TELKOMNIKA, Vol. 11, No. 8, August 2013, pp. 4817~4823
e-ISSN: 2087-278X

Segmentation, Clustering and Timing Relationship Analysis of MANET Traffic Flow

Huijun Chang*, Hong Shan, Tao Ma
Department of Network Engineering, Electronic Engineering Institution, Hefei, 230037, China
*Corresponding author, e-mail: changhj2417@126.com

Abstract

Users in mobile Ad Hoc networks (MANET) usually encrypt their data packets to resist the evasdroppers, which makes the network management and Intrution detection difficult. However, user behavior, ultimately displayed as traffic flow, shows regularity along time. This paper aims to study the regularity through studding the timing relationship between traffic flows, whose results provide the technical support for user behavior analysis. First, segment the end-to-end flows based on the information of time intervals and packet lengths. Second, cluster the segments by an improved maximum-distance method. Third, analyze the time relationship between the clusters, i.e., traffic flow types, based on the clustering results. Simulation results verify the effectiveness of the method.

***Keywords**: traffic flow segmentation, maximum-distance algorithm, apriori algorithm*

Copyright © 2013 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

The rapidness of networking and ease of communication make MANET widely used in military, commercial, emergency services. MANET is highly vulnerable to eavesdroppers due to its open medium, furthermore, it is more likely in MANET than other in Internet to accept new nodes, even malicious nodes because of dynamic topology. Therefore, the security problem of MANET is more serious.

Some MANET applications use encryption or anonymous communication techniques to prevent the attack of eavesdroppers [1-3]. However, the encryption mechanism could not prevent the analysis of traffic flow [4-6]. Some other MANET applications use authentication method [7-9] to avoid the joining of malicious nodes, nevertheless, some malicious nodes could still deceive the authentication node through replay attack [10, 11] and other means. As a result, encryption and identity authentication could not completely guarantee the safety of MANET, and a certain intrusion detection system is still required. Recently researchers have put forward some intrusion detection mechanisms [12-15] in MANET, but these mechanisms generally require one or more central node to run complicated testing procedures, which result in large energy consumption and hinders its wide application in MANET.

This paper presents an analysis method of traffic flow timing relationship to find user abnormal behavior, which is simple and easy to be realized, and does not produce much communication. Arrange monitoring agents in the network, all of which can monitor the information exchange of the entire network, and run the following algorithm on each agent. Firstly, based on the time interval and the packet length information, the user packet series is segmented. Then, based on the distribution of packet length, time interval and the length of the segment, the segments are clustered. Finally, based on Apriori algorithm, the timing relationships between the traffic flow types is analyzed, which can be used for user behavior analysis.

2. Model Statement and Parameter Definition

Figure 1 shows a MANET, whose nodes locate randomly and can move arbitrarily. Install multiple monitoring agents to cover the entire network, which can obtain the data transmission time series P between any two nodes in the network, where $P=(p_1, p_2, \dots p_n)$, $p_i=(t_i, l_i)$, with t_i and l_i representing the receiving time and packet length of each data packet,

Received January 18, 2013; Revised May 27, 2013; Accepted June 5, 2013

respectively. By this way, the processing of the collected data can be done on the local agent, without a sink node or the information interaction between agents.

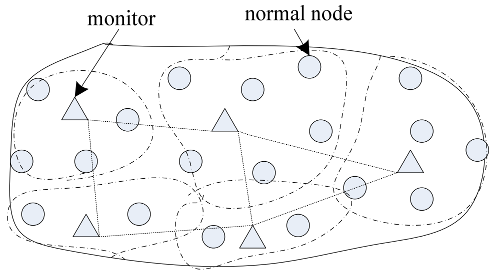

Figure 1. Target Network Diagram

According to the flow characteristic, P is supposed to be divided into K categories, represented as V_k, $k = 1, 2, ...K$. Each V_k is composed of segments of traffic flows with similar flow characteristics, i.e., $V_k = \{Q_j\}, Q_j = p_{j1}, p_{j2}, ...p_{jx}$. Then the timing relationships between V_i and V_j should be analyzed.

