Infrastructure-enabled Defense Methods against Data Poisoning Attacks on Traffic State Estimation and Prediction

Feilong Wang[†], Xin Wang[†], Jeff Ban[†,*]

[†]Department of Civil and Environmental Engineering, University of Washington, WA, United States

[*]Corresponding author: banx@uw.edu

1. Introduction

Data from various sources has become increasingly available, revolutionizing many aspects of transportation, including traffic state estimation and prediction (*TSEP*), traffic control, safety, and more. However, as we navigate this exciting data-driven landscape, we also encounter new challenges. Our growing dependence on data has inadvertently opened the floodgates to potential cyberattacks on transportation systems. Data poisoning attacks, for instance, are initiated by adversaries who introduce malicious perturbations into a dataset, leading to erroneous results when the dataset is used for offline training/learning of data-driven models or online decision-making processes. Current studies primarily focus on data poisoning attacks and defenses against vehicular data (e.g., GPS and cameras) and vehicle-to-everything (V2X) data (e.g., V2X-enabled platooning). TSEP relies on data collected at the infrastructure side and generates traffic measures, which are significant both in their own merit and as a critical input to many applications. Recently, it was discovered that data poisoning attacks could compromise TSEP, affecting traffic prediction, queue length estimation, and vehicle classification (Wang et al., 2024). However, the quest to effectively defend against such attacks is still ongoing, presenting an unresolved security issue that beckons for a solution.

This paper explores how secure data from the infrastructure can empower defending against data attacks on TSEP. The investigation is fueled by the rapidly evolving transportation infrastructure that has becomes “smarter” and is equipped with more intelligent sensing, data collection, and computing devices. Harnessing the infrastructure for defense is not only cost-effective but also to minimize the reliance on ubiquitous user-side investment, enhancing transportation equity. Previous studies have illuminated the path, showing that secure data from infrastructure can facilitate the development of defense algorithms, such as those defending against GPS spoofing (Wang et al., 2023). In this paper, we propose an *Infrastructure-Enabled Defense (IED)* framework to combat data poisoning attacks on TSEP. Designing the IED framework for TSEP applications involve coordinating *three key aspects*. The *first* is a strategic decision on which secure infrastructure data to use. Depending on the sources of the secure data, the levels of security and collection cost could vary, thereby influencing the design and performance of the defense. The *second* aspect is a tactical one, involving how to utilize the collected secure data to detect attacks, considering specific data characteristics. For existing infrastructure that can provide relatively secure data, we need to repurpose this data to obtain a benchmark TSEP solution that can be used to detect attacks. This step could be challenging: deriving the benchmark TSEP solution is complex due to the distinct measurement model, and the secure data from infrastructure are typically spatio-temporally sparse and contain noise or errors. The *third* aspect is to determine how to mitigate the adversarial impacts. After detecting an attack, our method goes a step further to discuss correcting the errors caused by the attack to minimize the disruption. Another significant advantage lies in the ability to tag and isolate individual "poisoned" data samples (e.g., vehicles or mobile devices). Such a step is beneficial as a vehicle identified as poisoned at one location could continue to cause disruptions or should be scrutinized more at other locations. Therefore, the IED framework can not only defend against poisoning attacks at the locations where they are deployed but also enhance network system-wide security.

This abstract firstly presents the proposed IED framework considering these three aspects, and then tests various attack strategies. The performance of IED is evaluated and compared with existing defense solutions to demonstrate its benefits. We illustrate the IED method by the queue length estimation model using mobile sensing data. Defending poisoning attacks on the queue length estimation model using secure infrastructure data from wired loop detectors and signal controllers is demonstrated. In the full paper, we will show details of the IED framework, its generality of handling different types of secure data, the flexibility for other types of TSEP applications, the performance of defending state-of-the-art attacks, and advantages over existing defense strategies.

