A Visual Analytics Approach for Anomaly Detection from a Novel Traffic Light Data

Glenn Turner, Guoning Chen, Yunpeng Zhang

Abstract

Traffic signals are part of our critical infrastructure and protecting their integrity is a serious concern. Security flaws in traffic signal systems have been documented and effective detection of exploitation of these flaws remains a challenge. In this paper we present a visual analytics approach to look for anomalies in traffic signal data (i.e., abnormal traffic light patterns) that may indicate a compromise of the system. To our knowledge it is a first time a visual analytics approach is applied for the processing and exploration of traffic signal data. This system supports level-of-detail exploration with various visualization techniques. Data cleaning and a number of preprocessing techniques for the extraction of summary information (e.g., traffic signal cycles) of the data are also performed before the visualization and data exploration. Our system successfully reveals the errors in the input data that would be difficult to capture with simple plots alone. In addition, our system captures some abnormal signal patterns that may indicate intrusions into the system. In summary, this work offers a new and effective way to study attacks or intrusions to traffic signal control systems via the visual analysis of traffic light signal patterns.

1 Introduction

Cyber attacks on traffic signals is a great threat to national security and local economies. If multiple traffic signals are compromised there could be great economic impact, crippling at least a portion of a city. This is not merely a theoretical threat, it has happened. Two men were charged in hacking Los Angles traffic signals and disabling several signals as part of a labor dispute [12].

In addition, Connected Vehicles (CV) are being studied by the US Department of Transportation (USDOT). In September 2016, the USDOT initiated a pilot study to deploy CV based transportation systems [3]. These vehicles are wirelessly connected to the infrastructure to broadcast position and speed to aid in signal timing. Researchers at the University of Michigan have demonstrated with simulations that spoofing the data from a single vehicles can cause significant traffic congestion by interfering with the timing of the traffic signal. They were able to cause 22% of vehicles to take over 7 minutes for a typical half minute trip [3].

1.1 Problem Statement

We demonstrate here a method to find the kind of anomalies that would be associated with attacks to a traffic control system. To imagine how attacks may be seen in this traffic control system data we must first understand how these systems work. In the following, we provide a brief introduction of the traffic signal systems, followed by a summary of the potential attacks to the traffic signal systems.

The Terminology for a Traffic Signal A typical traffic intersection signal system is a combination of sensors, controllers, and network devices in addition to the signals themselves [8]. There are sensors usually embedded in the roadway to detect traffic flow. Occasionally video or other methods are used. The sensors may have a wireless connection or may be wired to the controller. These provide data to controllers that have programmed instructions and determine the state of the traffic lights. Some controllers are stand alone and work independently, and others are connected together to act as a system to coordinate signal information, and some are connected to a central server that sends them commands. Each controller has a Malfunction Management Unit (MMU) that only allows certain states for the lights. This prevents the signals from entering an unsafe state as a result of malfunction of the controller or of tampering. For example, Green lights in both directions are not allowed. These states are not software dependent but are hardwired into the MMU. The system is only allowed to be in one of those states, so if the controller fails the MMU activates an acceptable state, such as blinking lights [8].

Traffic systems that accommodate Connected Vehicles are more complex. There is a subsystem to process and store trajectory data, position, and speed, from the vehicles, and algorithms to do signal planning based on that data. However, such data are not available in the presented study.

Types of Attacks That May Be Found Traffic signal attacks that we consider here include: Denial of Service (or Distributed Denial of Service, DDoS) attacks, Data Signal attacks, where sensors that provide data to the controller are compromised, or the signals sent to the controller are falsified, and Light Control attacks, where the control system is directly compromised. We do not consider major attacks such as shutting off the power to the signal controller or physical destruction of part of the system.

In **Denial of Service** attacks the system might be overwhelmed with network traffic and may not be able to function at all. This may cause a shutdown of the controller. It would be possible for systems connected to the internet to be attacked in this fashion. We have no evidence that this is even possible for the system that provided our data. We do see large gaps in the data, so we cannot completely rule out a system failure.

With **Data Signal Attacks** the devices sending data to the control systems may be compromised, or falsified data may be sent. Often sensor data are sent via radio connection which are easily counterfeited. In addition, Connected Vehicles (CV) may be spoofed as previously mentioned. We expect to see short, or very long, Red or Green lights with any of these attacks.