3. Traffic Flow Analysis

3.1. Top-down Time Series Segmentation

Existing segmentation (or summarization) methods are the methods of representing time series, which aim at decreasing the representation dimensions under the precondition of maintaining the basic characteristic of time series. They are mainly divided into three categories: sliding window method [16-18], top-down method [19-21], and down-up method [22-24]. The sliding window anchors the left point of a potential segment at the first data point of a time series, and attempts to approximate the data to the right with increasing longer segments, until the error between the represented and original time series exceeds a threshold. Top-down method is a divide-and-rule method. It recursively divides the time series into different sub-series until some stopping criterion is met. Down-up method combines the points or segments with the lowest combining cost until the cost exceeds a threshold. According to the objective, time series segmentation method can be divided into two categories: aiming at minimizing storage space [20, 22, 24] and approximation error [16]. These two existing methods are usually applied to the representation of time series, however, in this paper the time series segmentation problem is: Divide $P = (p_1, p_2, ...p_n)$ into k continuous segments $Q_1, \cdots, Q_k$, where $Q_j = p_{j1}, p_{j2}, ...p_{jx}$, while maximizing the similarity between each element in Q_j. Therefore, the segmentation method here is different from those existing methods in essence, resulting that those existing segmentation methods become invalid here.

Regarding the frame receive time interval and packet length as the principle standard, a simple traversing method is proposed to segment the data packet time series, which divides the neighboring packets possibly belonging to different traffic flows into different segments. The details of proposed segmentation method are shown in Algorithm 1. Since two nodes do not maintain the continuous communication, the time series before and after the time interval can be considered as two different traffic flows if the time interval exceeds T seconds. The time interval between traffic flows could be very short when communication is dense. Furthermore, two different traffic flows can be transmitted between two users coincidentally. Considering the above cases, only basing on the time interval of data packet cannot differentiate different traffic flows. Since different traffics use different protocols, the data packet lengths are different. Correspondingly, formula (1) can be used for determining the sudden change of length at p_i. Segment at p_i when formula (1) satisfies. Thus, the time series is segmented to different traffic flows after one traversal.

$$\frac{\max(l_i, l_{i+1})}{\min(l_i, l_{i-1})} > \Delta \quad (1)$$

Algorithm 1：Segmentation
Input：time series $P = (p_1, p_2, ..., p_n)$
k=1; $Q_k.s = p_1$, $Q_k.e = p_n$ //beginning and end
for i=2:n
if $t_i - t_{i-1} > T$ or $\frac{\max(l_i, l_{i+1})}{\min(l_i, l_{i-1})} > \Delta$,
k=k+1; $Q_{k-1}.e = p_{i-1}$;
$Q_k.s = p_i$; $Q_k.e = p_n$;
end if
end for
Output：segmented time series{ Q_i }

3.2. Clustering of Segments

In this section, we classify the segmentation results of time series. i.e., traffic flows. Due to the encryption mechanism, the traffic class number is unknown where clustering technique is a candidate for the classification. Clustering techniques classify the sample set into clusters. The samples inside the same cluster are similar to each other, while the samples between different clusters have the utmost difference. There are 3 key points in clustering techniques: the selection of sample properties, calculation method of sample distance, and selection of clustering method.

The flow characteristics are mainly shown in Table 1. Each sample, represented by these characteristics, stands for a traffic flow. Putting these samples into the clustering algorithm, those traffic flows can be divided into different categories.

Table 1. Flow Characteristics Related to the Traffic

data packet/control packet	packet number
	packet interval(average,max,min)
	packet length(average,max,min)

Since data packets that have obvious changes of packet interval and length have been divided into different segments during the aforementioned segmentation, the parameters of ‘max’ and ‘min’ are no longer considered.

