Integrating Granger Causality and Vector Auto-Regression for Traffic Prediction of Large-Scale WLANs

Zheng Lu[1], Chen Zhou[1], Jing Wu[1,2], Hao Jiang[1,2], Songyue Cui[1]
[1] School of Electronic Information, Wuhan University
Wuhan, 430072, China
[2] Collaborative Innovation Center for Geospatial Technology
Wuhan, 430079, China
[e-mail: wujing@whu.edu.cn]
*Corresponding author: Jing Wu

Received April 11, 2015; revised July 14, 2015; accepted November 29, 2015; published January 31, 2016

Abstract

Flexible large-scale WLANs are now widely deployed in crowded and highly mobile places such as campus, airport, shopping mall and company etc. But network management is hard for large-scale WLANs due to highly uneven interference and throughput among links. So the traffic is difficult to predict accurately. In the paper, through analysis of traffic in two real large-scale WLANs, Granger Causality is found in both scenarios. In combination with information entropy, it shows that the traffic prediction of target AP considering Granger Causality can be more predictable than that utilizing target AP alone, or that of considering irrelevant APs. So We develops new method - **G**ranger **C**ausality and **V**ector **A**uto-**R**egression (GCVAR), which takes APs series sharing Granger Causality based on Vector Auto-regression (VAR) into account, to predict the traffic flow in two real scenarios, thus redundant and noise introduced by multivariate time series could be removed. Experiments show that GCVAR is much more effective compared to that of traditional univariate time series (e.g. ARIMA, WARIMA). In particular, GCVAR consumes two orders of magnitude less than that caused by ARIMA/WARIMA.

Keywords: Wireless local area network, traffic prediction, Granger Causality, Vector auto-regression

This research is supported by National Natural Science Foundation of China(NSFC) under Grant No. 61371126,Fundamental Research Funds for the Central Universities under Grants Nos. 2042014gf020 and 2042014kf0256, National High Technology Research and Development Program of China (863 Program) under Grant No. 2014AA01A707, national basic research program of china(973 Program) Grant No. 2011CB707106, Satellite application research innovation fund projects of china aerospace science and technology corporation(CASC) under Grant No. 2014_CXJJ-TX_15, and the Natural Science Foundation of Hubei province under Grants Nos. 2014CFB716 and 2011CDA042.

1. Introduction

With the popularization of the mobile internet and the intelligent terminal, the mobile data traffic experience explosive growth. Large-scale WLANs are flexible as terminal bypass for Internet traffic and widely deployed in crowded and highly mobile places such as campus, airport, shopping mall and company etc. Nowadays, the number of AP in a medium-sized city may reach 20000. Meanwhile, network congestion due to huge traffic flow, collapses under the attack of malicious traffic makes the management of large-scale WLAN urgent and different to traditional WLANs, especailly the traffic prediction.

Traffic prediction is of great significance in network management [1], and it is usually performed through time series models [2]. Traditional univariate time series models, such as Auto-Regression (AR), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) [3], only consider the historical data of the target serie alone. Recently, wavelet transform expands the methods for traffic forecasting [4-7]. But the research of traffic prediction for large-scale WLAN is not kept up with the engineering construction. In a large-scale WLAN containing many APs, the power and channel are usually adjusted for optimization [8], so not all stations can hear each other as traditional WLAN. In this situation, throughput distributions among links may be highly uneven and severe unfairness can result under the standard greedy algorithm in medium access in which all stations try to grab as much bandwidth as possible from the network. So the relationship of APs in large-scale WLANs should be carefully considered in traffic prediction. But above models ignore the information provided by un-targeted series which may improve the prediction. Multivariate time series models, such as Vector Auto-Regression (VAR), Support Vector Machine (SVM) utilize statistical dependence of multiple time series for traffic prediction. However, irrelevant series will introduce redundant and noise information for prediction. But the impact could be relaxed through information entropy which is demonstrated thereinafter.

Granger Causality was first introduced by C. W. J. Granger [9], and it is widely used in statistics, economics, neural networks, etc. Paul *et al.* [10] brought the Granger Causality into the 3G mobile networks for the first time. However, Wireless Local Area Networks (WLANs) shows new features compared to cellular systems. On one hand, the coverage of a AP is much smaller than that of a base station (BS), and overlapping areas are quite common in WLANs. Thus the handoff in WLANs is much more frequent than that in cellular systems, and leading to traffic alteration in adjacent APs. On the other hand, in some scenarios, such as airport or shopping mall, the movement of users may be well and unconsciously organized due to their scheduling or habits. That may lead to the correlation between APs, which could benefit traffic prediction.

In the paper, we utilize the correlation between APs to predict network traffic based on multivariate time series. We find that Granger Causality is universal in two typical scenarios, which accounts for at least 85% of total APs. In addition, based on information entropy, we shows that Granger Causality makes the traffic prediction of target AP more accurate, compared to considering historical traffic of target AP alone, or taking irrelevant APs into account. So we propose the method integrating **G**ranger **C**ausality and **V**ector **A**uto-**R**egression (GCVAR) for traffic prediction in the paper. Simulation results show that GCVAR is more accurate and precise compared to univariate time series models such as

ARIMA/WARIMA. Moreover, GCVAR is more efficient since it consumes two orders of magnitude less than that caused by ARIMA/WARIMA.

The main contributions of this paper are:

1) Based on the traffic data of two real large-scale WLANs, we verifies the widespread of Grange Causality in our scenarios (airport: 432 APs, mall: 182 APs). To the best of our knowledge, no other researchers have provided as convincing proof as we did to reveal the Grange Causality in WLANs in such large-scale real networks.
2) Based on the Granger Causality in large-scale WLANs, this paper proposes traffic prediction method integrating **G**ranger **C**ausality and **V**ector **A**uto-**R**egression (GCVAR), to predict the traffic flow in large-scale WLANs. Extensive simulations show that GCVAR is more accurate and precise compared to ARIMA and WARIMA. In particular, GCVAR consumes two orders of magnitude less than that caused by ARIMA/WARIMA.

The rest of this paper is organized as follows: Related work is provided in section 2, the verification of Granger Causality in Large-scale WLAN is introduced in section 3, the details of GCVAR is offered in section 4, experiments are presented in section 5. Finally, section 6 concludes this paper.

