of advance predictive for network traffic volume series in the network scenarios of the whole network, chat tools and WEB video application. It indicates that the variation direction and amplitude of the network traffic series are highly consistent with the fluctuation of NoNGF series a few minutes earlier. Based on this finding, we can perceive the trend and burst of traffic volume several minutes in advance by NoNGF series, which provides the theoretical support that the feature of newly-generated network flows can improve network traffic prediction effectively.

IV. PREDICTION EXPERIMENT AND RESULT ANALYSIS

We derived the correlation between the network traffic volume and NoNGF by data analysis mining. In order to verify that the feature of NoNGF can indeed improve the accuracy of network traffic prediction, we build prediction models and take NoNGF series and network traffic volume series together as input for training and prediction, to compare with the prediction model only inputting traffic volume series.

A. Design of experiments

The most fundamental difference between the comparison experiments is the input features, which are divided into one-dimensional data with only network traffic time series and two-dimensional data that include a combination of NoNGF time series and network traffic time series. A deep-learning model with only one-dimensional data input is called a 1D-model. A model with two-dimensional data input is called a 2D-model, which is the network traffic prediction method based on NoNGF our proposed.

Meanwhile, we built two prediction models, LSTM and TCN, which are described in related work, shown in Fig. 5. The reason for using two models with quite different principles is to demonstrate the universality of the feature NoNGF we proposed. Prediction models can be divided into two groups, 1D-LSTM/2D-LSTM, and 1D-TCN/2D-TCN. Each group of models has the same hyperparameters for training and prediction. Only the input data dimensions are different. Table III shows the main parameters of these models.

First we performed single-step prediction experiments, outputting the predicted values for only one time-step. For a more comprehensive evaluation, we also designed multi-step

prediction experiments using the LSTM model with input of 60 time-steps to predict the next 10 time-steps single shot [27].

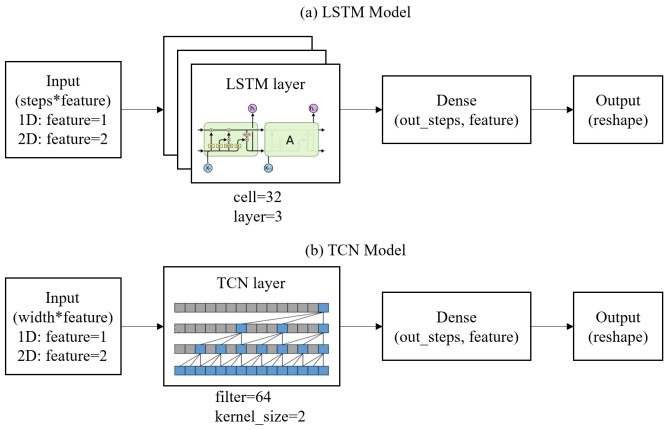

Fig. 5. LSTM and TCN model architectures

TABLE III. THE MAIN PARAMETERS OF LSTM AND TCN MODELS

Model	Main Parameters
LSTM	hidden_size=32; num_layers=3; input_steps=30; batch_size=32
TCN	nb_filter=64; kernel_size=3; dilations=[1,2,4,8]; input_width=30; batch_size=32;
Multi-Step LSTM	hidden_size=32; num_layers=3; input_steps=40; output_steps=20; batch_size=32

The dataset we use is the dataset introduced in Section 3.1, including the whole network, web video applications, chat tools, and cloud drive. The network traffic time series and NoNGF time series is 31 days long and 10 minutes time granularity. The first 70% of them are the training set, followed by 20% of the training set are the validation set, and the last 10% are the test set.