A **Light Control** attack is a direct attack on the system. This involves changing the controller and is described in [8]. This may

occur by intercepting and changing control signals from a central server [5], infecting with malicious software or physically accessing the hardware. This may present in many forms, such as signal shutdown, resulting in an MMU defined state, stuck light (no change of signal for a very long time), or long or short Green and Red light states as we would see in Signal Attacks.

1.2 Motivation

To our knowledge there has no method to detect signal anomalies that would indicate attacks beyond simple observation of obvious malfunction. In the meantime, it is natural to look at the change of the traffic signal patterns to identify potential attacks. This may be true for some instances where the traffic signal data has regular patterns and deviations from these patterns are easily identified. However, in the data given to us we found no regularity in the traffic signal cycle times, making the recognition of abnormal behaviors even more challenging with any existing approaches. Furthermore, it is not known exactly how different attack types will affect the traffic light patterns. Finally, the data for this kind of analysis can be large, the data set used here has multiple records per second for only a few hours, resulting in over a million records. These are the challenges we attempt to address in this work.

1.3 Contributions

To aid the identification of potential attacks to a traffic control system, we develop a visual analytics framework for the traffic light data collected from a field test (details provided in Section 3). Our method first pre-processes the traffic light data to correct obvious errors, such as records out of order, and detects missing entries. It then follows the paradigm of "overview first, zoom, and then details-on-demand" to support user's exploration. In particular, to generate the overview of the data, a number of calculations are performed, including the decomposition (or extraction) of traffic light cycles (e.g., Green-Yellow-Red cycles), the accumulation of traffic light cycles within a given time period (e.g., to get total times on Green light), and the computation of the statistics of the traffic light cycles within a period (e.g., the distribution of the traffic light cycles and/or the average of the traffic light cycles). After these computations, the obtained summary information is visualized using standard plotting techniques, such as plots of cumulative time in signal states and/or histograms. From this overview representation, the user can perform detailed inspection using a number of user interactions. To aid the effective detection of potential abnormal patterns of traffic lights, we also normalize the individual traffic light cycles. This normalization effectively separates the abnormal patterns from the typical ones, which can be easily spotted in the scatter plots visualizing the normalized cycles. Furthermore, to aid the informative exploration of the normalized data, we also compute the meta information for the individual data points (e.g., corresponding to the individual traffic light cycles) shown in the visualization.

We have applied our visual analytics system to the processing of the traffic light data and successfully revealed a number of errors hidden in the input raw data that will be difficult to capture otherwise. To evaluate the effectiveness of our system in identifying abnormal traffic light patterns, we apply it to the data that includes artificially modified long or short Green/Red lights that may correspond to the Data Signal or Light Control attacks (based on the above description of different attack types). Our visualization can effectively separate these abnormal traffic light cycles from the other regular ones.

A reference implementation (source code, not a working system) of the techniques and system described in this paper can be found at *https://github.com/glenntu15/MMITSS_archive.git.*

2 Related Work

Visual analytics is an active research field in data visualization. It is impossible to review all visual analytics systems and the relevant techniques. There are many commercial tools for visual data analysis. Some examples include Sisense [14], Google Data Studio [4], Tableau [15], R Studio based GGpplot2 [10] and others that also incorporate business intelligence. These produce common plotting techniques that are used for business data. Transportation incidents have been investigated using basic 2D visual techniques such as histograms, scatter plots, and parallel coordinate plots integrated into a specialized system [13]. The system developed by these authors not only displays categorical data but is also linked to maps to give spatial references.

Recently, Häussler et. al. [9] have introduced a prototype interactive visualization system for urban traffic data based on high-resolution and high-dimensional environmental sensor data. This system for visual log analysis employed some novel displays including a "clock metaphor" display showing various parameters such as engine speed and CO_2 emissions in a polar coordinate system showing time of day. They also used a 2D "dense pixel" display to covey speed and CO_2 emission data for a road intersection.

Visualization techniques and different visual analytics systems have also been developed for other traffic related data, such as traffic trajectories. For example, a few systems have been developed for exploring taxi trips to understand the cause and effect of traffic jams [7] [18]. Visualization has also been applied to the travel time and reachability in a city transportation network with identification of locations with increased traffic volumes [2][19]. However, even though visualization techniques have been applied to different urban traffic data, there does not exist a visual analytics system for the traffic light data.