2. Related Work

Generally, there are two kinds of models used in time series prediction, namely univariate and multivariate time series models. As for univariate time series models, *Jiang et al*. [2] made a comparison among AR, ARMA and other derived models in different time scale, while *Taylor et al.* [11] made similar comparison on the forecasting of intraday arrivals at a call center. *Chen et al.* [1] made some investigation on the seasonal property of WLAN traffic based on ARIMA model. *Dominguez et al*. [12] combined ARIMA models and wavelet tranform for time series forecasting. *Tan et al.* [13] proposed an aggregation method combining ARIMA model and neutral network model for weekly, daily and hourly traffic flow prediction. *Nakayama et al.* [14] improve ARIMA through a resolution adaptive method to increase accuracy. *Cuaresma et al.* [15] used linear univariate time-series models to predict electricity spot-prices hourly. Those derived algorithms of univariate time series enhanced prediction results, but the available information of other related time series is not considered to improve accuracy and precise.

Different from univariate time series models, multivariate time series models take other related series into account to predict the target series. *Medeiros et al.*[16] proposed a hybrid linear-neural model, which can naturally incorporating the thresholds of linear multivariate. *Holanda et al.*[17] utilized principal components analysis and K-means in traffic prediction. *Feng et al.*[18] applied support vector machine (SVM) to predict WLAN traffic. *Liang et al.* [19] utilized Ant Colony Optimization to obtain the parameter in SVR model for network traffic prediction. *Ahmed et al*.[20] proposed sample entropy to evaluate structural complexity. *Ghosh et al.*[21] introduced structural time-series model to reduce computation complexity for short-term traffic forecasting. *Kocak et al.*[22] combined univariate and multivariate time series for nonlinear time series prediction. Although multivariate time series models above show good completeness, other related time series are considered indiscriminate for target time series prediction, which introduces irrelevant information and noise, as well as unnecessary overhead in computation. Therefore, in this paper Granger Causality is used to enhance the accuracy and computation simpleness for multivariate time serie mode.Granger Causality test is a statistical hypothesis test for determining whether one time series is useful

in forecasting another. Compared with prediction based on entropy theory [23-26] in the latest researches, more direct correlation of different traffic series are considered in the prediction based on Granger causality.

3. Granger Causality in Large-scale WLANs

In this section, rigorous tests based on statistic analysis are provided to verify the existence of Granger Causality in our large-scale WLANs.

3.1 Stationary of WLANs Traffic Series

The stationary of traffic series is the premise for traffic modeling and Granger Causality Test. In particular, stationary is the key factor in selecting a proper order p and estimating the parameters used in traffic model. Un-stationary series should be turned to stationary one through integer or fractional order differential for further processing.

To examine the stationary of WLAN traffic series, we employ the ADF (Augmented Dickey-Fuller) test [27], which is widely used in statistics. The time series sample in ADF test is generated through an auto-regressive process $AR(p)$. Based on the given sample, a test for a unit root is carried out. The series is un-stationary if the test is failed. The validation of ADF test of two larger-scale WLAN (an airport and a shopping mall) are shown in **Table 1**.

Table 1. Validation of ADF test of two scenarios

	Airport	**Shopping Mall**
ADF test statistic	-3.773524	-11.09497
1% level	-3.449738	-2.571183
5% level	-2.869978	-1.941773
10% level	-2.571335	-1.616066

As shown in **Table 1**, the results of ADF test statistic are less than the significant confidence threshold (1%), which means that the traffic series in our scenarios are stationary.

3.2 Verification of Granger Causality in WLANs

According to investigation, Granger Causality exists for some stationary time series. In this paper we applies Granger Causality test to estimate the causality relationship between different traffic series generated by different APs in WLANs.

A time series X is regarded as the Granger-Cause of Y if the history values of both X and Y provide statistically significant information compared to that of Y alone, namely the series X is helpful in forecasting the future values in series Y.

$$\begin{aligned} Y(t) &= \sum_{i=1}^{p} A_{11,i} X(t-i) + \sum_{i=1}^{p} A_{12,i} Y(t-i) + \varepsilon_1(t) \\ X(t) &= \sum_{i=1}^{p} A_{21,i} X(t-i) + \sum_{i=1}^{p} A_{22,i} Y(t-i) + \varepsilon_2(t) \end{aligned} \tag{1}$$

Where $X(t)$ and $Y(t)$ are the value at current time step. $X(t-i)$ and $Y(t-i)$ are the i-th lagged values in original series, respectively. A_{11}, A_{12}, A_{21} and A_{22} are regression coefficients.

$\varepsilon_1(t)$ and $\varepsilon_2(t)$ are error terms.

Granger Causality is usually determined through a *F*-test. Based on the hypothesis that X is not the Granger-Cause of Y, we execute the two regression process, respectively including and not including the lagged terms of X. The residual sum of squares of former one is RSS_U, and that of later is RSS_R, then *F*-test is defined as:

$$F = (RSS_R - RSS_U) \times (N - 2n - 1) / RSS_U \times n \tag{2}$$

Where n is the number of lagged terms, N is the total sample number. The hypothesis is rejected if the value of *F* is less than threshold $F_a(n, N-2n-1)$ with the significant level α in *F* distribution. Then we can regard series X as the Granger-Cause of Y.

As mentioned above, we have verified the stationary of the traffic series of two larger-scale WLAN (an airport and a shopping mall). In the following, Granger Causality is tested for our data sets. The data sets are from real networks, an international airport equipped with 432 APs, and a shopping mall equipped with 182 APs. The popular matlab toolbox 'Granger Causal Connectivity Analysis' [28] is employed. Regression order is determined by AIC (Akaike's information criterion). The threshold of *F*-test is 0.05. The ratio of APs that have Granger Causality in airport and shopping mall respectively are shown in **Table 2**.

Table 2. The ratio of Granger Causality in airport and shopping mall scenarios

	Airport	**Shopping Mall**	**Total**
The number of APs	432	182	614
The number of APs that has at least one granger neighbor	370	161	531
ratio	85.6%	88.5%	86.5%

As shown in **Table 2**, the APs that have at least one granger neighbor account for at least 85% of the total, which means Granger Causality is universal in our scenarios. So we try to utilize Granger Causality to forecast the traffic flow in WLANs.

3.3 Impact of Granger Causality in Traffic Forecasting

In the previous section, Granger Causality is verified widespread in our scenarios. Therefore we investigate how this phenomenon affects the performance of prediction in this section.

