A Time-series Clustering Approach for Sybil Attack Detection in Vehicular Ad hoc Networks

Neelanjana Dutta
Department of Computer Science
Missouri University of
Science and Technology
Rolla, Missouri 65409
Email: nd2n8@mst.edu

Sriram Chellappan
Department of Computer Science
Missouri University of
Science and Technology
Rolla, Missouri 65409
Email: chellaps@mst.edu

***Abstract*—Sybil attack is a security threat wherein an attacker creates and uses multiple counterfeit identities risking trust and functionality of a peer-to-peer system. Sybil attack in vehicular ad hoc networks is an emergent threat to the services and security of the system. In the highly dynamic environment of vehicular ad hoc networks, due to mobility and density of nodes, it is challenging to detect the nodes that are launching Sybil attack. Existing techniques mostly use additional hardware or complex cryptographic solutions for Sybil attack detection in vehicular ad hoc networks. In this paper, we propose a fuzzy time-series clustering based approach that does not require any additional hardware or infrastructure support for Sybil attack detection in vehicular ad hoc networks. The proposed technique leverages the dispersion of vehicle platoons over time in a network and detects Sybil nodes as those which are traveling closely in a cluster for an unusually long time. Simulation results and analysis show that the approach is able to identify Sybil nodes with very low false positive and false negative rates even under varying intensity of attack.**

***Index Terms*—Sybil attack; vehicular ad hoc networks; platoon dispersion; fuzzy time series clustering.**

I. Introduction

Vehicular Ad hoc Network (VANET) is a type of ad hoc network that is comprised of vehicles and road transportation infrastructure. The application of VANETs in different emergency notification system, safety-related and infotainment purposes have increased over past few years, leading it to become the backbone of *Intelligent Transport System* (ITS). Alongside, new security threats in VANETs have been investigated as well [1], [2], [3]. In this paper a critical security problem, namely *Sybil attack*, has been addressed and a time-series clustering based approach is proposed for detection of nodes that are launching this attack in VANETs.

Sybil attack [4] is a security threat wherein an attacker creates and uses multiple counterfeit identities risking trust and functionality of a peer-to-peer system. Sybil attack in Vehicular Ad hoc Networks (VANET) is an emergent threat to the services and trust of the system. In the highly dynamic environment of a VANET, an attacker can easily create and use multiple fake identities, and exploit node mobility to exit the location of the attack. Consequently, detecting the presence of Sybil attack and identifying the Sybil nodes become a challenge considering the dynamic nature of vehicular networks, ephemeral neighborhood proximities and ad hoc mobility.

In this paper, we propose a fuzzy time-series technique to cluster mobile nodes' locations based on neighborhood proximity. The underlying principle behind our approach is as follows. As a Sybil node counterfeits multiple identities and presents them to the system, those *fake* vehicles (represented by the counterfeit identities) will generally be reported around the Sybil vehicle that uses the identities leading these vehicles to violate normal dispersion dynamics. The proposed technique leverages the dispersion of vehicle platoons over time in a network and detects Sybil nodes as those which are located closely in a cluster as they move for an unusually long time. Simulation results and analysis show that the approach is able to identify Sybil nodes with very low false positive and false negative rates even under varying intensity of attack.

A. Related Work

While Sybil attacks have been addressed in social networks, Wireless Sensor Networks and Mobile Ad Hoc Networks, solutions in these domains require long term observation, collaboration and verification which are not possible in ephemeral networks like VANETs, where associations are short and unlikely to repeat. However, there have been research for detection of Sybil attack and identification of Sybil nodes in VANETs as well. In [5], a physical signal characteristics based technique was discussed for Sybil node detection in VANETS. A pair of nodes could be distinguished from each other using estimate of relative node localization that gives an indication of the coherence of the received signal. A signal strength distribution based method for detection and localization of Sybil nodes is proposed in [6] too. In [7], authors propose to employ RSUs that issue temporally varying pseudonyms to vehicles near their vicinity. A cryptographic solution to the problem of Sybil attack detection is proposed in [8]. In [9], spatial and temporal correlation between vehicles and RSUs is used to detect Sybil nodes, exploiting the fact that two vehicles passing by multiple RSUs at exactly the same time is rare. In [10], the authors presented a general approach to validate the VANET data, even in the presence of a few Sybil nodes. Anomalies are detected by checking the validity of the VANET data with respect to the VANET model and adversarial model.