We use root mean squared error (RMSE) as metrics, where predicted values denote as y_i, true values denote as $\hat{y}_i$, and the sample size denotes as m, defined as:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{9}$$

B. *Single-step prediction*

We made single-step prediction on LSTM and TCN models. Table IV shows the traffic prediction error RMSE for the different network traffic categories and provides the average peak cross-correlation. Table IV also shows the two-dimensional input model accuracy improving rates (AIR) comparing with one-dimensional model for each group of models, defined as:

$$AIR = \frac{RMSE_{1D} - RMSE_{2D}}{RMSE_{1D}} \tag{10}$$

For traffic categories with a high TLCC peak (Avg TLCC Peak > 0.98) such as the whole network, web video and chat tool application, 2D input models with NoNGF to forecast traffic volume perform significantly more accurate and less RMSE error than 1D input models. In the whole network scenario, the 2D-LSTM reduces prediction RMSE error by 10.41% compared to the 1D-LSTM and the TCN error by 8.81%. In some single application traffic scenarios, such as Chat Tools, the improvement of each group of models is more than 10%, with LSTM being the most significant growth of 15.97%. The network traffic of Chat Tools has more bursts than other application traffic.

TABLE IV. COMPARISON OF 1D MODEL AND 2D MODEL RMSE FOR EACH NETWORK TRAFFIC CATEGORY

Network traffic category	The Whole Network	WEB Video	Chat Tool	Cloud Drive
Avg TLCC Peak	0.9932	0.9984	0.9873	0.8911
1D-LSTM RMSE	21.6875	9.5496	3.6345	0.9392
2D-LSTM RMSE	19.4301	8.6127	3.0542	0.9593
LSTM AIR	**10.41%**	**9.81%**	**15.97%**	-3.44%
1D-TCN RMSE	21.5526	9.6623	3.5913	0.9143
2D-TCN RMSE	19.6559	8.6007	3.2532	0.9393
TCN AIR	**8.81%**	**10.99%**	**9.42%**	-2.73%

The fact that NoNGF can better predict the burst condition is the main reason for the considerable accuracy improvement of 2D-input models. It reflects the effectiveness and robustness of our proposed feature series, NoNGF. In network traffic prediction for application categories with low TLCC peaks, such as Cloud drive, 2D-input models have no accuracy improvement over 1D-input even slightly decrease. The traffic of cloud drive accounts for 0.7% of the whole network traffic, so that it has minimal impact on the network traffic prediction.

The prediction experiments prove that it is essential to do TLCC for different categories between network traffic and the NoNGF before prediction. As the TLCC results show, for the application traffic categories with stable TLCC results, which means the peak offsets are almost fixed and the average peak CC is greater than 0.98, network traffic prediction RMSE errors of 2D-input models are much lower compared to the 1D-input. For the application traffic categories with insensitive TLCC performance, the traffic prediction effect of 2D-input models will not improve.

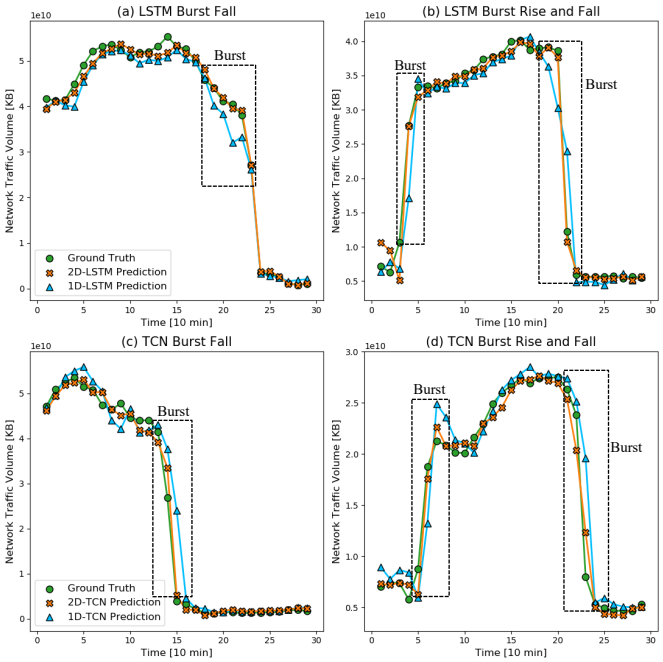

Fig. 6. Comparison of 1D/2D-LSTM and 1D/2D-TCN prediction result on network traffic burst prediction for Chat Tools application

Because the traffic of chat tools has more bursts, it can better show the advantage and effectiveness of network burst prediction based on the feature of newly-generated network flows. Fig. 6 shows the network traffic prediction results for the chat tool application traffic. The green dots are the ground truth, the orange crosses are the predicted values of 2D-Models, and the blue triangles are the predicted values of 1D-Models.