Information entropy is a metric to measure the uncertainly of a system. A lower information entropy value means the system is more predictable. So we uses the information entropy of time series to evaluate how prediction can be improved based on Granger Causality. Three kinds of information entropies values are calculated, namely the information entropy of the target traffic series, the conditional entropy of target traffic series and all the other traffic series, and the conditional entropy of target traffic series and its granger neighbors traffic series, respectively.

Firstly the traffic series are quantified to finite values. The quantitative levels m are set to 50, 100, 200 respectively. $D = \{d_1, d_2, \cdots d_n\}$ is the donation of data set. And d_i is the traffic series of *i*-th AP. Then we quantify each traffic series with level m and obtain $S_i = \{S_{i1}, S_{i2}, \cdots S_{im}\}$, where S_i is the quantization of d_i.

The information entropy of the traffic series of i-th AP is defined as:

$$H(S_i) = \sum_{j=1}^{m} p(s_{ij}) \log_2 (p(s_{ij})) \tag{3}$$

For the given j-th traffic series, the conditional entropy of i-th traffic series can be expressed as:

$$H(S_i / S_j) = \sum_{u=1}^{m} \sum_{t=1}^{m} p(s_{jt}) p(s_{iu} / s_{jt}) \log_2 (p(s_{iu} / s_{jt})) \tag{4}$$

Obviously, a lower $H(S_i / S_j)$ value means a higher definiteness of the system S_i conditioned on S_j, namely S_i is more predictable. Therefore, conditional entropy is a proper metric to measure how predictable the traffic series is [29].

Let $U = \{u_1, u_2, \cdots u_n\}$, u_i is the average conditional entropy of i-th AP at the condition of every other AP. So we can obtain:

$$u_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} H(S_i / S_j) \tag{5}$$

We define the sets $V = \{v_1, v_2, \cdots v_n\}$ and $W = \{w_1, w_2, \cdots w_n\}$. v_i is the average conditional entropy of i-th AP at the condition of granger neighbors and w_i is the APs set of the granger neighbors of i-th AP. Then we can obtain:

$$v_i = \frac{1}{|w_i|} \sum_{S_j \in w_i} H(S_i / S_j) \tag{6}$$

The three kinds of information entropies mentioned above (the information entropy of the target traffic series, the conditional entropy of target traffic series and all the other traffic series, and the conditional entropy of target traffic series and its granger neighbors traffic series) are donated as $H(S), u, v$ respectively. $H(S)$ is measure ofaveraged indefiniteness of traffic for an AP without any other referenced informtion. u is measure of averaged indefiniteness of traffic for an AP conditioned on one of its neighbors. v is measure of averaged indefiniteness of traffic for an AP conditioned on one of its granger neighbors. We calculate those three kinds of information entropies for 432 APs of airport and 182 APs of shopping mall, and find that in the airport scenario, 87.5% of the total APs follow the relationship as $H(S) > u > v$, while the ratio is 95.8% in the scenario of shopping mall.The results indicate that multivariate time series make system more predictable then univariate time series in our scenarios, in particular, with the help of Granger Causality, the target traffic series becomes much more predictable compared to take irrelevant AP into account.

For demonstration and explaination, we take 18 APs of airport scenario and 26 APs of shopping mall scenario for example, and the values of $H(S), u, v$ are shown in **Fig. 1** and **Fig. 2**.We find that some conditional entropy v is nearly zero, i.e. the traffic of target AP could be determined with the traffic of the granger neighbors , which indicates that the traffic of granger neighbors is in great relevance with target APand helpful for traffic predication. As showed in **Fig. 1** and **Fig. 2**, the values of $H(S), u, v$ vary over different quantitative levels, but the relative of $H(S), u$ and v is same.

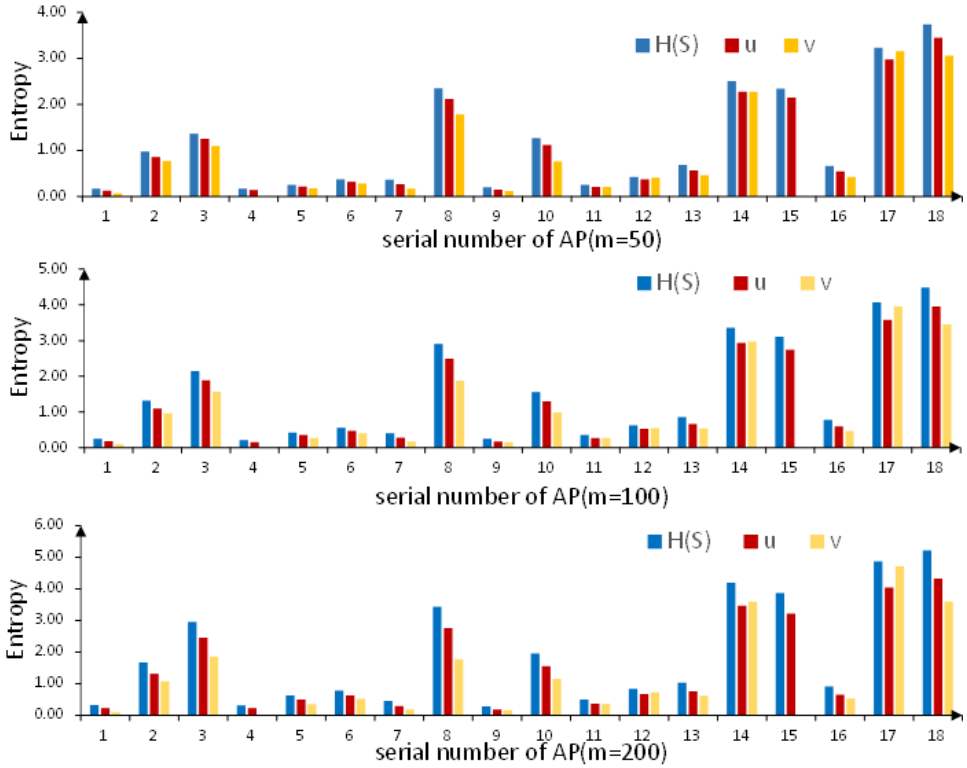


Fig. 1. The value of $H(S),u,v$ for airport scenario with different quantitative levels