Existing techniques for Sybil detection in VANETs mostly require additional hardware and overhead, but they do not use the available network physics, physical infrastructure information and statistics. The Sybil node detection technique proposed in this paper does not need any external support or complex algorithms, but rather relies on leveraging a basic mobility feature of nodes in VANET - the dispersion of vehicle platoons over time, or platoon dispersion [11]. Platoon dispersion indicates that in normal conditions, vehicles in proximity of each other at a certain time are unlikely to sustain their proximity clustering over time, i.e., proximity clusters are ephemeral. Based on this premise, the proposed solution in this paper uses fuzzy time series clustering for detection of Sybil nodes. We incorporate data preprocessing and feature extraction phases to make the algorithm more efficient. We perform theoretical analysis and simulations to derive threshold parameters and demonstrate performance of the technique. We also take into consideration various intensities of attack, which the attacker can adopt by utilizing only a part of its available counterfeit identities at a time. Such a variation in attack model makes it all the more difficult to estimate consistent association of nodes with one another. Simulation results show that the proposed technique succeeds in identifying most of the Sybil nodes over a period of time under such conditions as well.

The rest of the paper is arranged as follows. In Section II, the problem addressed is formally defined. In Section III, our time-series clustering based solution is presented. Detailed discussion about data preprocessing and feature extraction phases, the clustering algorithm and theoretical analysis are parts of this section. Performance evaluation of the proposed method is presented in Section IV. Section V presents concluding remarks and future work.

II. Problem Definition

The network model, attack model and the problem addressed are defined in this section.

Network Model - The main components of the VANET are - vehicles, Road Side Units (RSUs) and Certification Authority (CA). Vehicles are alternatively referred to as “nodes” in this paper. Nodes in VANETs are equipped with On Board Units (OBUs) to communicate and compute messages. Nodes may also have sensors, navigation device or Global Positioning System (GPS), computing devices, display units, etc. Each node is aware of its own location and the map of the network area and usually communicates using short range wireless communication technology, such as Dedicated Short Range Communication (DSRC), bluetooth, etc. RSUs usually comprise of cheap embedded devices including sensors, smart traffic controllers, etc. RSUs store secure information such as its secure communication keys, traffic information, safety-related information etc. CA is a central authority, which authenticates vehicles and RSUs using the secure authentication infrastructure like public key infrastructure. Each node is given a unique identity or *ID* by the CA.

Attack Model - A Sybil node is defined to be one which uses multiple counterfeit identities to pretend to be some other node(s). As discussed in Section I, the benefits of the attacker in launching such attacks are multi-fold. A group of malicious nodes can subvert the trust and reputation system of the network if they conduct Sybil attack on the network for some time. Eventually this can deteriorate the overall performance of the system. In our model, we consider that a malicious vehicle, with original id V, has n different identities, V_n = V_0, V_1, ..., V_{n-1}. V can determine the intensity of attack by choosing to use only a certain percentage of the counterfeit identities at a time. Intuitively, using lesser number of ID’s at a time will lower its chance of getting detected, but at the same time it will mitigate the intensity of attack as well. In our model, V uses $x\%$ of these ids over a time duration Δt where $x \in [0, 100]$. It is assumed that V randomly selects i different ids from the set V_n such that $x = \frac{100i}{n}$ and uses them to communicate for the next Δt, and then again repeats the same process. We assume that the vehicles follow predefined speed limits on the roads.

In the very dynamic environment of a VANET, it is challenging to identify a Sybil node due to the high mobility and density of nodes. In other words, a node can escape one part of the network and reach another part very fast. The large number of nodes in a network makes it all the more difficult to identify malicious node(s). These challenges warrant the need of a lightweight and efficient approach to detect Sybil nodes in VANET. The objective of the paper is to propose an efficient method to detect Sybil nodes without using additional hardware or infrastructure support.

