Secure Connected Vehicle-based Traffic Signal Systems Against Data Spoofing Attacks

Tianye Ma
University of Delaware
Newark, DE, USA
matianye@udel.edu

Rui Zhang
University of Delaware
Newark, DE, USA
ruizhang@udel.edu

Mark Nejad
University of Delaware
Newark, DE, USA
nejad@udel.edu

2021 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-7281-9505-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/WCNC49053.2021.9417524

***Abstract*—The emerging Connected Vehicle (CV) technology is widely expected to greatly enhance traffic safety and efficiency by enabling vehicles, pedestrians, and infrastructures to communicate with one another. As a promising CV application, CV-based traffic signal control aims to improve the traffic efficiency at intersections by dynamically optimizing traffic signal control plans based on the mobility information submitted by surrounding CVs. Effective CV-based traffic control relies on accurate estimation of the queue length i.e., the number of vehicles waiting at intersections, to determine the optimal traffic signal control plans. Despite significant efforts on accurate queue length estimation, the robustness of queue length estimation has so far received very limited attention. A recent study has demonstrated that it is possible for malicious CVs to significantly manipulate the queue length estimation by reporting false mobility data, which can cause severe traffic congestion. To tackle this challenge, we introduce a robust queue length estimation mechanism that first utilizes the mobility data reported by all the CVs waiting in the queue to calculate multiple preliminary queue length estimates. Then, the robust statistical methods are adopted to derive a resulting estimated queue length whose accuracy is kept at an acceptable level even though there exist multiple malicious CVs in the queue. The simulation results confirm the effectiveness of the proposed mechanism.**

***Index Terms*—Security, Connected Vehicles, Intelligent Transportation Systems, Data Spoofing Attack**

I. Introduction

Connected vehicle (CV) technology is widely expected to greatly improve traffic efficiency and safety by enabling vehicles to communicate with other vehicles, transportation infrastructures, and pedestrians. The CV-based traffic signal control is one of the emerging CV applications, which relies on wireless communication between CVs and traffic control infrastructures to reduce congestion and improving traffic mobility at road interactions. In a CV-based traffic signal control system, vehicles equipped with communication capabilities periodically report their speed, location, heading, etc. to the infrastructures via the dedicated short-range communications (DSRC), and the traffic control system determines the optimized traffic signal plans according to the current traffic conditions at intersections.

The queue length at a signalized intersection, i.e., the number of vehicles waiting in line, is one of the most crucial parameters for determining optimal traffic signal control plans. In particular, the optimal traffic signal plan is largely affected by the estimated queue length, as the traffic signal control system needs to allocate sufficient time for the waiting vehicles to pass the intersection. In the absence of CV-technology, the queue length is currently estimated with the assistance of vehicular detectors, such as inductive loop, video cameras, and microwave sensors [1]. They not only incur high maintenance costs but also fail to produce accurate estimates during heavy traffic jams or bad weather. In contrast, the CV-based traffic signal control system can operate normally under oversaturated traffic flow conditions and low visibility conditions.

The current low market penetration rate of CVs makes queue length estimation a challenging problem. Ideally, if all the vehicles waiting at the intersection are equipped with CV technologies, queue lengths would be easily obtained by counting the number of CVs in queues. However, as the market penetration of CV-based vehicles remains low and is not expected to reach 0.95 before 2045 [2], queue length estimation needs to be based on the reported data from sporadic CVs. Consider as an example the Intelligent Traffic Signal System (I-SIG), which is an arterial traffic signal application developed in the Dynamic Mobility Applications (DMA) program launched by the USDOT [3]. The I-SIG adopts the Estimation of Location and Speed (EVLS) algorithm to estimate the trajectory data of the non-connected vehicles [4]. Queue length needs to be estimated in this process. To this end, the EVLS algorithm utilizes the information of stopping positions and stopping times reported by the last two CVs in the queue [5].

The overreliance on the stopping position of the last connected vehicle for queue length estimation makes CV-based traffic signal control systems vulnerable to data spoofing attacks. In particular, a recent study [4] demonstrates that even a single malicious CV can deceive the CV-based traffic control system, I-SIG, into accepting a significantly inflated queue length through reporting false mobility data. The inflated estimated queue length can cause the I-SIG to allocate unnecessarily a long period for the lane with the malicious CV and cause congestion or disrupt traffic flow at intersections, resulting in worse traffic mobility than that without using the I-SIG system. There is thus a pressing need for developing a robust queue length estimation mechanism resilient to data spoofing attacks to fully unleash the potential of CV-based traffic signal control.

In this paper, we tackle this challenge by introducing the design and evaluation of a robust queue length estimation

mechanism for CV-based traffic signal control system. We observe that the key to thwarting data spoofing attacks is to fully utilize the mobility data of all the available CVs waiting in the queue instead of the last one alone. Specifically, our mechanism estimates the queue length based on each individual CV's report and then aggregates multiple estimates to produce a final estimated queue length using robust statistical methods. Our contributions in this paper can be summarized as follows.