As shown in Fig. 6 (a) and (b), it is clear that the 2D-LSTM with the NoNGF series as input has better performance, and the predicted values are closer to the ground truth with less error compared to 1D-LSTM. Observing the traffic burst moments in the dashed box, the 1D-model prediction results change inaccurately and slowly during bursts. In the worst case, the 1D-models perform non-robust prediction. The predicted value is very close to the previous true value. As shown in Fig. 6 (c) and (d), the predicted values of 1D-TCN lag the ground truth by exactly one time-step. This is the most common and most difficult problem for traffic burst prediction, but our approach effectively solves this problem. The prediction results of 2D-models we proposed rise or fall accurately during traffic bursts, even almost overlap with the ground truth, without any advance or lag. In conclusion, network traffic prediction based on the feature of newly-generated network flows can significantly improve accuracy in the case of network traffic bursts.

C. *Multi-step prediction*

We also built the LSTM single-shot multi-step prediction model and compared multi-step traffic prediction between 1D-input and 2D-input, with inputting 60 time-steps and single-shot predicting 10 time-steps. Table V shows the multi-steps prediction RMSE errors of 1D-LSTM and 2D-LSTM for each network traffic category, and the improving rate of 2D-LSTM compared with 1D-LSTM. We can see that the RMSE of multi-step prediction is slightly higher than that of single-step prediction, and the accuracy improvement obtained with our method is greater. 2D-LSTM prediction has 20.41% less RMSE error than 1D-LSTM for the whole network and the marvelous 32.53% less RMSE error for chat tools. We can conclude that the multi-step prediction based on the feature of newly-generated network flows is still valid and can significantly improve the accuracy of predicting the network traffic for categories with high TLCC peaks.

TABLE V. MULTI-STEP PREDICTION RMSE ERRORS OF 1D-LSTM AND 2D-LSTM FOR EACH TRAFFIC CATEGORY

Network traffic category	The Whole Network	WEB Video	Chat Tool	Cloud Drive
Avg TLCC Peak	0.9932	0.9984	0.9873	0.8911
1D-LSTM RMSE	36.6113	16.7462	5.2715	1.0109
2D-LSTM RMSE	29.1398	14.8380	3.5564	0.9795
Multi-Step AIR	**20.41**%	**11.39**%	**32.53**%	3.11%

Fig. 7 shows the multi-step prediction results for chat tools. The black dashed line indicates the starting point of multi-step prediction, the green dots are the ground truth, the blue triangles are the 1D-LSTM prediction results, and the orange crosses are the 2D-LSTM prediction results. We can see that the prediction result of 2D-LSTM is much better than that of 1D-LSTM and the gap in curve fitting is noticeable. The prediction results of 2D-LSTM are closer to the true values when bursts occur, while those of 1D-LSTM can only predict general trends. The 1D-LSTM prediction lags during the traffic burst rise and the burst peak value of prediction is much lower than the true burst peak. In the case of traffic burst fall, the 1D-model predicted values drop early and slowly. However, the 2D-LSTM predicted values rise or fall accurately and abruptly within one time step, and the amplitude of variation is consistent with the true traffic burst. In the burst peak period, the predicted values of the 2D-model are much closer to the ground truth than the 1D-model. Network traffic prediction based on the feature of newly-generated network flows works better for the burst and fluctuation of the network traffic.