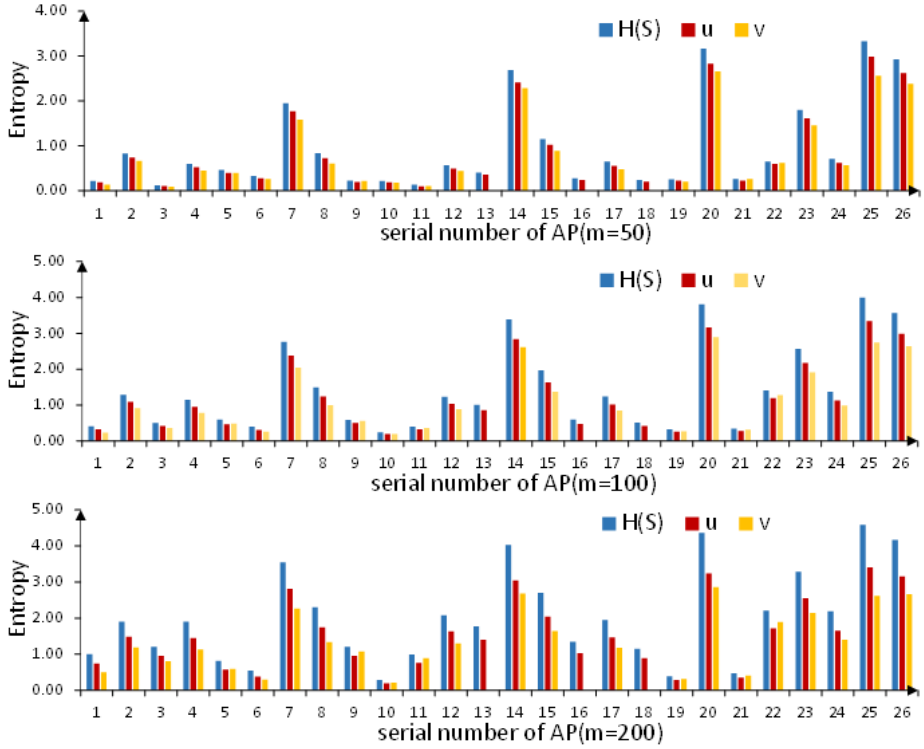


Fig. 2. The value of $H(S),u,v$ for shopping mall scenario with different quantitative levels

3.4 The Causing of Granger Causality in WLANs

The above analysis shows that Granger Causality is universal in our scenarios. Moreover, the prediction of network traffic can be improved by such relationship. Here we try to explain this physical phenomenon in the view of human behavior. Wireless channel is open and shared by all APs. Adjacent APs may share overlapping areas. The movement of users in these areas can cause the re-association between users and APs, and then produce the causal traffics in adjacent APs. So Granger Causality in WLANs attributes to the movement of users and shared channel of APs.

For larger-scale WLAN, the network is heavy dense, with the people show strong organization in mobility. For example, in the airport, people transfering abroad usually go from flight connections to International terminal. So if a domestic flight with many passengers arrives, many users associate with AP in flight connections, which leads to the incrased traffic within half an hour. But after then the traffic in the AP decreases gradually while the traffic of AP in the entrance of International terminal increases progressively, because many users de-associate with AP in flight connections and re-associate with AP in International terminal .

Thus the correlation between APs may be generated by the mobility of human, which results in the handoff between different APs, and then the causality between traffic of different APs. In other words, the causality between different traffics also reflects the activity regularity of people.

Literature [13] also revealed the similar Granger Causality in cellular network, which indicates that the correlation between nodes in wireless channel is universal. This discovery is useful in network management and traffic prediction in wireless network.

4. Traffic Prediction Method

As it mentioned above, the Granger Causality is widespread in our scenarios. The performance of prediction towards target node can be enhanced by utilizing nodes that shares Granger Causality with it. As VAR (Vector Auto-Regression) is constraints-free in a predict process, it is quite suit for our problems. Compared to univariate time series, multivariate time series is more complete and identified, as well as introducing redundant and noise information. Therefore, we employ the APs sharing Granger Causality with target AP to avoid redundant and noise information, and applies VAR model to predict traffic flow of target AP. So we develop the traffic prediction method – GCVAR (Granger Causality and Vector Auto-regression). There are two primary steps for GCVAR. In the first step, neighboring APs sharing Granger Causality with the target AP is computed, which was introduced above. In the second step, the traffic of target AP is predicted using the past traffic of causal APs through VAR. In the following, the second step would be introduced in details.

4.1 Brief Introduction to VAR

VAR model is widely used in statistics and forecast. It is the simultaneous form of AR(Auto-Regression) model in univariate time series. Without exogenous variables, a VAR model with order p can be expressed as:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t \tag{7}$$

Where c is a $n \times 1$ constant vector. y_t is $n \times 1$, in which each element is the value of corresponding variable in model at time t. A_i is the $n \times n$ parameter matrix to be predicted, indicating the transforming relation between y_t and y_{t-1}. e_t is an $n \times 1$ error vector, which

satisfies to:

(1) $E(e_t)=0$

(2) $E(e_t e_t')=\Omega$, and Ω is a $n\times n$ positive definite matrix

(3) $E(e_t e_{t-k}')=0$, with $k\neq 0$, which means error terms are independent.

4.2 Application of VAR in Traffic Prediction

Let $V(t)=(V_1(t),V_2(t),\cdots V_k(t))$ denotes the traffic series of network, $V_k(t)$ is the traffic series of *k*-th AP. Note that all traffic series are synchronous. So the VAR model with order *p* of network traffic can be expressed as Eq(7).

$$V(t)=C+A_1V(t-1)+\cdots+A_pV(t-p)+u(t) \tag{8}$$

Where $u(t)=(u_1(t),u_2(t),\cdots u_k(t))'$, in which each component is identically distributed, i.e. $u(t)\sim N(0,\Omega)$. Ω is the covariance matrix of $u(t)$. $C=(C_1,C_2,\cdots C_k)'$ is the constant vector. $A_i(i=1,2,\cdots p)$ is the parameter of VAR model which is obtained based on the estimation from history traffic. It is the component of coefficient matrix, with the form as follows:

$$A_i=\begin{bmatrix} a_{11,i} & \cdots & a_{1k,i} \\ \vdots & \ddots & \vdots \\ a_{p1,i} & \cdots & a_{pk,i} \end{bmatrix}$$

We apply MLE (Maximum Likelihood Estimation) to estimate the parameters of VAR model. According to Eq(8), we can obtain:

$$V_t-\mu=A_1(V_{t-1}-\mu)+\cdots+A_p(V_{t-p}-\mu)+u_\tau \tag{9}$$

When rewrite Eq(8) to a matrix form, we can obtain:

$$WS=A\cdot W+U \tag{10}$$

Where $WS=(V_1-\mu,\cdots,V_T-\mu)_{K\times T}$, $A=(A_1,\cdots,A_p)$, $WS_t=\begin{bmatrix} V_t-\mu \\ \vdots \\ V_{t-p+1}-\mu \end{bmatrix}$, and $W=(WS_0,\ldots,WS_{T-1})$, $U=(u_1,\ldots,u_T)$.