Considering the different metric scales of the above three parameters, we use Mahalanobis distance to calculate the sample distance. It involves the relationship between various characteristics and shields the scale difference, which can efficiently compute the similarity degree between two unknown data sets. Let *X*, *Y* be two samples from the sample set with mean μ and covariance $\sum$. The Mahalanobis distance between *X* and *Y* is:

$$d_m^2(X,Y) = (X-Y)'\sum{}^{-1}(X-Y) \quad (2)$$

Clustering with category number unknown mainly consist two categories: adaptive sample set construction method [25] and maximum-distance clustering method [26]. Adaptive sample set construction method requires relatively compact sample set and large distance between samples from different groups. Due to the encryption mechanism, the differentiation between different traffic flow characteristics on the link layer is not distinct, resulting in the invalidity of this method. The maximum distance clustering method can guarantee each new cluster center relatively far from the existing cluster centers, and intelligently determine the number of initial cluster centers. However, its performance depends on the choice of the initial cluster centers, so the cluster center choosing part of the method is slightly improved in this paper. The detailed clustering algorithm is as follows.

(1) Extract *n* samples from the segments set. For each sample, compute the sum of M-distances between it and other samples. Among all the sums, choose the sample X_1 which makes the maximum sum as the first clustering center $Z_1=X_1$.

(2) From the segments set, choose the sample farthest from Z_1 as the second clustering center Z_2.

(3) For each X_i of the rest samples, compute the Mahalanobis distances from Z_1 and Z_2 respectively, and let the smaller one be D_{Xi}.

(4) If the maximum value in $\{D_{Xi}\}$ is not less than α of M-distance between Z_1 and Z_2, then X_i is another clustering center, turn to (5). Otherwise, turn to (6).

(5) Redo (3) and (4).

(6) Classify the rest samples to their nearest clustering centers.

3.3. Timing Relationship Analysis

The packet series can be represented as Figure 2 shows after clustering, where ABC represent different traffic flows, respectively.

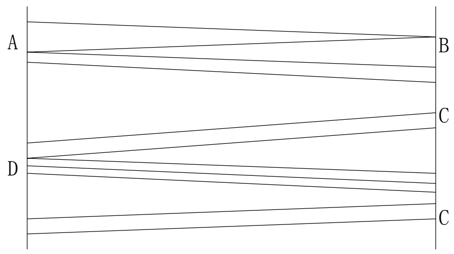

Figure 2. Traffic Flow Interchange between Nodes

Treating each segment of traffic flow as an ensemble, the whole sample data set can be represented as a time series on the time axis. If different traffic flows are time related, e.g., B appears after A, this relationship can be applied in attacking. Information exchange probably has some certain modes, for example, the response of situation information appears after the request of situation information and then the commander would probably transmit some control commands. If this possible mode can be discovered, replaying the preceding situation or control information according to the request time points can confuse the network users. Since the nodes in some MANETs periodically transmit situation information, relaying attack can be easier to apply if the traffic flow has time periodicity.

The overall of discovering the time relationship between traffic flows is as follows: Traverse one time, calculate the numbers that other segment types appear after each segment type, and utilize A-priori algorithm [27] to compute the belief of each type of time relationship based on these numbers. The two kinds of segments satisfying a certain confidence are considered to be time related. Taking segment types A and B for example, the belief of A appearing after B is $support(A,B)/support(A)$, where $support(A)$ means the number that A appears in the whole segment series and $support(A,B)$ means that A appears after B in the whole segment series. That is what we call timing relationship between A and B. If the traffic flows newly come do not meet the regularities, there may be some abnormal behavior.

4. Simulation and Results

10 volunteers provide 10 notebooks comprised a MANET. These notebooks generated communication between them, and access the Internet through a gateway. Monitor and record the communication between the two of them, 2 and 3, e.g., for a period of 10 days. Take the packet series of the 10 days as input, run the analysis algorithm on Matlab platform, we can test the algorithm.