2. Methodology

2.1. Overview

The proposed IED framework is composed of three key components: secure data acquisition and benchmark TSEP solution, detection of data poisoning attacks, and error correction in TSEP solution along with tagging of attacked data. Each component is briefly outlined below. Figure 1 demonstrates an example of defense against attacks on queue length estimation. In this example, two virtual trip lines are established at an intersection to wirelessly gather a vehicle's travel time. This data can be processed to create delay patterns for queue length estimation (Ban et al., 2011). However, this mobile, wireless data is susceptible to poisoning, while secure data from loop detectors offer benchmark estimation. Machine learning (ML) and deep learning (DL) models can subsequently be developed to detect attacks, correct potentially compromised queue length estimates, and identify poisoned vehicles.

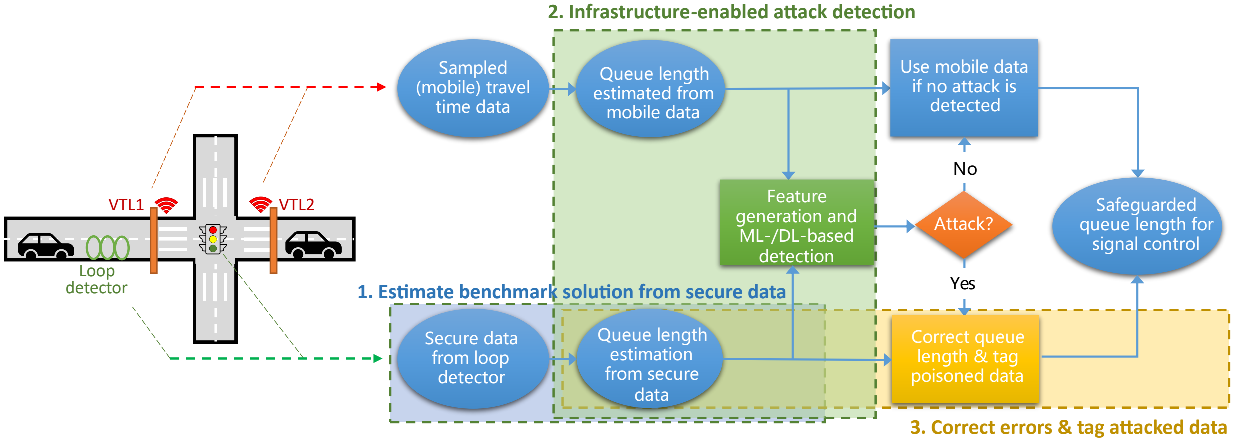
2. Infrastructure-enabled attack detection
Sampled (mobile) travel time data
Queue length estimated from mobile data
Use mobile data if no attack is detected
No
Feature generation and ML-/DL-based detection
Attack?
Safeguarded queue length for signal control
VTL1
VTL2
Loop detector
Yes
1. Estimate benchmark solution from secure data
Secure data from loop detector
Queue length estimation from secure data
Correct queue length & tag poisoned data
3. Correct errors & tag attacked data

Figure 1. Overview of the IED framework using queue length estimation as an example.

Before delving into each component, we present two types of threat models to facilitate a better understanding of the problem setup: 1) *False data injection attack* that manipulates the TSEP model's input data, causing TSEP to produce erroneous outputs: $x = x_c + \Delta x$. Here, x_c and x stand for the clean data and the data resulting from perturbation Δx. 2) *Sybil attack* that creates multiple fake identities to gain disproportionate influence or control over victim systems: $x = x_c \cup x_p$, with x_p standing for the created fake data samples. The semi-derivative-based attack (SDA) model developed in our previous work (Wang et al., 2024) is applied to *implement these attacks*. The SDA model optimally manipulates data to achieve an adversarial goal while meeting specific constraints.

2.2. Obtaining Secure Data and Benchmark TSEP Solution

Legacy sensors such as loop detectors are assumed to be cyberattack-free as they rely on wired communications for data collection and transmission. They can thus be used to obtain benchmark TSEP solutions. Typically, there are readily available TSEP models built on legacy, secure infrastructure data (x_{secure}). These are represented as a general TSEP solution mapping, taking into account the uncertainty associated with the solution Q_S due to data noises and biases in the TSEP model. For newly deployed infrastructure sensors designed to support defense, the communication channels for data acquisition will be secured using state-of-the-art security measures (Wang et al., 2023).