The log-likelihood function can be expressed as:

$$\ln l(\mu,a,\Omega)=-\frac{KT}{2}\ln 2\pi-\frac{T}{2}\ln|\Omega|-\frac{1}{2}tr[(WS-A.W)'\Omega^{-1}(WS-A.W)] \tag{11}$$

Where $a=vec(A)$. Then we take partial derivative with μ, a, Ω respectively, and set the equations to 0. The maximum likelihood estimation of the three parameters are:

$$\tilde{\mu}=\frac{1}{T}(I_k-\sum_{i=1}^{p}\tilde{A}_i)^{-1}\sum_{t=1}^{T}(V_t-\sum_{i=1}^{p}\tilde{A}_iV_{t-i}) \tag{12}$$

$$\tilde{a}=[(\tilde{V}\tilde{V}')^{-1}\tilde{V}\otimes I_k](V-\tilde{\mu}) \tag{13}$$

$$\tilde{\Omega}=\frac{1}{T}(\tilde{WS}-\tilde{A}\tilde{V})(\tilde{WS}-\tilde{A}\tilde{V})' \tag{14}$$

The lag order used in equations above is decided by AIC (Akaike's information criterion):

$$AIC(p) = \ln\left|\tilde{\Omega}(p)\right| + \frac{2pK^2}{T} \tag{15}$$

We set the maximum lag order $p_{\max}$, and set p from 1 to $p_{\max}$. Finally, the p that minimizes $AIC(p)$ is selected as the best lag order for the modeling of VAR.

5. Experiment

In this section, we apply GCVAR to predict the AP traffic in two above real networks, and compare the performance of GCVAR, ARIMA, and WRIMA. The comparison of accuracy and complexity are provided.

5.1 Scenarios

The data used in this paper are from real networks, an international airport equipped with 432 APs, and a shopping mall equipped with 182 APs. We utilize the management platform to collect traffic data once an hour in two areas. And finally we acquire traffic data with 672 time moments (28 days). Note that both scenarios have the same timelines. Both scenarios support 802.11 b/g/n protocols.

To demonstrate the performance of GCVAR, we selected 18 APs from data set of airport scenario, and 26 APs from data set of shopping mall scenario. Both regions have relatively dense crowd with high mobility and the accurate traffic prediction is difficult. We apply GCVAR model, WARIMA model and ARIMA model for these two senerios and compare the results.

5.2 Performance Comparison

In purpose to compare the accuracy of these three methods, we collect the percentage of absolute errors between prediction results and the real values in the airport and mall for every AP. The cumulative distribution of absolute errors in the airport and in the mall are shown in **Fig. 3** and **Fig. 4** respectively. The CDF of other APs are in similar trend.

As it shows in **Fig. 3**, the prediction results of GCVAR model are better than that of ARIMA model and slightly worse than that of WARIMA model in airport scenario. Among the prediction results based on WARIMA model , 97% of absolute errors are less than 10%. As for GCVAR and ARIMA model, the percentage declines to 94%. While in **Fig. 4**, the prediction results of GCVAR model is better than that of ARIMA and WARIMA. In GCVAR, 86% of absolute errors is less than 5%, while the percentage of ARIMA and WARIMA are only 80%.

We also note that WARIMA costs 3396.6s in the prediction process running, and ARIMA cost 596.62s, GCVAR only costs 27.689s. The running time of GCVAR model is shorter than the other two models by two orders of magnitude, which indicates that GCVAR model is much effective compared with ARIMA and WARIMA. GCVAR sacrifices a little accuracy for a much lower time consumption. Decomposition in multiple scales is needed in WARIMA, so it is much time-consuming. In GCVAR, causal neighbors and related information are computed before VAR algorithm is used, so the time consumption is least.