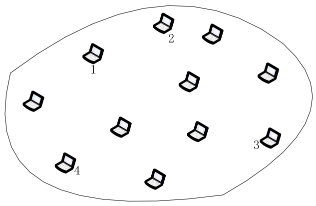

Figure 3. Diagrammatic Sketch of Experimental Scene

The result of timing relationship analysis is highly related to the clustering accuracy, which is the percentage of correctly clustered records in the total communication records. Clustering accuracy is mainly affected by the following parameters: interval T , message length mutation threshold Δ, distance coefficient α , and sample number n. The experiments were divided into 4 groups, to examine the effect of above 4 parameters on clustering accuracy, respectively. Parameter setting for 4 groups of experiments is shown in Table 2, where n =1 means to choose a random sample as the first cluster center.

Table 2. Setting of Experimental Parameters

	the fixed parameters	the varied parameters
1	$\Delta = 6$, $\alpha = 0.3$, n =15	T =5,10,15,20,25,30,40,50,60
2	T =10, $\alpha = 0.3$, n =15	Δ =3,4,5,6,7,8,9,10
3	$\Delta = 6$, T =10, n =15	α =0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9
4	$\Delta = 6$, T =10, $\alpha = 0.3$	n =1,3,5,9,12,15,18,20

Figure 4, 5, 6 and 7 show the clustering accuracy under the setting of above 4 groups of parameters, respectively.

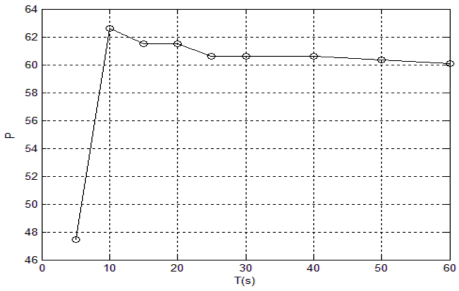

Figure 4. Clustering Accuracy with T

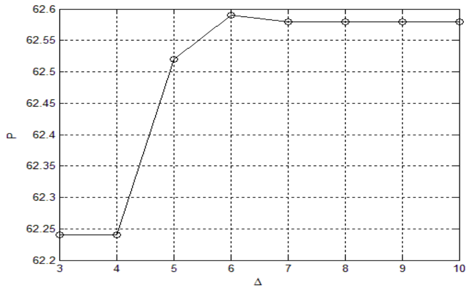

Figure 5. Clustering Accuracy with Δ

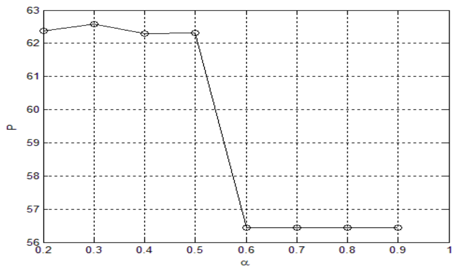

Figure 6. Clustering Accuracy with α

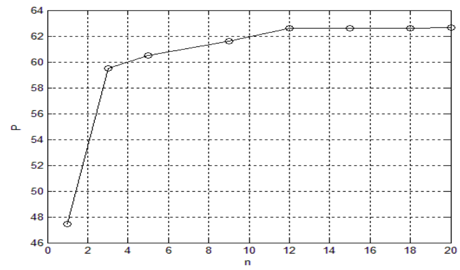

Figure 7. Clustering Accuracy with n

Figure 4 shows that with T increasing, the clustering accuracy first increases, then decreases, and finally flattens. That is because a smaller interval could split the same traffic flow into two segments, and a larger interval could merge two different flows into one segment, both of which increase the clustering error. In Figure 5, when the $\Delta = 6$, the clustering accuracy rate reaches the highest. In Figure 6 the clustering coefficient decreases fast when α is greater than 0.5. Figure 7 shows that the sample selection method works better than random selection method and with the increase of the sample number, clustering accuracy first improves and then tend to be stable.