2.3. Infrastructure-Enabled Attack Detection

With the benchmark TSEP solution in place, we develop a detection algorithm that continuously monitors the data to detect potential attacks. This could be an ML-/DL-based model, comprising two steps as outlined below.

Feature Generation. This step aims to use secure data to alleviate the task of sanitizing data and to create robust features. A straightforward feature can be obtained by examining the inconsistency between the TSEP results (solutions) based on the secure data and the target (possibly poisoned) data. To create robust features, we address several challenges mentioned earlier, including leveraging the strength of spatio-temporal prediction models (e.g., graph convolutional network) to handle the *sparsity* of secure data, utilizing the dynamics of TSEP to address *noise* issues, and accounting for the uncertainty associated with the TSEP solution to handle potential model *biases*. Details will be provided in the full paper.

Detection Algorithm Development. By framing the attack detection as a real-time anomaly detection problem, we employ unsupervised ML-/DL-based models to learn anomalies based on the features constructed from the data, as outlined above. This generally involves several steps, including *model selection and training* depending on the characteristics of the TSEP application and the features, *metric design and hyperparameter tuning* (e.g., determining a threshold for the anomaly score above which TSEP data are considered anomalous), and *maintenance of the defense model* (e.g., updating data's security level, features, and the detection algorithm in the evolving attack-defense games).

2.4. *Correcting Errors in TSEP Solution and Tagging Poisoned Data*

Correction. With the benchmark TSEP solution, one could simply substitute the compromised estimate Q_M with Q_S. Yet, as Q_S often suffers from noise and errors, an alternative method involves combining Q_M with Q_S. This can be achieved using a statistical or ML-based data fusion model to balance their uncertainty and trustworthiness.

Tagging poisoned data. Tagging potentially poisoned vehicles is not a straightforward task due to the close distributions of the pristine and poisoned data. Drawing inspiration from the Random Sample Consensus (RANSAC) algorithm from robust statistics for robust data modeling, we develop an iterative algorithm to identify potentially poisoned data samples in a TSEP model.

3. Experiments

3.1. *Experiment Setting*

We test the proposed IED method to defend against attacks on the queue length estimation model using both simulation data and real-world traffic data (Ban et al., 2011). The simulation model was created using SUMO, simulating an intersection with fixed signal configurations. The real-world data was from the Next Generation Simulation (NGSIM) dataset. Mobile travel time data was gathered by setting virtual trip lines at each intersection and processing vehicle trajectories. Further details on the estimation model and implementing the attacks can be found in (Wang et al., 2024).

An isolation forest model is applied for attack detection. The performance of defense is evaluated using several perspectives: (i) *The IED should detect a data poisoning attack with high accuracy*. (ii) *The compromised queue length estimate should be corrected.* (iii) *The poisoned vehicles should be accurately tagged*. The accuracies are evaluated with F1 score, precision, and recall, while Mean Absolute Percent Error (*MAPE*) measures the errors in queue length estimations before and after correction.

3.2. *Results*

Table 1 summarizes the performance of the proposed IED solution when faced with false injection attacks. Similar observations can be made from sybil attacks. The attack detection algorithm based on secure infrastructure data can effectively identify signal cycles under attack. A recall of 0.96 in the simulation data tests indicates that 96% of attacked signal cycles in the simulation were successfully identified. However, the detection accuracy from the real-world data slightly decreases, as indicated by the lower F1 score. This decrease may be due to the larger data noise in the real-world data, which makes it more difficult to distinguish from perturbations added by attacks.