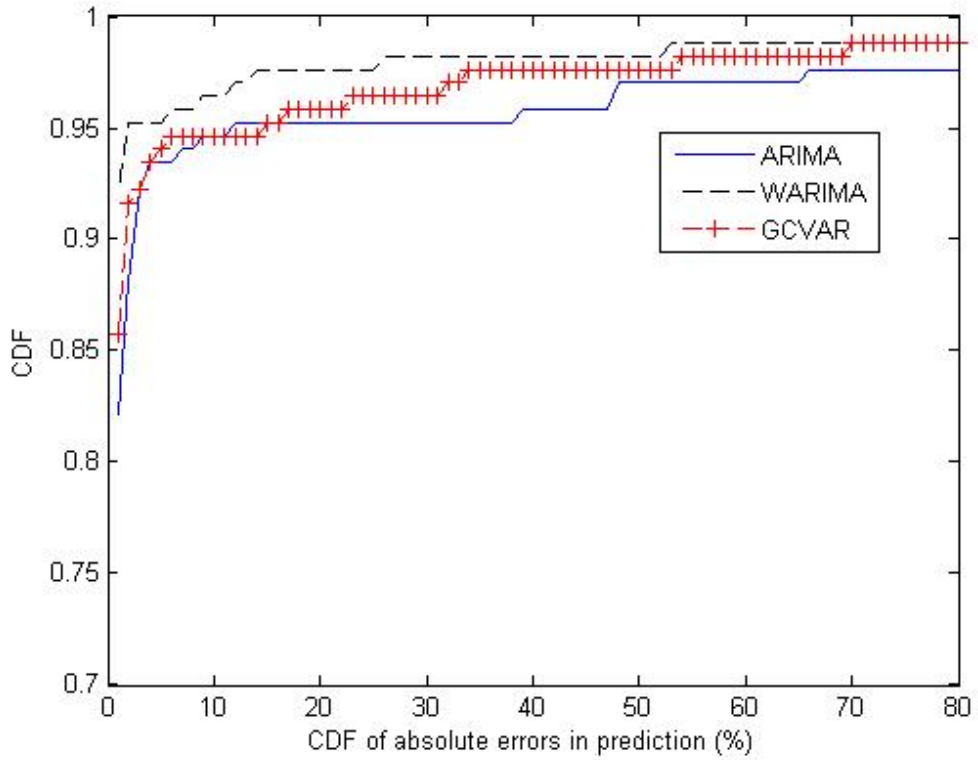


Fig. 3. Cumulative distribution of absolute errors in the airport

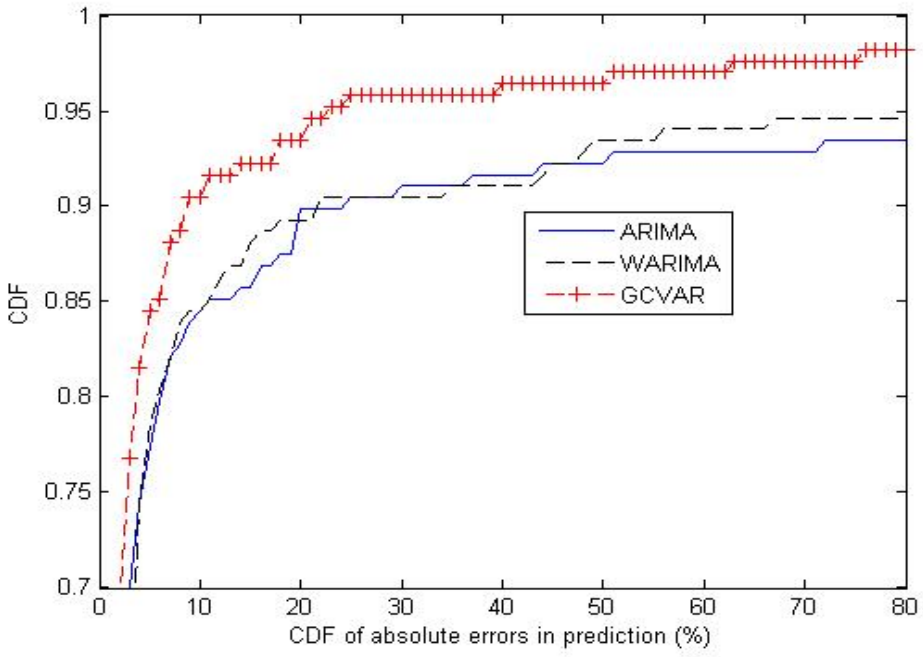


Fig. 4. Cumulative distribution of absolute errors in the mall

Take AP18 in the airport for example, the comparison of real traffic and prediction results are shown in **Fig. 5**. In the traffic spike, WARIMA is obviously more accurate than GCVAR, and GCVAR is more accurate than ARIMA.

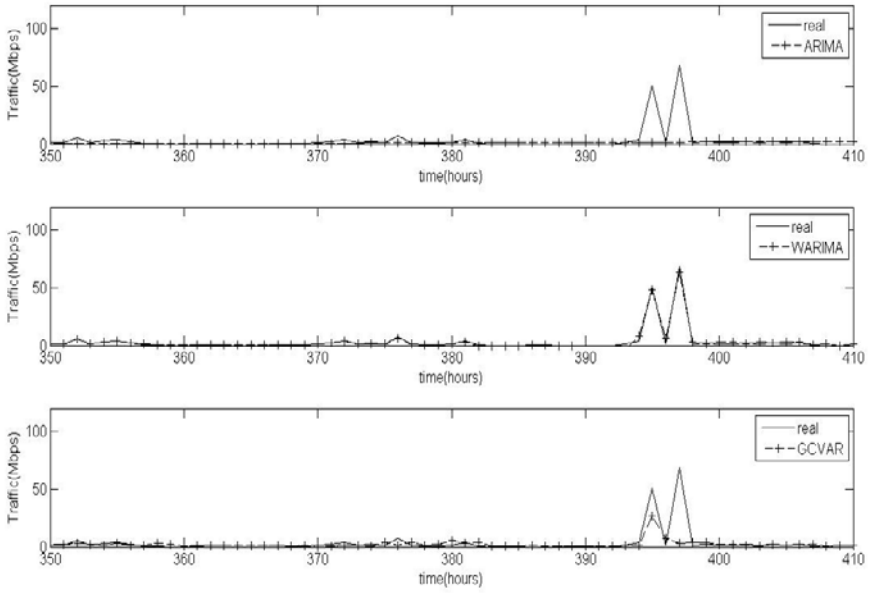


Fig. 5. Comparison of traffic prediction of AP18 in the airport

In order to compare the stability of different algorithms, we analyze the time percentages for different percentage of absolute errors in all APs in two scenarios. For percentage of absolute errors less than 10%, the time percentages of the correct prediction in the airport are shown in **Fig. 6**. For percentage of absolute errors less than 5%, the time percentages of the correct prediction in the mall are shown in **Fig. 7**.

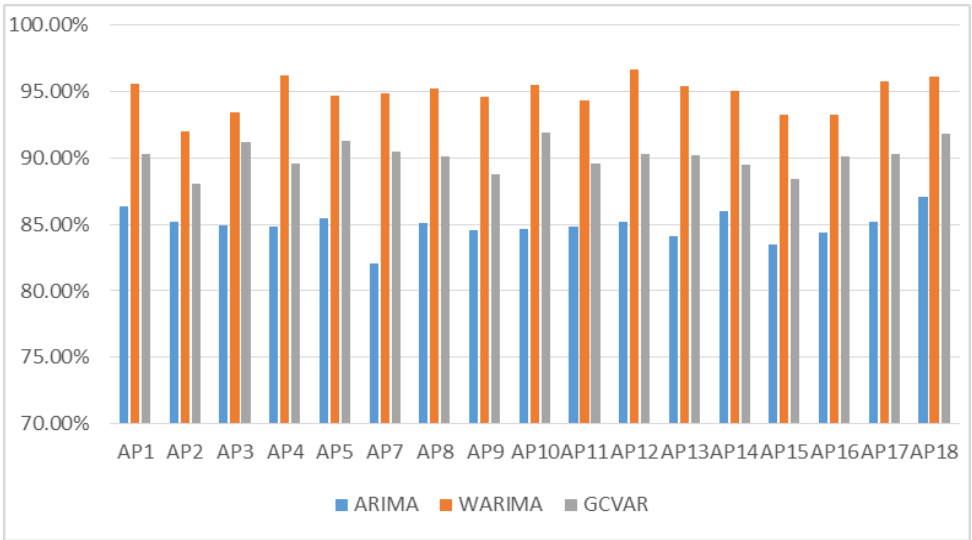


Fig. 6. The time percentage of correct prediction in the airport with absolute errors less than 10%