In general, the traffic flows between 2 and 3 are clustered into 6 categories, and the 6 clustering centers, denoted by packet number, average packet length, and average packet interval, are shown in Table 3. Compared to the actual traffic flows, the corresponding application type could be single confirmation message, connection control flow or short message interaction flow, picture transmission flow, file transfer, video transmission, and interactive voice flow, respectively.

Table 3. The Results of Traffic Clustering

type	packet number	average packet interval	average packet length	corresponding application type
1	1	0	77	single confirmation message
2	10	0.35433	79.2	connection control
3	279	0.0095282	87.226	video transmission
4	11	0.0032541	575.623	file transfer or big picture
5	8	0.079918	88.567	short message interaction or small picture
6	20	0.14152	78.2	interactive voice

The timing relationship between the 6 kinds of traffic flows is shown in Table 4. Each matrix element (i, j) represents the probability of type i flow followed by type j flow. As the table shows, type 2 flow is probably followed by type 4 flow, while the type 4 flow is definitely followed by type 1 flow, and type 6 flow is probably followed by type 6 flow. When the monitor agent find there are large amount of other type flows, it can infer that there is some abnormity on node 2 or 3.

Table 4. Sequence Rule between 6 Type of Flows

	1	2	3	4	5	6
1	0	0.014085	0.056338	0.29577	0.042254	0.028169
2	0	0	0	**0.71429**	0.33333	0
3	0.10714	0	0	0	0.17857	0
4	**1**	0	0	0	0	0
5	0.5	0.4	0.1	0	0	0
6	0.40383	0.00638	0	0	0	**0.63333**

5. Conclusion

We put forward a traffic flow timing relationship analysis method in this paper. Firstly, we decompose the end-to-end flow into segments according to the interval and the packet length information. Secondly, we classify these segments into clusters based on the average frame size, frame interval and the flow length. Finally, we obtain the timing relationships between different kinds of traffic flows based on the Apriori algorithm. The simulation results show that this method can effectively classify the end-to-end traffic flows and give their timing relationships. Timing relationships could be used user abnormity detection.

References

[1] Tehrani Avissa Hosseini, Shahnasser Hamid. *Anonymous communication in MANET's, solutions and challenges*. 2010 IEEE International Conference on Wireless Information Technology and Systems (ICWITS). Honolulu. 2010; 1 - 4.

[2] Lianyu Zhao, Shen Haiying Helen. *ALERT: An Anonymous Location-Based Efficient Routing Protocol in MANETs.* 2011 International Conference on Parallel Processing (ICPP). Taipei. 2011; 703 - 712 .

[3] Ronggong Song, Tang H. LAA: Link-layer anonymous access for tactical MANETs. 2012 MILITARY COMMUNICATIONS CONFERENCE, Orlando. 2012; 1 - 7.