Table 1. IED performance of defending false injection attack on simulation and real-world data

		Simulation data	**Real-world data**
Accuracy of Detecting Queue Length under Attack	F1 score	0.96	0.93
	Precision	0.96	0.91
	Recall	0.96	0.95
MAPE of estimated queue length	Before correction	26	54
	After correction	4 (-85%)	8 (-85%)
Tagging Vehicle Accuracy (one intersection)	F1 score	0.94	0.87
	Precision	0.97	0.89
	Recall	0.92	0.74

The MAPEs of queue length estimates improve considerably after the detection and correction: for tests on both the simulation and real-world datasets, the MAPE can be improved by 85%. It is relatively challenging to identify the poisoned vehicles. Comparing with detecting the existence of an attack, the accuracy of identifying poisoned vehicles is lower, especially for the test on the real-world data that gives a F1 score of 0.87. The recalls of tagging the poisoned vehicles are even lower (0.92 in the simulation and 0.74 in the real-world data). The low accuracy could be due to the stealthiness of the attack model: the small perturbations added to the attacked vehicles make it difficult to identify the root-of-trust by differentiating the poisons from the measurement noises. Considering that only the delay pattern is used to detect attacks (for privacy protection), one potential way to improve the detection accuracy is to account for more information, such as scrutinizing vehicles at other locations. Our test shows that by scrutinizing data and checking the anomaly scores at two intersections, the recall increases by about 30% for tagging the poisoned vehicles.

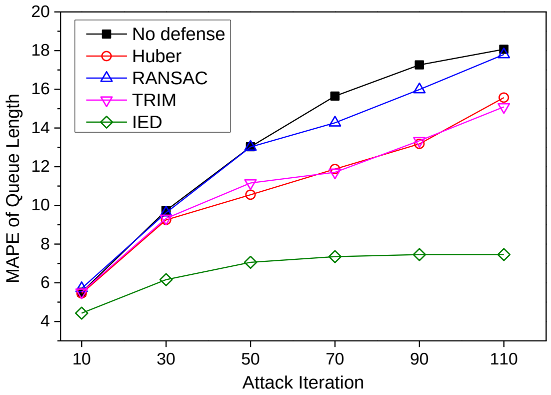

Figure 2. Comparison with existing defenses methods.

Figure 2 presents a comparison of the performance of the IED method with existing defenses, including Huber regression, RANSAC, and TRIM. The first two are well-established methods in robust statistics, while TRIM is a recently developed defense method against data poisoning attacks in the cybersecurity field (Jagielski et al., 2018). The baseline represents the MAPE of queue length estimation without any defense under varying numbers of attack iterations. A higher number of attack iterations implies more repetitive perturbations added to the data, resulting in a larger MAPE. The comparison suggests that existing defenses cannot safeguard the TSEP model: with more perturbations added to the data, the MAPE significantly increases. Compared with the baseline, none of these defenses can mitigate the adversarial impacts under relatively small poisons. *In contrast*, the IED method shows two advantages. First, it maintains the MAPE of queue length within 8% under all attack scenarios. Second, the MAPE is more controlled instead of increasing dramatically as more adversarial perturbations are added to the data. This confirms the advantage of using secure infrastructure data as an independent data source for defense.

4. Future Works in the Full Paper

Initial experiment shows that the proposed IED framework can effectively defend against data poisoning attacks on queue length estimation model. In the full paper, we plan to further conduct the following investigations:

1) Present more details of the IED framework including the detection algorithm and mitigation methods, as well as more numerical experiment results.
2) Apply the proposed IED framework to other TSEP models, such as crowdsourcing data-based traffic flow estimation and prediction on a corridor. We will need to consider suitable secure data from infrastructure and DL-based TSEP models.
3) Investigate the scenarios where the secure data are spatially or/and temporally sparse. We will analyze how the data sparsity may affect the defense and test the idea of using spatio-temporal prediction model to augment the secure data.

Reference

Ban, X. (Jeff), Hao, P., Sun, Z., 2011. Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C: Emerging Technologies 19, 1133–1156.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., Li, B., 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, in: 2018 IEEE Symposium on Security and Privacy (SP). Presented at the 2018 IEEE Symposium on Security and Privacy (SP), IEEE, San Francisco, CA, pp. 19–35.

Wang, F., Hong, Y., Ban, X., 2023. Infrastructure-Enabled GPS Spoofing Detection and Correction. IEEE Transactions on Intelligent Transportation Systems 1–15.

Wang, F., Wang, X., Hong, Y., Tyrrell Rockafellar, R., Ban, X. (Jeff), 2024. Data poisoning attacks on traffic state estimation and prediction. Transportation Research Part C: Emerging Technologies 104577.