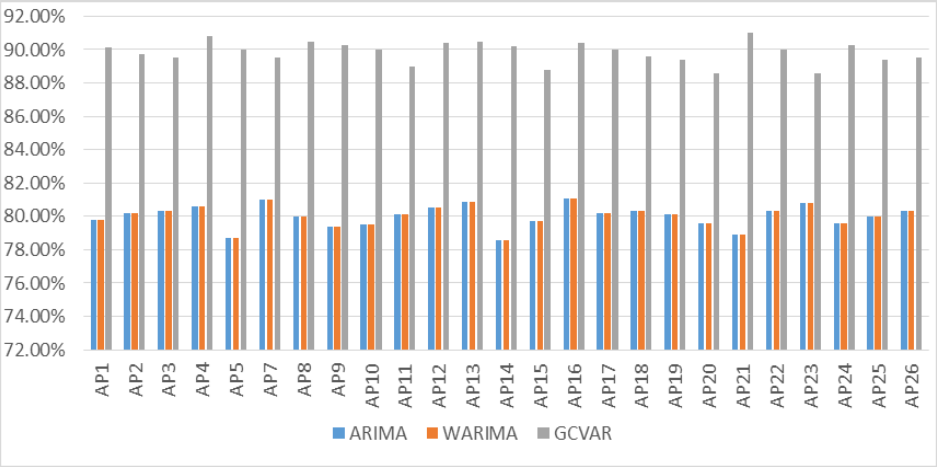


Fig. 7. The time percentage of correct prediction in the mall with absolute errors less than 5%

As it shown in **Fig. 6** and **Fig. 7**, the prediction results of ARIMA model and WARIMA model fluctuate obviously in different scenarios, while GCVAR model performs relatively stable. The small difference of prediction accuracy among the three models also suggests that the traffic patterns of the APs under a certain scenario are close. **Fig. 6** and **Fig. 7** also show that in the airport, WARIMA performs the best, however, WARIMA is the most time-consuming. While in the mall, GCVAR outperforms other algorithms, as well as most time-saving. It is also verified in the experiment that although GCVAR is less accurate than WARIMA in some cases, it is the most stable and time-saving.

In WARIMA, network traffic series are decomposed with more complex factor into sub-series on different scales, so both profile and details could be extracted, but it is very time-consuming. If the burst of traffic is prominent or non-stationary, WARIMA shows correct prediction obviously in more time because the details with high frequency information could be predicted in small scale, such as in the airport scenario. Otherwise the performance of WARIMA could be similar to ARIMA because only profile information needed to be captured in both algorithms, such as in the mall scenario. But only the traffic of target AP is used for prediction in WARIMA, while traffics of causal neighboring APs are used in GCVAR, so in the mall scenario GCVAR is obviously better than others. Also details could not be predicted very accurately in GCVAR, so GCVAR performs worse than WARIMA in the airport scenario. But GCVAR is more stable no matter the traffic is rich in high frequency information or not.

6. Conclusion

In this paper traffic prediction based on GCVAR in large-scale WLANs is performed. Based on two real networks, we reveal the widespread of Granger Causality in our scenarios. In combination with information entropy, the prediction performance of target AP can be enhanced by taking the history data of its granger neighbors into account. We proposes GCVAR, which takes APs series sharing Granger Causality based on VAR (Vector Auto-regression) into account, to predict network traffic for larger-scale WLAN. Simulation shows that GCVAR is more effective compared to ARIMA and WARIMA. Besides, GCVAR

is more practical since it consumes two orders of magnitude less than that caused by ARIMA/WARIMA.

GCVAR reduces the time-consumption to a significant level by sacrificing little accuracy, however, the correlation between APs is obtained from the perspective of statistics. In our future work, we will try to combine GCVAR with principal component analysis, or other clustering algorithms, so that physical meaning is more obvious, as well as to improve the performance of GCVAR.

References

[1] Chen C, Pei Q, Ning L, "Forecasting 802.11 Traffic Using Seasonal ARIMA Model," in *Proc. of IEEE Int. Forum on Computer Science-Technology and Applications*, pp.347-350, December 25-27, 2009. Article (CrossRef Link)

[2] JIANG Ming, WU Chun-ming et al, "Research on the Comparison of Time Series Model for Network Traffic Prediction," *ACTA Electronica Sinica*, Vol.37, No.11, pp.2353-2357, November, 2009. Article (CrossRef Link)

[3] G. E. P Box, G. M. JENKINS, *Time Series Analysis Forecasting and Control*, 3rd Edition, Prentice Hall, Upper Saddle River, New Jersey, 1994. Article (CrossRef Link)

[4] Peng W, Yuan L, "Network traffic prediction based on improved BP wavelet neural network," in *Proc. of. IEEE 4th Int. Conference on Wireless Communications, Networking and Mobile Computing*, pp.1-5, October 12-14, 2008. Article (CrossRef Link)

[5] Di C., Hai-Hang F., Qing-jia L. et al, "Multi-scale Internet traffic prediction using wavelet neural network combined model," in *Proc. of 1st IEEE International Conference on Communications and Networking in China*, pp.1-5, October 25-27, 2006. Article (CrossRef Link)

[6] Chen X T, Liu J X, "Network traffic prediction based on wavelet transformation and FARIMA," *Journal on communications*, Vol.32, No.4, pp.153-157, 2011. Article (CrossRef Link)

[7] Daqiang Zhang, Zhijun Yang, Vaskar Raychoudhury, Zhe Chen, Jaime Lloret, "An Energy-efficient Routing Protocol Using Movement Trend in Vehicular Ad-hoc Networks," *The Computer Journal*, vol. 56, no. 8, pp. 938-946, 2013. Article (CrossRef Link)

[8] Jiang H, Zhou C, Wu L, et al, "TDOCP: A two-dimensional optimization integrating channel ass gnment and power control for large-scale WLANs with dense users," *Ad Hoc Networks*, Vol.26, pp.114-127, March, 2015. Article (CrossRef Link)

[9] Granger, C. W. J., "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, Vol.37, No.3, pp.424-438, 1969. Article (CrossRef Link)

[10] Paul U., Subramanian A. P., Buddhikot M. M., Das S. R., "Understanding traffic dynamics in cellular data networks," in *Proc. of IEEE INFOCOM*, pp. 882-890, April 10-15, 2011. Article (CrossRef Link)

[11] Taylor, J. W., "A comparison of univariate time series methods for forecasting intraday arrivals at a call center," *Management Science*, Vol.54, No.2, pp.253-265, 2008. Article (CrossRef Link)