III. Proposed Solution

In this paper, a time-series clustering method called FSTS [12] is used to detect Sybil nodes in VANETs under varying attack intensity. Time-series clustering helps to identify the nodes that are moving in proximity of each other over a time period based on the location traces of the nodes. Because of the large number and density of nodes in a typical VANET, it is likely that a node can be part of multiple clusters at the same time, making fuzzy clustering algorithms suitable for the scenario. In this section we first discuss the location data collection method, followed by different steps of the proposed technique for Sybil node detection.

The key idea behind the proposed solution comes from a vehicular network phenomenon called *platoon dispersion* [11] as mentioned in Section I. A platoon is a group of vehicles traveling together. If all vehicles in an existing platoon maintain their speeds, a platoon will never disseminate. However, due to physical factors like road friction, vehicle characteristics and signaling, along with human factors like car following pattern, lane changes, fatigue, there is inherent randomness in driver behavior, and platoons tend to disperse over time. Intuitively, longer the travel time between points, greater is the dispersion, since there is more time for drivers to deviate from current speeds. We use this idea to derive a threshold probability P_{Th} of two vehicles being within a specified distance after a given time if their initial locations were same.

By the virtue of platoon dispersion, different vehicles in a network are not likely to travel together for very long. Towards this end, the threshold duration for which vehicles are likely to travel with each other can be estimated theoretically. If any two or more vehicles surpass this threshold duration, they are likely to be the same node faking identities as different nodes. The clustered time-series correspond to the identities of the vehicles which are likely to be Sybil nodes.

A. Location Data Collection by Peer Nodes

Standard DSRC communication allows vehicles to update its location and other physical parameters using periodic messages at a short, regular interval (usually 20 ms). However, in the scenario considered in this paper, any node can be a malicious Sybil node and it can also falsify its own location information to avoid detection. So, location data of vehicles over time is collected through peer vehicles through via messages or *report*. All nodes send *report* messages to the base station on a periodic basis in a fixed time interval. The purpose of *reports* is to inform the base station about the nodes which V_x has heard communicating in the last time interval. Because only a part of the nodes could be malicious, this collaborative process of reporting assures that the real location of a node is reported. For instance, if node V_x receives message from a node V_y at time t when V_x was at location l, it will incorporate this information in its next *report* to the base station. The location data of V_y collected by peers over time is represented in form of a time series $L_{V_y} = l_{V_y}(0), l_{V_y}(1), ..., l_{V_y}(t)$. It can be noted that the RSUs deployed along the road serve as local base stations that can execute the clustering algorithm and collaborate with each other as needed too.

B. Preprocessing Collected Data

After base station collects the location data from nodes in the network, all the following steps are executed by the base station for detection of Sybil nodes. Clustering algorithms are usually used for evenly distributed sampling for time-series, or can handle unevenly sampled data to some extant. But handling the ad hoc nature of data in VANET, specially when the Sybil node uses only a part of it's Sybil ID's at a time, becomes an orthogonal challenge. In simulation based experimentations it is feasible to collect data with regular sampling rates, but it is unlikely to do so in practical scenario. For instance, locations of V_x can be reported by peer nodes time instants t_0, t_1, t_5, t_9, t_{10} and t_{20} whereas locations of V_y can be reported by peer nodes time instants t_0, t_1, t_2, t_3, t_9, t_{11}, t_{12} and t_{15}. Clustering of these two time-series becomes due to the irregularity of sampling rate and size. In this paper, the effect of linear interpolation in time-series clustering of data is studied. Subsequently in Section III-D, a prediction technique is proposed to estimate locations of vehicles when no report is obtained. Although the time-series clustering algorithm used in this paper supports clustering of unevenly sampled time-series data, preprocessing of collected data is done for better results.