[4] Ting He, Ho Yin Wong, Kang-Won Lee. *Traffic Analysis in Anonymous MANETs.* Military Communications Conference (MILCOM 2008). San Diego. 2008;1-7.
[5] Yang Qin, Dijiang Huang. *A Statistical Traffic Pattern Discovery System for MANET*. Military Communications Conference (MILCOM 2009). Boston. 2009; 1-7.
[6] Jinsub Kim, Lang Tong. *Detection of Time-Varying Flows in Wireless Networks.* Military Communications Conference, San, Jose. 2010;2199-2204.
[7] Tameem Eissa, Shukor Abd Razak, MD Ngadi . Towards providing a new lightweight authentication and encryption scheme for MANET. *Wireless Networks*. 2011; 17(4): 833-842 .
[8] Poonam Gera, Kumkum Garg, Manoj Misra. *Trust based multi-path routing for end to end secure data delivery in manets*. SIN '10: Proceedings of the 3rd international conference on Security of information and networks. 2010; 81-89 .
[9] Soumyadev Maity, RC Hansdah. *Membership Models and the Design of Authentication Protocols for MANETs.* 2012 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA) . 2012; 544-551.
[10] Ryan Pries, Wei Yu, Xinwen Fu, Wei Zhao. *A New Replay Attack Against Anonymous Communication Networks* . IEEE International Conference on Communications. Beijing. 2008; 1578-1582.
[11] KA Rahman, KS Balagani, VV Phoha. Snoop-Forge-Replay Attacks on Continuous Verification With Keystrokes. *IEEE Transactions on Information Forensics and Security*. 2013; 8(3): 528-541 .
[12] AHM Rezaul Karim, RMAP Rajatheva, Kazi M Ahmed. *An efficient collaborative intrusion detection system for MANET using Bayesian Approach*. MSWiM '06: Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems. 2006; 187-190 .
[13] Yi-an Huang, Wenke Lee. *A cooperative intrusion detection system for ad hoc networks*. SASN '03: Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks 2003; 135 - 147 .
[14] V Manoj, N Raghavendiran, M Mohammed Aaqib, R Vijayan. *An approach for detection of malicious node using fuzzy based trust levels in MANET*. ACWR '11: Proceedings of the 1st International Conference on Wireless Technologies for Humanitarian Relief. ACM. 2011; 477-480.
[15] Adnan Nadeem, Michael Howarth. *Adaptive intrusion detection & prevention of denial of service attacks in MANETs.* IWCMC '09 Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly. ACM. 2009; 926-930.
[16] Gattal, Abdeljalil, Chibani Youcef . *Segmentation and Recognition Strategy of Handwritten Connected Digits Based on the Oriented Sliding Window*. International Conference on Frontiers in Handwriting Recognition (ICFHR). Bari, Italy. 2012; 297-301 .
[17] Won-Hee Kim, Tae-Il Jeong, Jong-Nam Kim. *Video segmentation algorithm using threshold and weighting based on moving sliding window*. 11th International Conference on Advanced Communication Technology. Phoenix Park. 2009; 1781-1784.
[18] Jiaojiao Wu, Bo Yin, Wenjuan Qi. Video Motion Segmentation Based on Double Sliding Window. 2011 Fourth International Symposium on Computational Intelligence and Design (ISCID). Hangzhou. 2011; 232 - 235.
[19] Yongsheng Pan. *Top-down image segmentation using the Mumford-Shah functional and level set image representation*. IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei. 2009; 1241-1244.
[20] Jiang-Ping Li, Yang SX, Gregori S. *Combining Top-Down and Ncut Methods for Figure-Ground Segmentation*. International Conference on Apperceiving Computing and Intelligence Analysis. ICACIA 2008. Chengdu. 2008; 216-219.
[21] N Vasconcelos, G Carneiro. *Weakly Supervised Top-down Image Segmentation*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006; 1001-1006
[22] S Alpert, M Galun, A Brandt, R Basri. Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012; 34(2): 315 - 327.
[23] Yongsheng Pan, JD Birdwell, SM Djouadi. *An Efficient Bottom-Up Image Segmentation Method Based on Region Growing, Region Competition and the Mumford Shah Functional*. IEEE 8th Workshop on Multimedia Signal Processing, Victoria, BC. 2006; 344-349.
[24] J Douglas Birdwell, Seddik M Djouadi, Yongsheng Pan. Efficient Bottom-Up Image Segmentation Using Region Competition and the Mumford-Shah Model for Color and Textured Images. *Eighth IEEE International Symposium on Multimedia*. San Diego, CA. 2006; 376-390.
[25] Yanhuang Jiang, Qiangli Zhao. Machine Learning Techniques. Beijing: Publishing House of electronics industry. 2009; 249-251.
[26] Tiemei Chen, Daoping Huang, Guxin Lu, etc. Study of pattern clustering in the data pretreatment of the pulp cooking. *Computers and Applied Chemistry*. 2003; 20(3); 241-243.
[27] Pang-Ning Pan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining (Ming Fan, Hongjian Fan). Beijing: People's Posts and Telecom Press. 2011; 201-237.