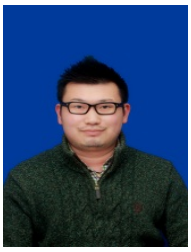
[12] Dominguez G., Guevara M., Mendoza M., Zamora J., "A wavelet-based method for time series forecasting," in *Proc. of 31st International Conference of the Chilean Computer Science Society*, pp.91-94, November 12-16, 2012. Article (CrossRef Link)

[13] Man-Chun Tan, Wong S.C., Jian-Min Xu, Zhan-Rong Guan, et al., "An Aggregation Approach to Short-Term Traffic Flow Prediction," *IEEE Transactions on Intelligent Transportation Systems*, Vol.10, No.1, pp.60-69, 2009. Article (CrossRef Link)

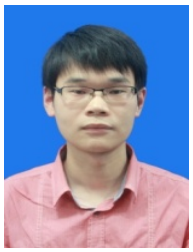
[14] Nakayama H., Ata S., Oka I., "Predicting time series of individual trends with resolution adaptive ARIMA," in *Proc. of IEEE International Workshop on Measurements and Networking*, pp. 143-148, October 7-8, 2013. Article (CrossRef Link)

[15] J.C. Cuaresma, J. Hlouskova, S. Kossmeier, M. Obersteiner, "Forecasting electricity spot-prices using linear univariate time-series models," *Applied Energy*, Vol.77, No.1, pp.87-106, 2004. Article (CrossRef Link)

[16] Medeiros, M. C., Veiga A., "A hybrid linear-neural model for time series forecasting," *IEEE Transactions on Neural Networks*, Vol.11, No.6, pp.1402-1412, 2000. Article (CrossRef Link)
[17] Holanda Filho, R. and J. E. B. Maia, "Network traffic prediction using PCA and K-means," in *Proc. of IEEE Network Operations and Management Symposium*, pp.938-941, April 19-23, 2010. Article (CrossRef Link)
[18] H. Feng, Y. Shu, S. Wang, M. Ma, "SVM-based models for predicting WLAN traffic," in *Proc. of IEEE Int. Conference on Communications*, pp.596-602, June 11-15, 2006. Article (CrossRef Link)
[19] Liang, Yonglin, and Lirong Qiu. "Network Traffic Prediction Based on SVR Improved By Chaos Theory and Ant Colony Optimization.," *International Journal of Future Generation Communication and Networking*, Vol.8, No.1, pp.69-78, 2015. Article (CrossRef Link)
[20] M.U. Ahmed, D.P. Mandic, "Multivariate multiscale Entropy analysis," *IEEE Signal Processing Letters*, Vol.19, No.2, pp.91-94, 2012. Article (CrossRef Link)
[21] B Ghosh, B Basu, M O'Mahony, "Multivariate short-Term traffic flow forecasting using time-series analysis," *IEEE Transactions on Intelligent Transportation Systems*, Vol.10, No.2, pp.246-254, 2009. Article (CrossRef Link)
[22] K Koçak, L Şaylan, J Eitzinger, "Nonlinear prediction of near-surface temperature via univariate and multivariate time series embedding," *Ecological Modelling*, Vol.173, No.1, pp.1-7, 2004. Article (CrossRef Link)
[23] Xiang, Zhengtao, et al. "Predictability of Aggregated Traffic of Gateways in Wireless Mesh Network with AODV and DSDV Routing Protocols and RWP Mobility Model," *Wireless Personal Communications*, Vol.79, No.2 , pp.891-906,2014. Article (CrossRef Link)
[24] Li, Rongpeng, et al. "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice," *IEEE Communications Magazine*, Vol.52, No.6, pp.234-240, 2014. Article (CrossRef Link)
[25] Daqiang Zhang, Hongyu Huang, Jingyu Zhou, Feng Xia, and Zhe Chen, "Detecting Hot Road Mobility of Vehicular Ad Hoc Networks," *ACM/Springer Mobile and Network Applications*, vol. 18, no. 6, pp. 803-813, 2013. Article (CrossRef Link)
[26] Daqiang Zhang, Jiafu Wan, Zongjian He, Shengjie Zhao, Ke Fan, Sang Oh Park, "Identifying Region-wide Functions Using Urban Taxicab Trajectories," *ACM Transactions on Embedded Computing Systems*, 2015. Article (CrossRef Link)
[27] RID Harris, "Testing for unit roots using the augmented Dickey-Fuller test: some issues relating to the size, power and the lag structure of the test," *Economics Letters*, Vol.38, No.4, pp.381-386, 1992. Article (CrossRef Link)
[28] A. K Seth, "A MATLAB toolbox for Granger causal connectivity analysis," *Journal of Neuroscience Methods*, Vol.186, No.2, pp.262-273, 2010. Article (CrossRef Link)
[29] Z. Chun-Tao, M. Qian-Li, P. Hong, J. You-Yi, "Multivariate chaotic time series phase space reconstruction based on extending dimension by conditional entropy," *Acta Physical Sinica*, Vol.60, No.2, pp.1-8, 2011. Article (CrossRef Link)



Zheng Lu received his double Bachelor's degree (major at telecommunication and business administration) in the Wuhan University, China, in 2006. He received his Master's degree (MBA) at Wuhan University, China, in 2013. He is currently a Ph.D. candidate in the School of Electronic Information, Wuhan University, China. His research interests are in the area of data mining and machine learning.



Chen Zhou received his Bachelor's degree in the College of Physical Science and Technology, Huazhong Normal University, China, in 2011. He is currently a Ph.D. candidate in the School of Electronic Information, Wuhan University, China. His research interests are in the area of wireless communication, data mining of complex networks.

Jing Wu received the BS degree and the Ph.D. degree in School of Electronic Information in Wuhan University, Wuhan, China, in 2002 and 2008, respectively. She is currently an Associate Professor in School of Electronic Information, Wuhan University. Her research interests include wireless ad hoc network, satellite network and network simulation.

Hao Jiang received the BS degree and the Ph.D. degree in communication networks from Wuhan University, Wuhan, China, in 1998 and 2004, respectively. He is currently a Professor at the School of Electronic Information, Wuhan University. His research interests include wireless LANs, wireless ad hoc network, and vehicle ad hoc network.

Songyue Cui received his Bachelor degree in the School of Electronic Information at Wuhan University, in 2012, and he is pursuing a master degree in the School of Electronic Information at Wuhan University. His research interests are in the area of data analysis.