Linear Interpolation - Referring back to Section III-A, the time-series data for V_y can be represented as, $L_{V_y} = l_{V_y}(0), l_{V_y}(1), ..., l_{V_y}(t)$, where, $l_{V_y}(i) = (x_{V_y}(i), y_{V_y}(i))$. The linear interpolation between points $(x_{V_y}(i), y_{V_y}(i))$ and $(x_{V_y}(j), y_{V_y}(j)) \forall (i, j)$ and $(j - i) > 1$, can be given by,

$$y = y_{V_y}(i) + (x - x_{V_y}(i)) \frac{y_{V_y}(j) - y_{V_y}(i)}{x_{V_y}(j) - x_{V_y}(i)} \qquad (1)$$

The data points between $l_{V_y}(i)$ and $l_{V_y}(j)$ can be constructed on the line represented by Equation 1 at regular distances $\Delta d = \frac{||(l_{V_y}(i), l_{V_y}(j))||}{p-1}$, where $(j - i) = p$ and $||.||$ refer to Euclidean distance.

C. Estimation of Number of Sybil Nodes

Association rule mining is used as a basic feature extraction step in this paper, in order to have an idea about how many Sybil nodes are likely to be present in a part of network. Association Rule Learning mines relation between multiple attributes of an entity based on their frequency of co-occurrence in a dataset [13]. Let $I = i_1; i_2; i_3, ..., i_r$ be a set of r binary attributes called items. Let $\tau = \tau_1, \tau_2, \tau_3, ..., \tau_s$ be a set of s transactions called a database. Each transaction in τ contains a subset of the items in I. The problem here is to identify association rules in the database, which is an implication of the form $X \implies Y$, where $X, Y \in I$ and $X \bigcap Y = \emptyset$. Reverting back to Sybil detection, consider a Vehicle V_x that has communicated with peers over time. The dataset τ_x of V_x is a row of transactions with each time-stamped row consisting of vehicle ids with which V_x has communicated at that time. Recall from platoon dispersion that a group of vehicles is highly unlikely to be consistently associated geographically (i.e., as a platoon) over a long time period. When a consistent association of two or more vehicles is seen, those vehicles can be suspected to be Sybil. Using this technique, different peer nodes in a network can predict how many Sybil nodes are likely to be present in its vicinity and report to the base station. Also, the base station itself can use this technique to gauge which nodes could possibly be Sybil. However, it is not possible to draw a conclusion from their analysis when the Sybil node uses only a part of forged identities over time and changes them over next time period. This step is only useful for the base station to predict expected number of clusters, w_{ij}, which is an input to the clustering algorithm as discussed in Section IV-A.

D. Fuzzy Time-Series Clustering

Fuzzy time-series clustering involves fuzzy clustering of time-series data collected over time with even or uneven sampling rate. In this paper, *Fuzzy Short Time-Series* (FSTS) clustering of location traces of mobile nodes over a time period is used for Sybil attack detection. The proposed technique is based on the FSTS algorithm presented in [12]. It can be noted that the proposed short time-series based piecewise slope distance clustering seems intuitively appropriate for the application considered in this paper. The type of location data obtained from vehicles in a VANET can be enormous in size, but the Sybil detection technique deals with data over a comparatively shorter period of time. However, there are several differences in the two approaches. Firstly, in this paper, two-dimensional data (location) is considered for clustering over time. So, the time-series data considered is three dimensional unlike the two dimensional clustering performed in [12]. Besides, this technique is further extended in Section III-D to leverage the advantages of estimation techniques in the domain of time-series clustering.

Fuzzy short time-series (FSTS) technique proposed in [12] is a variation of fuzzy C-means clustering for time-series data. The basic idea is to perform a slope distance computation of time-series, which can be used for clustering the time-series in FSTS method. In this paper, the distance considered includes the three dimensional data (x and y coordinates of location and time) obtained from VANETs. For time-series of vehicle $l_{V_x} = l_{V_x}(0), l_{V_x}(1), ..., l_{V_x}(t_n)$, the linear function between $L_{V_x}(t)$ two consecutive time instants t_k and $t_{(k+1)}$ are defined as,

$$L_{V_x}(t) = m_k(t) + b_k, \tag{2}$$

where $t_k < t < t_{k+1}$, and

$$m_k = \frac{||l_{V_x}(k+1) - l_{V_x}(k)||}{t_(k+1) - t_k}, \tag{3}$$

$$b_k = \frac{t_(k+1) l_{V_x}(k+1) - t_k l_{V_x}(k)}{t_(k+1) - t_k} \tag{4}$$

Consequently, a set of equations can be derived as both x and y coordinate are separately considered in Equation 4. The short time-series distance between time-series vector of vehicle V_x and prototype vector V_y is computed as below -

$$\begin{aligned} d^2_{STS}(V_x, V_y) = \sum_{k=0}^{n_t - 1} \Bigg(& \frac{V_y(k+1) - V_y(k)}{t_{k+1} - t_k} \\ & - \frac{V_x(k+1) - V_x(k)}{t_{k+1} - t_k})^2 \end{aligned} \tag{5}$$