- We introduce a novel robust queue length estimation mechanism against data spoofing attacks for CV-based traffic signal control systems.
- Detailed simulation studies confirm the effectiveness of the proposed mechanism. For example, our mechanism can reduce the capability of the attacker in terms of skewing the resulting estimated queue length by 86.6%, 79.3%, and 70.4%, when the number of attacking CVs in the queue is 1, 2, and 3, respectively.

The rest of the paper is structured as follows. We review the related work in Section II and introduce the problem formulation in Section III. We then present the proposed mechanism in Section IV and report the simulation result in Section V. This paper is finally concluded in Section VI.

II. Related Work

As a serious threat to CVs and intelligent transportation systems, data spoofing attacks have drawn growing attention in recent years. Besides the position spoofing attack studied in [4], the impact of arrival time spoofing attacks on different backpressure-based scheduling algorithms in traffic signal control (TSC) was studied in [6]. In addition, Dedinsky *et al*. [7] introduced a vision system against the data spoofing attacks by monitoring the position of incoming vehicles and verifying their behaviors. Moreover, Ta and Dvir [8] presented a secure traffic congestion detection and management system to defend against data spoofing attacks that using a vehicular public key infrastructure. Li *et al*. [9] designed a blockchain-based and decentralized architecture to secure the CV-based traffic signal control systems. None of these works consider robust queue length estimation.

Besides data spoofing attacks targeting traffic control systems, the vulnerabilities of connected/autonomous vehicles (C/AV) have been exploited to attack a platoon of vehicles [10], [11] or a single vehicle [12], [13]. Amoozadeh *et al*. [10] and Abdo *et al*. [11] studied different security attacks on Cooperative Adaptive Cruise Control, which can affect a group of vehicles. Sun *et al*. [12] explored the vulnerability of current LiDAR-based perception architectures in AVs and perform the LiDAR spoofing attack. Shen *et al*. [13] showed that the Multi-Sensor Fusion (MSF) algorithms in AVs are vulnerable to the strategically performed GPS spoofing attacks. These works address different problems and are thus orthogonal to our work.

The subject of queuing at signalized intersections has been studied extensively in the past. As early as the 1940s and 1950s, Clayton [14], Wardorp [15], and Beckmann *et al*. [16] discussed the queues at fixed-cycle traffic light. The inductive-loop detector was introduced in the early 1960s and has become the most widely used traffic sensor [17]. The traffic flow data collected by inductive-loop detectors have been used to estimate the queue length at intersections [18], [19]. Recently, the emerging ITS and CV technologies have given rise to new mechanisms for queue length estimation. There have been two major approaches for queue length estimation in CV, including shockwave theory approach [20]–[25] and statistical approach [26]–[29]. Our proposed queue length estimation mechanism belongs to the statistical approach. Comert and Cetin [26] proposed to use the location information of the last CV in the queue to estimate the queue length. Tiaprasert *et al.* [27] applied the least-mean-square-error (LMSE) method estimate the queue length. However, none of these solutions can withstand the data spoofing attack addressed in this paper.

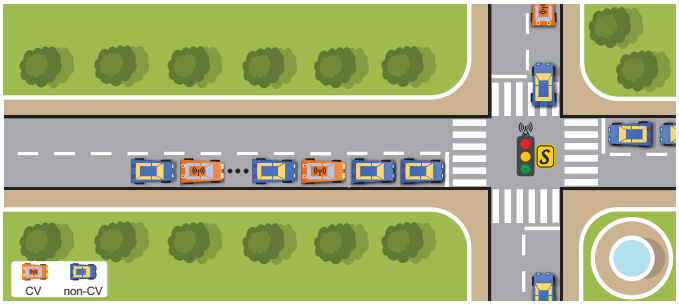


Fig. 1. Illustration of a CV-based traffic control system at a four-arm intersection.

III. Problem Formulation

A. System Model

We consider a CV-based traffic signal system at a four-arm intersection shown in Fig. 1. The CV-based traffic signal system periodically receives mobility report from nearby CVs whereby to estimate the queue length of each lane to determine the signal control plans. We focus on queue-length estimation in this paper, and how to determine the optimal traffic signal plan based on the estimated queue length is out of the scope of this work.

We assume that the time is divided into epochs of the same length. Our subsequent discussion considers a single lane at a given epoch t. Assume that there are l vehicles in the lane waiting after the stop line, including m CVs denoted by $V_1, \ldots, V_m$, and $l - m$ non-connected vehicles. Each CV V_i periodically broadcasts a status report containing its location and velocity at a frequency of 10Hz. We denote the status report broadcasted by CV V_i during epoch t by $R_{i,t} = \langle ID_i, p_i, v_i, t \rangle$, where ID_i is the unique ID of V_i assigned by a trusted authority, e.g., the DMV, and p_i and v_i are the position and velocity of V_i in epoch t, respectively. We assume that the length of each epoch is sufficiently small, e.g., 100ms, such that each CV only broadcasts one status report. In addition, for the CVs that are waiting in the queue, its speed is typically low, e.g., lower than 2m/s, and the distance it travels

during each epoch is negligible. Moreover, we assume that every beacon message is digitally signed by the sender with its private key to ensure the integrity of the message. We also assume that the current CVs' market penetration rate $\rho \in (0, 1)$ is known to the traffic control system. Given the status reports from the m CVs, $R_{1,t}, \ldots, R_{m,t}$, the traffic control system $\mathcal{S}$ intends to produce an estimated queue length $\hat{l}$.