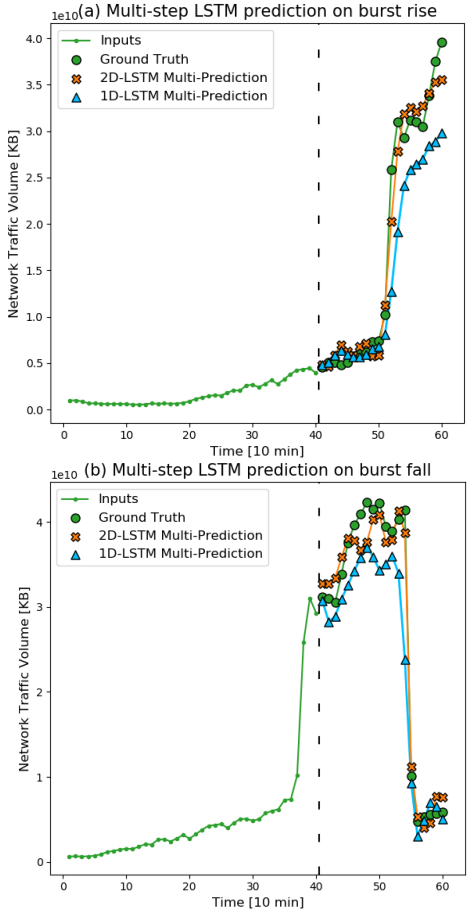

Fig. 7. Multi-step prediction results for chat tools network traffic

In addition, we counted the RMSE error per prediction step in the multi-step prediction under the chat tool traffic. As shown in Table VI, the RMSE error of the fifth time-step in 2D-LSTM predicting is lower than that of the first time-step in 1D-LSTM predicting.

TABLE VI. COMPARISON OF 1D-LSTM AND 2D-LSTM PER SINGLE STEP RMSE FOR MULTI-STEP PREDICTION IN CHAT TOOL APPLICATION NETWORKS TRAFFIC

Time Step	1	2	3	4	5	6
1D-LSTM RMSE	**3.87**	4.30	4.52	4.75	4.86	4.94
2D-LSTM RMSE	**3.12**	**3.44**	**3.59**	**3.74**	**3.80**	4.01

From the comparison we infer that the network traffic prediction based on NoNGF is much more precise than the original method in predicting longer future network traffic for multi-step predictions and able to anticipate network burst traffic earlier. Our traffic prediction method accurately detects network emergencies earlier and allows for more timely notification to the network control system.

V. Conclusion

In this paper, we propose a novel approach to network traffic prediction based on the feature of newly-generated network flows. We provide an inspiring idea of network traffic prediction by mining external features of traffic from network flow data as covariates to assist in traffic prediction.

We mine an effective feature series, the number of newly-generated network flows (NoNGF), from the network flow data. We demonstrate that the network traffic time series and NoNGF time series are strongly correlated in the whole network traffic and most of the application-level network traffic. Through TLCC analysis, we prove that we can anticipate the trend and burst of network traffic several minutes in advance by NoNGF. Then we build LSTM and TCN prediction models, and perform prediction experiments with multiple inputs and settings for the whole network and three application network traffic. Our traffic prediction method significantly improves prediction accuracy, with RMSE reduction of more than 10% on average. For some application network traffic with frequently burst such as chat tool, our multi-step prediction RMSE error is reduced by more than 30%. It is worth mentioning that our prediction method performs much better for burst traffic prediction with nearly perfect accuracy.

This work is just to throw light on a different way of thinking about network traffic prediction. Network traffic prediction accuracy can be improved not only by optimizing prediction models, but also by mining the relevant external features of network flows as covariates to assist network traffic prediction. It may be possible to explore other network flow features that can also effectively assist in network traffic prediction. In future work, we will customize advanced models and techniques, such as attention mechanism and transformer [28, 29], to effectively inject more network flow features into the prediction model to achieve better results.

VI. Acknowledgments

This work is supported by the National Nature Science Foundation of China (Grant No.U1909204) and Youth Innovation Promotion Association of Chinese Academy of Sciences (2021168).

References

[1] Statista R D. Internet of Things - Number of connected devices worldwide 2015 - 2025[J]. Statista Research Department, 2019.

[2] Cisco, Cisco visual networking index: forecast and trends, 2017-2022, 2018.

[3] Xie J, Yu F R, Huang T, et al. A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges[J]. IEEE Communications Surveys & Tutorials, 2018, 21(1): 393-430.

[4] Donevski I, Vallero G, Marsan M A. Neural networks for cellular base station switching[C]//IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2019: 738-743.