Rest of the FSTS algorithm is similar to fuzzy C-means algorithm . The cost function is defined as,

$$J(V_x, V_y, u) = \sum_{i=1}^{n_k} \sum_{i=1}^{n_v} u_{ij}^w d^2(V_x(j), V_y(i)), \tag{6}$$

where n_k is the number of clusters, n_v is the number of vehicles and w is the weight factor. All these values are user-defined. The value of u determines the membership value of the element in the cluster. Updating of the partition matrix is done in the same way as described in [12], where u_{ij}^w is updated as,

$$u_{ij}^w = \frac{1}{\sum_{q=1}^{n_k} (d_{STS_{ij}} / d_{STS_{qj}})^{\frac{1}{w-1}}} \tag{7}$$

Further details of this algorithm is abstracted in the current paper and can be found in [12].

E. Derivation of P_{Th}

In this section, the objective is to derive P_{Th}, the probability of two vehicles traveling in each other's vicinity so that the expected time of observation for Sybil node detection can be estimated. Towards this end, first theoretical analysis is performed to determine P_{Th} and then the outcome is tested using simulation studies.

Let us consider that two vehicles are moving on a straight road. They are initially (time $t = 0$) at a distance d_0 apart. In a time interval δt, the vehicles can move any distance within a range of D_H and D_L on the road. The range is represented as D_{range}. At every time instance the vehicles update their velocities based on past velocities and thus the distances to be covered (denoted by D_1 and D_2) in next time interval, δt. D_1 and D_2 are chosen from D_{range} using uniform distribution. Our initial objective is to figure out the probability that the two vehicles are within a distance α of each other after a time interval $n\delta t$.

As mentioned above, we assume uniform distribution for D_1 and D_2. For simplicity of computation, we assume d_0 = 0 throughout this derivation. Now, using normal approximation of uniform distribution, if $D_1 \sim Unif(D_H, D_L)$ and $D_2 \sim Unif(D_H, D_L)$, then $\sum_{i=1}^{n} D_{1i} \sim N(\frac{n(D_L+D_H)}{2}, \frac{n(D_H-D_L)^2}{12})$ and $\sum_{i=1}^{n} D_{2i} \sim N(\frac{n(D_L+D_H)}{2}, \frac{n(D_H-D_L)^2}{12})$.
So, $(\sum_{i=1}^{n} D_{1i} - \sum_{i=1}^{n} D_{2i}) \sim N(0, \frac{2n(D_H-D_L)^2}{12})$.

Now, the probability that the condition $|\sum_{i=1}^{n} D_{1i} - \sum_{i=1}^{n} D_{2i}| \leq \alpha$ holds true can be written as

$$P(|\sum_{i=1}^{n} D_{1i} - \sum_{i=1}^{n} D_{2i}|) \tag{8}$$

$$\begin{aligned} &= P(-\alpha \leq \sum_{i=1}^{n} D_{1i} - \sum_{i=1}^{n} D_{2i} \leq \alpha) \\ &= P(\frac{-\alpha - 0}{\sqrt{\frac{2n(D_H-D_L)^2}{12}}} \leq Z \leq \frac{\alpha - 0}{\sqrt{\frac{2n(D_H-D_L)^2}{12}}}) \\ &[where\ Z = \sum_{i=1}^{n} D_{1i} - \sum_{i=1}^{n} D_{2i}] \\ &= P(-z \leq Z \leq z) \\ &\qquad [where\ z = \frac{\alpha}{\sqrt{\frac{2n(D_H-D_L)^2}{12}}}] \end{aligned} \tag{9}$$