B. Adversary Model

We consider an adversary whose goal is to deceive the traffic control system into producing an inflated estimated queue length whereby to cause suboptimal traffic signal plans and significant traffic congestion in other lanes and directions. In particular, the prior study has shown that it is much easier for the adversary to inflate the estimated queue length using forged mobility reports containing fake locations that are far away from the stop line. In contrast, it is much difficult for the adversary to mislead the traffic control system into significantly underestimating the queue length as long as there is at least one legitimate CV reporting a location behind the fake CVs.

We assume that the adversary has control over $c > 0$ attacking CVs which may launch data spoofing attack by submitting forged mobility reports under its instruction. A forged mobility report may contain a fake location of the adversary's choice but must include a valid CV's ID and appropriate digital signature. The adversary may launch the attack in different ways. First, it may have the CVs under its control be physically present at the target lane. Second, it may use a mobile device with a powerful transmitter to impersonate the CVs and send forged mobility reports on their behalf from a nearby location. In both cases, the adversary has the valid security credentials, e.g., private keys issued to the IDs, and can send mobility reports with proper digital signatures that can pass the verification at the traffic control system, and we thus will not differentiate the two cases hereafter.

C. Design Goals

We seek to design a robust queue length estimation mechanism to meet the following goals.

- *Resilience against data spoofing attack*: The estimated queue length should be sufficiently accurate in the presence of data spoofing attacks.
- *Accuracy in the absence of attack*: The queue length estimated by the proposed mechanism should be close to existing solutions in the absence of data spoofing attacks.

IV. Robust Queue Length Estimation

In this section, we present a novel robust queue length estimation mechanism that is resilient to the data spoofing attacks.

A. Overview

We observe that the vulnerability of existing queue length estimation techniques to data spoofing attacks stems from their reliance on the reported position of the last CV in the queue. In particular, the estimated queue length is largely affected by the position of the last CV, which can be easily manipulated by even a single malicious CV. To achieve robust queue length estimation against data spoofing attacks, it is thus important to minimize the impact of the last CV's position. Based on this observation, our mechanism estimates the queue length from each individual CV report based on the CV's reported location, its ranking among all CVs, and the total number of CVs using maximum likelihood estimation. The estimated queue length from individual CV report is thus not affected by the last CV's location. Given a set of estimated queue lengths, we then compute a final estimated queue length by aggregating them using a robust estimator that is resilient to outliers.

In what follows, we detail to the two phases of the proposed mechanism.

B. Queue Length Estimation from Individual CV Report

In this subsection, we introduce how to estimate the queue length based on a single CV's mobility report through maximum likelihood estimation.

First, we estimate the rankings of each CV among all the CVs and among all the vehicles based on the m reports $R_{1,t}, \ldots, R_{m,t}$. Specifically, we sort all the CVs according to their distances to the stopping lines. For each report $R_{i,t} = \langle ID_i, p_i, v_i, t \rangle$ received in epoch t, we first compute the distance between its position p_i and the stopping line as d_i. Without loss of generality, assume that $d_1 < d_2 < \cdots < d_m$. It follows that CV V_i is ranked ith among all CVs. Moreover, we estimate CV V_i's ranking in the queue as

$$r_i = \left\lfloor \frac{d_i}{h} \right\rfloor ,$$

where h is the empirical value of the space headway, which is the average distance between the front bumpers of two successive vehicles and equals to the length of a vehicle plus the gap between two successive vehicles.

Second, for each CV V_i, we estimate a queue length $\hat{l}_i$ based on CV V_i's rankings of each CV among all the CVs and among all the vehicles. Let R_c and R be the random variables representing the rankings of a CV among all the CVs and among all vehicles, respectively. Also let L and M be the random variables representing the queue length and the total number of CVs in the queue, respectively. Assume that each vehicle in the queue is equally likely to be a CV with probability ρ, i.e., the penetration rate. Observing a CV with a ranking $R_c = i$ among all the CVs and a ranking $R = r_i$ among all the vehicles is equivalent to the event that there are $i - 1$ CVs out of the first $r_i - 1$ vehicles, the r_ith vehicle in the queue is a CV, and $m - i$ out of the last $l - r_i$ vehicles are CVs. The likelihood of the event given there are l vehicles waiting in the queue is given by

$$\begin{aligned}\mathsf{L}(R_c = i, R = r_i | M = m, L = l) \\ &= \binom{r_i - 1}{i - 1} \rho^{i-1} (1-\rho)^{r_i - i} \rho \\ &\quad \cdot \binom{l - r_i}{m - i} \rho^{m-i} (1-\rho)^{l - r_i - (m-i)} \\ &= \binom{r_i - 1}{i - 1} \binom{l - r_i}{m - i} \rho^{m} (1-\rho)^{l-m} .\end{aligned} \tag{1}$$