[5] Kang M, Song J, et al. Survey of Network Traffic Forecast Based on Deep Learning[J]. Computer Engineering and Applications, 2021, 57(10): 1-9.

[6] Xu F, Lin Y, Huang J, et al. Big data driven mobile traffic understanding and forecasting: A time series approach[J]. IEEE transactions on services computing, 2016, 9(5): 796-805..

[7] Aue A, Norinho D D, Hörmann S. On the prediction of stationary functional time series[J]. Journal of the American Statistical Association, 2015, 110(509): 378-392.

[8] Cheong, J. H. (2020, December 8). Four ways to quantify synchrony between time series data. https://doi.org/10.17605/OSF.IO/BA3NY

[9] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[10] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv:1803.01271, 2018.

[11] Oliveira T P, Barbar J S, Soares A S. Computer network traffic prediction: a comparison between traditional and deep learning neural networks[J]. International Journal of Big Data Intelligence, 2016, 3(1): 28-37.

[12] Moayedi H Z, Masnadi-Shirazi M A. Arima model for network traffic prediction and anomaly detection[C]//2008 International Symposium on Information Technology. IEEE, 2008, 4: 1-6.

[13] Nikravesh A Y, Ajila S A, Lung C H, et al. Mobile network traffic prediction using MLP, MLPWD, and SVM[C]//2016 IEEE International Congress on Big Data (BigData Congress). IEEE, 2016: 402-409.

[14] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[15] Ramakrishnan N, Soni T. Network traffic prediction using recurrent neural networks[C]//2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018: 187-193.

[16] Jin X, Yu X, Wang X, et al. Prediction for Time Series with CNN and LSTM[C]//Proceedings of the 11th International Conference on Modelling, Identification and Control (ICMIC2019). Springer, Singapore, 2020: 631-641.

[17] Bi J, Zhang X, Yuan H, et al. A hybrid prediction method for realistic network traffic with temporal convolutional network and lstm[J]. IEEE Transactions on Automation Science and Engineering, 2021.

[18] Lu H, Yang F. A network traffic prediction model based on wavelet transformation and lstm network[C]//2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2018: 1-4.

[19] Vinchoff C, Chung N, Gordon T, et al. Traffic prediction in optical networks using graph convolutional generative adversarial networks[C]//2020 22nd International Conference on Transparent Optical Networks (ICTON). IEEE, 2020: 1-4.

[20] He K, Chen X, Wu Q, et al. Graph attention spatial-temporal network with collaborative global-local learning for citywide mobile traffic prediction[J]. IEEE Transactions on mobile computing, 2020.

[21] Su Y, Zhao Y, Xia W, et al. CoFlux: robustly correlating KPIs by fluctuations for service troubleshooting[C]//Proceedings of the International Symposium on Quality of Service. 2019: 1-10.

[22] Derrick T, Thomas J. Time series analysis: the cross-correlation function[J]. 2004.

[23] Mindrila D, Balentyne P. Scatterplots and correlation[J]. Retrieved from, 2017.

[24] Curriero F C, Shone S M, Glass G E. Cross correlation maps: a tool for visualizing and modeling time lagged associations[J]. Vector-Borne & Zoonotic Diseases, 2005, 5(3): 267-275.

[25] Boker S M, Rotondo J L, Xu M, et al. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series[J]. Psychological methods, 2002, 7(3): 338.

[26] Wang F, Wang L, Chen Y. Detecting PM2. 5's correlations between neighboring cities using a time-lagged cross-correlation coefficient[J]. Scientific reports, 2017, 7(1): 1-11.

[27] Wang Y, Zhu S, Li C. Research on multistep time series prediction based on LSTM[C]//2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE). IEEE, 2019: 1155-1159.

[28] Li M, Wang Y, Wang Z, et al. A deep learning method based on an attention mechanism for wireless network traffic prediction[J]. Ad Hoc Networks, 2020, 107: 102258.

[29] Lim B, Arik S O, Loeff N, et al. Temporal fusion transformers for interpretable multi-horizon time series prediction[J]. arXiv preprint arXiv:1912.09363, 2019.