Using standard normal distribution of Z, i.e., $\Phi(Z)$, it is evident that, $\Phi(Z) = P(Z \leq z)$. So, our probability expression, (in equation 9) = 2 $\Phi(Z)$-1.
Using the standard normal CDF table, the probability for different values of D_H, D_L, n and α can be found out. From this derivation, it is straight forward to derive the expected time, t_{exp}, that two vehicles will take to reach a threshold probability P_{th} that they are traveling in each other's vicinity. It can be noted that in real life, based on several physical and human factors, any other distribution other than uniform

distribution can be used to model vehicle's distance traveled over a time period. However, similar derivation can be done using other probability distributions too.

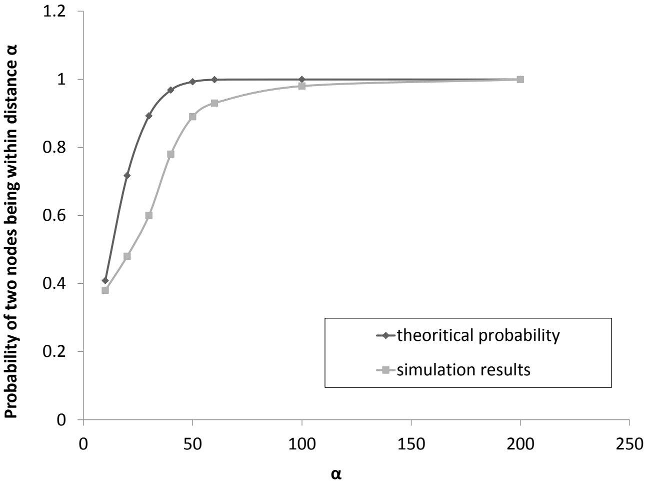

Fig. 1. Determination of Input Parameters where D_H = 50 m, D_L = 0 m, n = 10, δt = 1 s

A theoretical probability of two nodes moving within a given distance over a time period can be obtained by plugging in values of different input parameters into the expression derived above. In Figure 1, probability values derived through theoretical analysis and simulation results are plotted against different values of α where D_H = 50 m, D_L = 0 m, n = 10, δt = 1 s. This figure shows a case where simulation data is plotted along with theoretical results to show that the results match closely. Thus from this derivation, for a given time period, the probability of two vehicles being in a same cluster (or within a given distance) for a given time period can be obtained. For different experiments performed with different values of network parameters (like D_H, D_L etc.), we derived the probability threshold for which two nodes can be in a cluster for a given time duration. If the output of FSTS algorithm yielded a higher cluster membership than the probability threshold derived, the node in the cluster are detected as Sybil nodes. Derivation of threshold parameter through this process helped us differentiate among nodes traveling together for long time and malicious Sybil nodes.

IV. Performance Evaluation

In this section, the performance of the proposed technique is presented. Most of the work for this section is still ongoing and only preliminary results are presented. So far, the main effort in this section has been in developing a customized simulator for data collection, interpolation and association rule mining based estimation, and FSTS clustering.

A. Experimental Setup

SUMO (Simulator of Urban Mobility) was used to generate mobility traces of nodes and this data was used as input to the network. A C++ simulator is developed to emulate the vehicular network where the nodes move following the mobility traces from SUMO, thereby mimicking real traffic patterns. The final clustering experiments are done using the C++ simulator. By default, there were 100 vehicles with an average speed of 50mph, and sources and destinations were randomly chosen for each vehicle. There were 10 Sybil nodes among them, and each had 10 identities. Each vehicle was assumed to report it's location once every second, and the transmission range was assumed to be 250 m. The simulation was run for 1000 seconds and the default clustering distance was 400 m. All simulations were conducted 10 times and results were averaged.

Different sets of experiments were run in different phases. First the collected time-series data is preprocessed using linear interpolation using Matlab and then association rule mining is used for feature extraction phase estimating expected number of clusters in the data using WEKA (Waikato Environment for Knowledge Analysis). For the first phase of the study with association rule mining, each Sybil node used all its counterfeit identities during query response. Later the cases were studied when only a smaller percentage of identities are used by a node during a time period.