Further define the likelihood function as

$$f_i(l) = \binom{r_i - 1}{i - 1} \binom{l - r_i}{m - i} \rho^{m} (1-\rho)^{l-m}. \tag{2}$$

The queue length estimated from report R_i is then given by

$$\hat{l}_i = \underset{l \in \{r_i, \ldots, l_{\max}\}}{\arg\max} f(l) \,, \tag{3}$$

where $l_{\max}$ is the maximum queue length determined by the physical road condition known in advance. The problem in Eq. (3) can be solved using either exhaustive search or the Newton's method.

C. Final Queue Length Estimate via Huber's M-Estimators

Given estimated queue lengths $\hat{l}_1, \ldots, \hat{l}_m$, we compute a final estimated queue length using a robust estimator. The most intuitive way is to estimate the queue length as the mean of the m estimated queue lengths $\hat{l}_1, \ldots, \hat{l}_m$. However, the mean is not a robust measure and can be easily affected by a small number of outliers, which makes the estimates vulnerable to data spoofing attacks.

We choose the location M-estimator with Huber's function Ψ [30] to produce a final estimated queue length. Specifically, given m estimates of the queue length $\hat{l}_1, \ldots, \hat{l}_m$, we compute the final estimated queue length $\hat{l}$ as the Huber's M-estimator [31], which is a robust estimator that generalizes sample mean and sample median. Specifically, the Huber's M-estimator of $\hat{l}_1, \ldots, \hat{l}_m$ is the solution of the following problem

$$\sum_{i=1}^{m} \Psi \left(\frac{\hat{l}_i - \hat{l}}{\sigma} \right) = 0 \tag{4}$$

where function Ψ is defined as

$$\Psi(x) = \begin{cases} K & \text{if } x > K, \\ x & \text{if } |x| \le K, \\ -K & \text{if } x < -K, \end{cases} \tag{5}$$

$K > 0$ is a factor which can be adjusted to balance the accuracy and robustness of the estimate [32], and σ is a robust measure of statistical dispersion.

A typical choice is the Normalized Median Absolute Deviation about the median (MADN). Specifically, given $\hat{l}_1, \ldots, \hat{l}_m$, the Median Absolute Deviation about the median (MAD) is defined as

$$\mathsf{MAD}(\hat{l}_1, \ldots, \hat{l}_m) = \mathsf{Median}(|\hat{l}_1 - \hat{l}_{\text{med}}|, \ldots, |\hat{l}_m - \hat{l}_{\text{med}}|). \tag{6}$$

where $\hat{l}_{\text{med}}$ is the median of $\hat{l}_1, \ldots, \hat{l}_m$. The Normalized MAD (MADN) is then defined as

$$\mathsf{MADN}(\hat{l}_1, \ldots, \hat{l}_m) = \frac{\mathsf{MAD}(\hat{l}_1, \ldots, \hat{l}_m)}{\Phi^{-1}(\frac{3}{4})} \tag{7}$$

where $\Phi^{-1}(\cdot)$ is the quantile function for the standard normal distribution, and $\Phi^{-1}(\frac{3}{4}) \approx 0.6745$ is the MAD of a standard normal random variable [33].

The Huber's M-Estimators generalizes both mean and median. In particular, $\hat{l}$ would be the median and mean of $\hat{l}_1, \ldots, \hat{l}_m$ if $K = 0$ and ∞, respectively. Note that there is no closed-form expression for $\hat{l}$ in Eq. (4), and we use the Newton's method to compute $\hat{l}$.

V. Simulation Evaluation

In this section, we evaluate the performance of the proposed queue length estimation mechanism via detailed simulation studies using MATLAB R2019b.

A. Simulation Settings

We compare the several variants of the proposed mechanism with a *Baseline* mechanism [3], which is the I-SIG system evaluated in [4]. As mentioned earlier, the I-SIG system estimates the queue length using the last CV's stopping location. We subsequently refer to *Huber K=1* and *Huber K=2* as the proposed mechanism with factor K set to 1 and 2, respectively. In addition, we use *Mean* and *Median* to denote the two special cases of the Huber's M-estimator with the factor K set to ∞ and 0, respectively. In our simulations, we set the maximum queue length $l_{\max}$ to 50, which serves as the farthest stopping position any attacking CV can report. Without such constraint, the attacker can report a false stopping position that is infinitely far away from the stopping line, which is impractical.

We use the Mean Absolute Percentage Error (MAPE) to evaluate the performance of the different queue length estimation methods, which is defined as

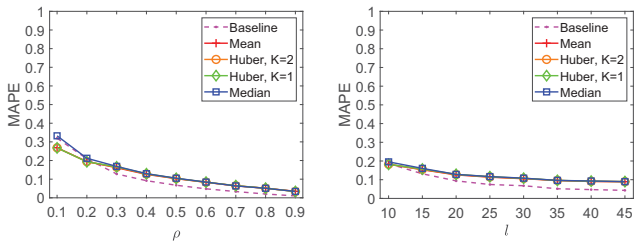
$$\mathsf{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{l}^i - l|}{l},$$

where n is the total number of runs, and l and $\hat{l}^i$ are the ground truth queue length and the estimated queue length of the ith run, respectively.

For each run of the simulation, we first randomly choose the $m - c$ vehicles out of the total l vehicles and then enumerate all possible combinations of the remaining c attacking CVs' positions to find the one that leads to the largest Absolute Percentage Error.

B. Simulation Results

We now report the simulation results where every data point represents the average of 1,000 runs unless mentioned otherwise.



(a) $l = 20$ (b) $\rho = 0.4$

Fig. 2. Comparison of the five queue length estimation methods with different market penetration rates of CVs ρ and different ground truth queue length.

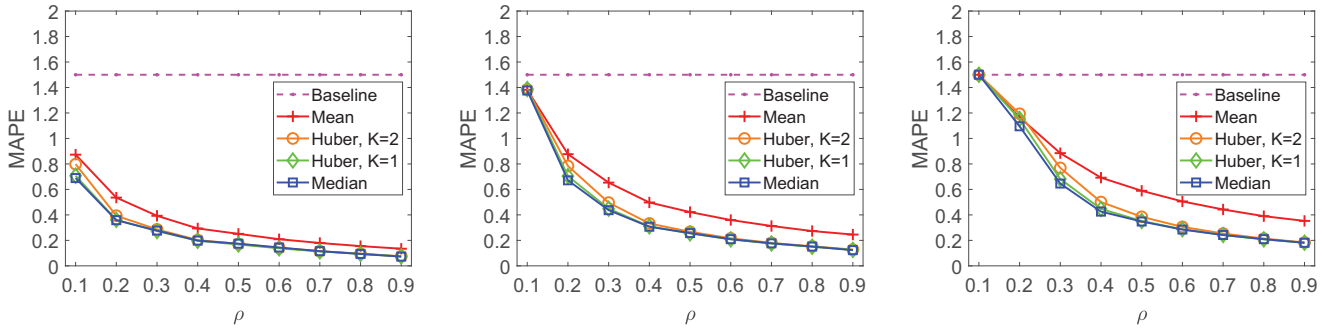
1) Performance in the Absence of Attack: Fig. 2a shows the MAPEs of the five queue length estimation methods where the actual queue length $l = 20$ and CVs' market penetration rate ρ varies from 0.1 to 0.9, and Fig. 2b shows the MAPEs of the five queue length estimation methods when CVs' market penetration rate $\rho = 0.4$ and the actual queue length l varies from 10 to 45. Fig. 2a shows that in the absence of attack, the proposed queue length estimation mechanisms, including *Mean*, *Huber K=1*, *Huber K=2*, and *Median* achieve similar MAPEs as the *Baseline* that relies on the stopping location of the last CV. The MAPEs of the five methods are also lower than 0.34 even when the CVs' market penetration rate ρ is only 0.1. If ρ reaches 0.5, which means half of vehicles on the road are CVs, the MAPEs of the proposed mechanism will be around 0.1. When the actual length is 20, an MAPE of 0.1 represents an error of two vehicles, which is quite acceptable in practice. Fig. 2b shows that the proposed mechanism and the *Baseline* method achieve similar MAPE under various actual queue lengths with the *Baseline* method slightly outperforming the other four methods when $l \geq 10$.

2) Impact of CVs' Market Penetration Rate ρ: Fig. 3 compares the MAPEs of the *Baseline*, *Mean*, *Huber K=2*, *Huber K=1*, and *Median* with CVs' market penetration rate ρ varying from 0.1 to 0.9, where the actual queue length $l = 20$. As we can see, the MAPE of the *Baseline* method is not affected by the change in the CVs' market penetration rate as the estimated queue length is determined by the last CV's position in the queue and the attacking CV will always report the maximum value 50. In contrast, the MAPEs of the proposed mechanisms including *Mean*, *Huber K=2*, *Huber K=1*, and *Median* decrease as ρ increases, especially when ρ increases from 0.1 to 0.3. This is because the proposed mechanisms estimate the queue length by using the stopping positions of all the CVs in the queue. If the number of attacking CVs is fixed, the more normal CVs in the queue, the weaker the impact attacking CVs can have on the estimated queue length $\hat{l}$. Moreover, for any fixed l, the expected number of normal CVs in a queue increases as ρ increases. As we can see from Fig. 3c, when $\rho = 0.1$, the MAPEs of the *Mean*, *Huber K=2*, *Huber K=1*, and *Median* methods are similar to that of the *Baseline* method. This is because when $\rho = 0.1$ and $l = 20$, the expected number of normal CVs in the queue equals to 2 which is smaller than the number of attacking CVs. If the number of attacking CVs is close to, or even greater than the number of normal CVs in the queue, the estimated queue length $\hat{l}$ will be dominated by the attacking CVs, which is similar to the *Baseline* method. However, generally speaking, the proposed mechanisms including *Mean*, *Huber K=1*, *Huber K=2*, and *Median* outperform the *Baseline* method.