Apriori algorithm implemented in WEKA tool was used as feature extraction technique to estimate number of abnormally repeating associations or clusters in the dataset of each vehicle. The success rate was 100% in Sybil vehicle detection without false positives. However when the percentage of ID's used by the Sybil node varied, only 60% of the Sybil nodes were detected and equal number non-Sybil nodes were detected as Sybil nodes. It means that the false positive and true positive rates were equal, which is not a desired performance. Clearly there is need of further analysis, which is conducted subsequently. However, several association rule experiments help get an *feel* or estimate of how many clusters to look for and the probable number of Sybil nodes in a set of nodes. For instance, in the case mentioned above, the results of feature extraction show that there are likely to be 12 clusters. In reality, there were 10 clusters that had Sybil nodes in them in that case. Hence in our experiment, we put the input number of clusters between 9 and 15, getting the best results when the number of clusters was 10. It can be noted that usually all clustering algorithm (including FSTS) require preprocessing and feature extraction of data or some sort of prior knowledge to estimate number of clusters. However, the results from association rule mining are not conclusive, warranting further experiments using the FSTS technique to determine the Sybil nodes from past location traces.

B. Clustering of Data

Recall from Section III-E, theoretical analysis can be used to derive P_{Th} for different input parameters and the output can be used to determine whether the concerned nodes are Sybil or not based on their cluster membership values determined using FSTS. In the clustering process, firstly the binary connection metric is clustered using the FSTS algorithm. Figure 2 shows the detected number of false positives and false negatives aver-

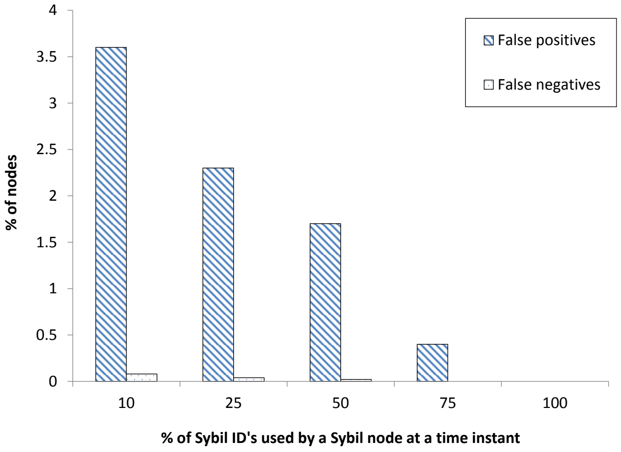

Fig. 2. False positive and false negative rate for varying percentage of Sybil ID's used by a Sybil node at a time instant

aged over 10 runs of simulation each. The X axis represents the percentage of available fake IDs that a Sybil node is using at a time instant. As mentioned earlier in this section, there were 10 Sybil nodes using 10 Sybil IDs each and there were 100 vehicles in total in the simulation setup. If all of the available IDs are used for transmission at every time instant, the false positive and false negative rates are both zero, indicating that all the Sybil nodes are identified. However, as the percentage decreases, both false positives and false negatives increase, although a major part of the Sybil nodes are detected over time. The reason behind the increase in false positives and false negatives is that with lesser number of counterfeit IDs being used at a time, it becomes increasingly difficult to cluster such IDs. This figure demonstrates the effectiveness of the proposed technique in detecting Sybil nodes in VANETs.

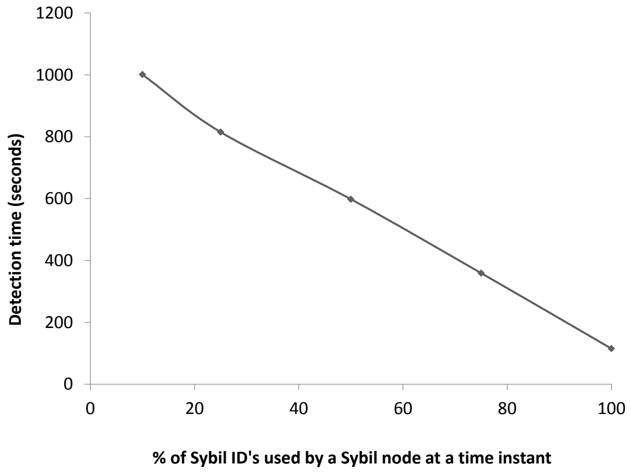

Fig. 3. Detection time in seconds for varying percentage of Sybil ID's used by a Sybil node at a time instant