3) Impact of the Actual Queue Length: Fig. 4 compares the MAPEs of the *Baseline*, *Mean*, *Huber K=2*, *Huber K=1*, and *Median* with the actual queue length l varying from 10 to 45, where CVs' market penetration rate $\rho = 0.4$. The MAPEs of the five methods all decrease as l increases. The reason is that the longer the actual queue length, the smaller the relative estimation error, and vice versa. Moreover, we can see from Fig. 4a, Fig. 4b, and Fig. 4c, the more attacking CVs, the larger the MAPEs of *Huber K=1*, *Huber K=2*, and *Median*, and *Mean*, which is expected. Furthermore, the *Huber K=2*, *Huber K=1*, and *Median* methods outperform the *Mean* method, as they all use robust estimator to compute the final queue length. Generally speaking, as long as there are sufficient normal CVs in the queue, the impact of attacking CVs can be greatly mitigated by normal CVs using robust estimators. In addition, as l increases from 10 to 30, the MAPEs of *Mean*, *Huber K=1*, *Huber K=2*, and *Median* are significantly lower than that of the *Baseline* method as they estimate the queue length by aggregating individual estimated queue lengths instead of relying on the position of the last CV. These results confirm the advantage of the proposed method over the *Baseline* method.

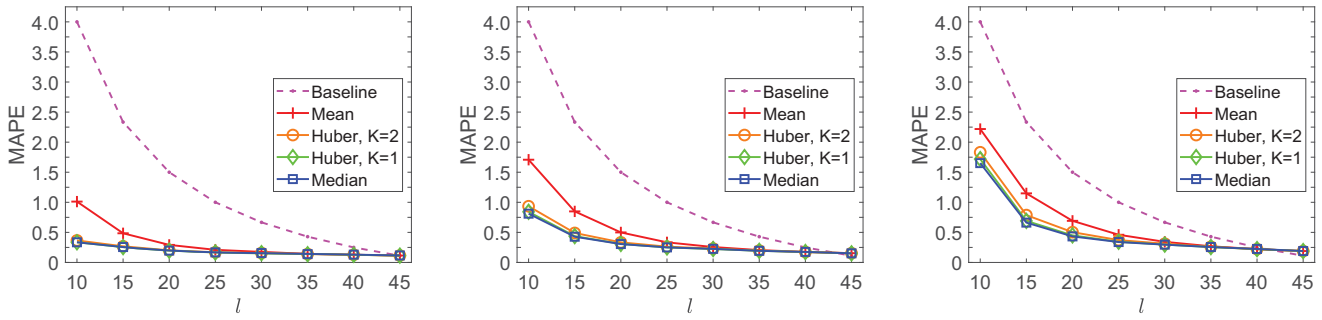
4) Impact of the Number of Attacking CVs: Fig. 5 compares the MAPEs of the *Baseline*, *Mean*, *Huber K=1*, *Huber K=2*, and *Median* methods with the number of attacking CV(s) varying from 0 to 3, where different CVs' market penetration rate $\rho = 0.3$ and 0.4 and the actual queue length $l = 15$ and 20. As Fig. 5 shows, the MAPE of the *Baseline* method depends on whether there is an attacking CV or not but is not affected by the number of attacking CV(s). The reason is that the *Baseline* method estimates the queue length based on the last attacking CV's position. In contrast, the MAPEs of the *Mean*, *Huber K=2*, *Huber K=1*, and *Median* methods all increase as the number of attacking CVs increases, which is expected. As we can see from Fig. 5b, Fig. 5c, and Fig. 5d, the MAPE of the *Baseline* method is slightly lower than that of *Mean*, *Huber K=2*, *Huber K=1*, and *Median* in the absence of the attack. However, the *Mean*, *Huber K=2*, *Huber K=1*, and *Median* significantly outperform *Baseline* when there is at least one attacking CV. Meanwhile, the performance of *Huber K=2*, *Huber K=1*, and *Median* are much better than that of the *Mean* method in the presence of at least one attacking CV.