Figure 3 plots the time required to reach 100% true positive rate (that is, detects all Sybil nodes) for varying percentage of ID's used by the Sybil nodes at a time instant. With increasing percentage of ID's used, the detection is faster. But as very less percentage of ID's are used by a Sybil node at a time instant, it still reaches 100% true positive rate in longer time. Clearly, this trend is exhibited due to the fact that with lesser number of counterfeit IDs being used at a time, clustering of such IDs become increasingly harder resulting in longer detection time.

V. Conclusions

This paper proposes a technique for Sybil attack detection in VANETs, based on fuzzy time-series clustering. The method leverages the principle of dispersion of vehicle platoons in a VANET and detects the nodes clustered with each other for longer than expected. Theoretical analysis has been conducted to derive input parameter to the algorithm and simulation results are presented to evaluate performance of the proposed method. The proposed method has achieved very low false positive and false negative rates even when the Sybil nodes use a small percentage of the counterfeit identities at a time instant. Future work involve derivation of threshold parameters considering different mobility models of vehicles and further investigation of platoon dispersion models to incorporate physical and human factors into the current analysis.

References

[1] J. T. Isaac, S. Zeadally, and J. Camara, "Security attacks and solutions for vehicular ad hoc networks," *Communications, IET*, vol. 4, no. 7, pp. 894–903, April 2010.

[2] L. Cheng and R. Shakya, "Vanet worm spreading from traffic modeling," in *IEEE Radio and Wireless Symposium (RWS)*, 2010, pp. 669–672.

[3] N. Dutta, R. Kotikalapudi, and M. Bhonsle, "A formal analysis of protocol-independent security threats in vanets," in *IEEE Students' Technology Symposium (TechSym)*, 2011.

[4] J. R. Douceur, "The sybil attack," in *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, 2002, pp. 251–260.

[5] M. S. Bouassida, G. Guette, M. Shawky, and B. Ducourthial, "Sybil nodes detection based on received signal strength variations within vanet." *International Journal on Network Security*, vol. 9, no. 1, pp. 22–33, 2009.

[6] B. Xiao, B. Yu, and C. Gao, "Detection and localization of sybil nodes in vanets," in *Workshop on Dependability issues in Wireless Ad hoc Networks and Sensor Networks*, 2006, pp. 1–8.

[7] T. Zhou, R. Choudhury, P. Ning, and K. Chakrabarty, "P2DAP- Sybil Attacks Detection in Vehicular Ad Hoc Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 3, pp. 582 – 594, 2011.

[8] M. Rahbari and M. Jamali, "Efficient detection of sybil attack based on cryptography in vanet," *International Journal of Network Security and Its Applications*, vol. 3, 2011.

[9] S. Park, B. Aslam, D. Turgut, and C. Zou, "Defense against sybil attack in vehicular ad hoc network based on roadside unit support," in *Military Communications Conference, 2009. MILCOM 2009. IEEE*, October 2009, pp. 1 – 7.

[10] P. Golle, D. Greene, and J. Staddon, "Detecting and correcting malicious data in vanets," in *Proceedings of the ACM First International Workshop on Vehicular Ad Hoc Networks*, 2004, pp. 29–37.

[11] R. Denney Jr, "Traffic platoon dispersion modeling," *Journal of transportation engineering*, vol. 115, pp. 193 – 207, 1989.

[12] C. Mller-Levet, F. Klawonn, K. Cho, and O. Wolkenhauer, "Fuzzy clustering of short time-series and unevenly distributed sampling points," *Advances in Intelligent Data Analysis V, Lecture Notes in Computer Science*, vol. 2810, pp. 330–340, 2003.

[13] R. Agrawal and T. Imielinski, "Association rules between sets of items in large databases," in *Proceedings of ACM SIGMOD*, 1993, pp. 207 –216.