5) Comparison of Mean, Huber $K = 2$, Huber $K = 1$, and Median: We now compare the *Mean*, *Huber K=1*, *Huber K=2*, and *Median* methods. As shown in Fig. 3 and Fig. 4, with different number of attacking CVs, the curves of *Mean* and *Median* represent the upper and lower bounds of the MAPE that the proposed mechanism can achieve through adjusting the factor K, as they represent the Huber's M-estimator with $K = \infty$ and 0, respectively. Moreover, we can see from Fig. 3 and



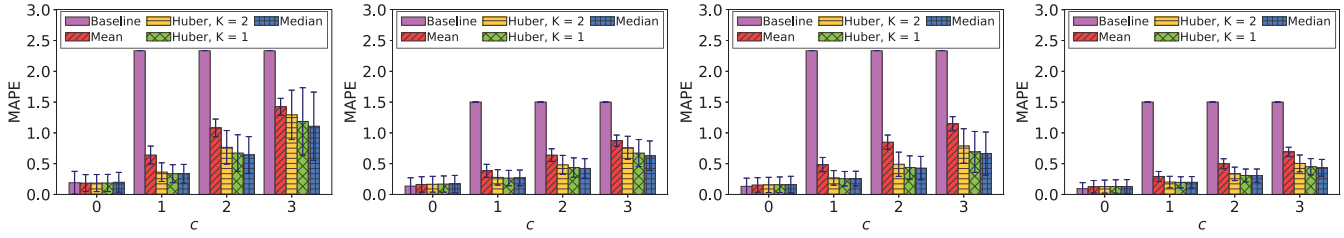
(a) 1 Attacking CV (b) 2 Attacking CVs (c) 3 Attacking CVs

Fig. 3. Comparison of the five queue length estimation methods with different market penetration rates of CVs ρ, where $l = 20$ and $c = 1, 2$, and 3.



(a) 1 Attacking CV (b) 2 Attacking CVs (c) 3 Attacking CVs

Fig. 4. Comparison of the five queue length estimation methods with different actual queue lengths, where the market penetration rate of CVs $\rho = 0.4$ and $c = 1, 2$, and 3.)



(a) ρ = 0.3, l = 15. (b) ρ = 0.3, l = 20. (c) ρ = 0.4, l = 15. (d) ρ = 0.4, l = 20.

Fig. 5. Comparing the the five queue length estimation methods with ρ and l, where $c = 1, 2$, and 3.

Fig. 4 that the proposed mechanism performs better with small K than with large K. Therefore, the smaller the factor K, the more robustness of the proposed queue length estimation mechanism, and vice versa.

On the other hand, the factor K also affects the estimation accuracy in the absence of the attack. As shown in Fig. 2a, when there is no attacking CV, the *Median* method incurs higher MAPE than the other three methods, especially when the CVs' market penetration rate ρ is between 0.1 and 0.3. When ρ is above 0.3, all four methods have similar MAPEs. Moreover, Fig. 5 shows that the *Mean* method has a smaller standard deviation in MAPE than the other three methods. Fig. 2a and Fig. 5 reflect the fact that the accuracy of the proposed mechanism increases as the factor K increases. In addition, Fig. 2a shows that, in the absence of the attack, *Huber K=1* achieves the same level of estimation accuracy as the *Mean* method. Meanwhile, Fig. 3 and Fig. 4 show that *Huber K=1* is as competitive as the *Median* method in the presence of the attacks. Generally speaking, *Huber K=1* is a good option for the proposed mechanism. Fig. 5 shows that the *Median* method is more robust in the presence of the attacks, although its estimation accuracy is slightly lower than the other methods in the absence of the attack. Therefore, when CVs' market penetration rate $\rho = 0.3$ and 0.4, the *Median* method is a preferable choice for the proposed mechanism.

VI. Conclusion

In this paper, we have presented a novel robust queue length estimation mechanism for CV-based traffic control systems. Unlike prior schemes that estimate the queue length based on the last CV's position, the proposed mechanism estimates the queue length from each individual CV's report based on their rankings among the CVs and all the vehicles and then aggregates them to produce a final estimate using Huber's M-Estimators. By doing so, we greatly mitigate the impact of attacking CVs reporting fake faraway positions from the stopping line. Detailed simulation studies confirm that the proposed mechanism outperforms prior solutions in the presence of data spoofing attacks at a slight sacrifice of the estimation accuracy in the absence of the attacks.

Acknowledgement

The authors would like to thank anonymous reviewers for their constructive comments that have helped us improve the quality of this work. This work was supported in part by the US National Science Foundation under grants CNS-1933047 CNS-1718078 CNS-1651954 (CAREER).

References

[1] R. L. Gordon, W. Tighe, I. Siemens *et al.*, "Traffic control systems handbook," United States. Federal Highway Administration. Office of Transportation ..., Tech. Rep., 2005.

[2] "ITS Research 2015-2019 Connected Vehicle," https://www.its.dot.gov/research_areas/WhitePaper_connected_vehicle.htm.

[3] "USDOT: Multi-Modal Intelligent Traffic Safety System (MMITSS)," https://www.its.dot.gov/research_archives/dma/bundle/mmitss_plan.htm.

[4] Q. A. Chen, Y. Yin, Y. Feng, Z. M. Mao, and H. X. Liu, "Exposing congestion attack on emerging connected vehicle based traffic signal control," in *Network and Distributed Systems Security (NDSS)*, 2018.

[5] Y. Feng, K. L. Head, S. Khoshmagham, and M. Zamanipour, "A real-time adaptive signal control in a connected vehicle environment," *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 460–473, 2015.

[6] C.-C. Yen, D. Ghosal, M. Zhang, C.-N. Chuah, and H. Chen, "Falsified data attack on backpressure-based traffic signal control algorithms," in *IEEE Vehicular Networking Conference (VNC)*, 2018, pp. 1–8.

[7] R. Dedinsky, M. Khayatian, M. Mehrabian, and A. Shrivastava, "A dependable detection mechanism for intersection management of connected autonomous vehicles," in *1st International Workshop on Autonomous Systems Design*, 2019.

[8] V.-T. Ta and A. Dvir, "A secure road traffic congestion detection and notification concept based on V2I communications," *Vehicular Communications*, p. 100283, 2020.

[9] W. Li, M. Nejad, and R. Zhang, "A blockchain-based architecture for traffic signal control systems," in *IEEE International Congress on Internet of Things (ICIOT)*, 2019, pp. 33–40.

[10] M. Amoozadeh, A. Raghuramu, C.-N. Chuah, D. Ghosal, H. M. Zhang, J. Rowe, and K. Levitt, "Security vulnerabilities of connected vehicle streams and their impact on cooperative driving," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 126–132, 2015.

[11] A. Abdo, S. M. B. Malek, Z. Qian, Q. Zhu, M. Barth, and N. Abu-Ghazaleh, "Application level attacks on connected vehicle protocols," in *22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, 2019, pp. 459–471.

[12] J. Sun, Y. Cao, Q. A. Chen, and Z. M. Mao, "Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures," in *USENIX Security*, 2020, pp. 877–894.

[13] J. Shen, J. Y. Won, Z. Chen, and Q. A. Chen, "Drift with devil: Security of multi-sensor fusion based localization in high-level autonomous driving under GPS spoofing," in *USENIX Security*, 2020, pp. 931–948.

[14] A. J. H. Clayton, "Road traffic calculations." *Journal of the Institution of Civil Engineers*, vol. 16, no. 7, pp. 247–264, 1941.

[15] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research." *Proceedings of the institution of civil engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[16] M. J. Beckmann, C. B. McGuire, and C. B. Winsten, *Studies in the Economics of Transportation*. Santa Monica, CA: RAND Corporation, 1955.

[17] L. A. Klein, M. K. Mills, D. R. Gibson *et al.*, "Traffic detector handbook: Volume I," Turner-Fairbank Highway Research Center, Tech. Rep., 2006.

[18] A. Skabardonis and N. Geroliminis, "Real-time monitoring and control on signalized arterials," *Journal of Intelligent Transportation Systems*, vol. 12, no. 2, pp. 64–74, 2008.

[19] H. X. Liu, X. Wu, W. Ma, and H. Hu, "Real-time queue length estimation for congested signalized intersections," *Transportation research part C: emerging technologies*, vol. 17, no. 4, pp. 412–427, 2009.

[20] X. J. Ban, P. Hao, and Z. Sun, "Real time queue length estimation for signalized intersections using travel times from mobile sensors," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1133–1156, 2011.

[21] Y. Cheng, X. Qin, J. Jin, and B. Ran, "An exploratory shockwave approach to estimating queue length using probe trajectories," *Journal of intelligent transportation systems*, vol. 16, no. 1, pp. 12–23, 2012.

[22] M. Cetin, "Estimating queue dynamics at signalized intersections from probe vehicle data: Methodology based on kinematic wave model," *Transportation research record*, vol. 2315, no. 1, pp. 164–172, 2012.

[23] M. Ramezani and N. Geroliminis, "Queue profile estimation in congested urban networks with probe data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 6, pp. 414–432, 2015.

[24] F. Li, K. Tang, J. Yao, and K. Li, "Real-time queue length estimation for signalized intersections using vehicle trajectory data," *Transportation Research Record*, vol. 2623, no. 1, pp. 49–59, 2017.

[25] K. Gao, F. Han, P. Dong, N. Xiong, and R. Du, "Connected vehicle as a mobile sensor for real time queue length at signalized intersections," *Sensors*, vol. 19, no. 9, p. 2059, 2019.

[26] G. Comert and M. Cetin, "Queue length estimation from probe vehicle location and the impacts of sample size," *European Journal of Operational Research*, vol. 197, no. 1, pp. 196–202, 2009.

[27] K. Tiaprasert, Y. Zhang, X. B. Wang, and X. Zeng, "Queue length estimation using connected vehicle technology for adaptive signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2129–2140, 2015.

[28] G. Comert and M. Cetin, "Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 563–573, 2011.

[29] Y. Zhao, J. Zheng, W. Wong, X. Wang, Y. Meng, and H. X. Liu, "Various methods for queue length and traffic volume estimation using probe vehicle trajectories," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 70–91, 2019.

[30] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[31] P. J. Huber *et al.*, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[32] H.-w. Jeng, "On small samples and the use of robust estimators in loss reserving," in *Casualty Actuarial Society E-Forum, Fall 2010*, 2010.

[33] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. John Wiley & Sons, Inc., 2009.