Custom visualization software packages for intrusion analysis, based on visual data mining of log data, have been developed using commercial packages mentioned above [6] [11]. These are targeted to network security concerns. Teoh et. al. have specifically developed novel visualization for this purpose [17]. Our approach is similar to their work, as well as the others cited here, in that we rely on human judgement with observation enhanced by data visualization. Our work is specifically tailored to show details in traffic signal data.

3 Background of the Data

The data for this study is from a field test associated with the Multi-Modal Intelligent Signal System, which is a system designed to improve mobility through the use of communications between vehicles and traffic signals to give the vehicles priority [1]. The data used are from one of two field tests. This one is in Arizona. There are several data sets produced by this test, but we specifically focus on the one described in a document section titled: "Detailed Description for Signal Plans for Roadside Equipment (RSE) Data", from the document "Multi-Modal Intelligent

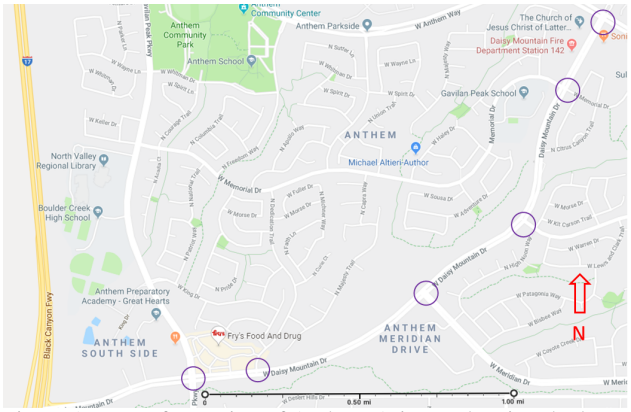


Figure 1: Map of a portion of Anthem Arizona showing the location of the test intersections based on a map provided by MMITSS documentation which was taken from Google Maps 2015. The first intersection (bottom left) is at **latitude 33.8, longitude -112.1** and we see a scale of 1.0 miles, per Google Maps 2019.

Traffic Signal Systems (MMITSS)-Sample Data, from Anthem Arizona". These data are traffic signal data at a number of intersections. There is additional data we did not have sufficient experimental details to understand, but did not seem related to the signal states. Each intersection is identified by an RSE (Road Side Equipment) number. The data was collected from March 3, 2015 to March 4, 2015.

The MMITSS document provides a map of the locations of the intersections where the data was collected. The test bed is described as six intersections extending from west to east. See Figure 1 for the map of the intersections (highlighted by circles). The six intersections are labeled as RSE25, RSE26,..., RSE30 from the bottom left to the top right, respectively. There is no signal data for the secondary road (or Minor road) at the intersections RSE27 and RSE28 (corresponding to the third and the fourth circles). Minimal examination is made for these two intersections.

Since this data are from a test using vehicles specifically equipped to give themselves priority in the traffic control system, we do indeed expect to see some Data Signal attacks that are part of this system testing. In the final report [1] the authors state that one of the performance measures of the system was to evaluate the total travel times of equipped vehicles. The document states:

"The MMITSS system determines the best priority timing based on the prevailing traffic conditions and the level of priority requested by the truck. The RSE either holds the green for the truck's direction of travel if the level of requested priority indicates the truck cannot make a safe stop, or decides if the phase should terminate based on prevailing traffic conditions."[sic]

We conclude, that the experimental vehicles can either hold the green light for the approaching vehicle or interrupt, i.e. short cycle, the traffic signals. We therefore *expect to see interruption of the normal signal cycle on occasion or very long Red and Green lights*. It is not clear if this priority timing only affects the major road. The final report [1], prepared by the Virginia Tech Transportation Institute, on traffic flow using this system references traffic on the Minor road for one intersection (RSE25). In addition, due to the lack of the information/data of the connected vehicles for the triggering of changes of the traffic light patterns, we are not able to verify whether our findings are indeed caused by the experimental vehicles, which is a limitation of this study.

4 Our Method

We follow closely the pipeline of classic visualization system as in [16]. Our exploration also follows the paradigm of overview first, zoom and filter, then details-on-demand. In the following we provide details for the individual components of the system specifically developed for this traffic signal data.

4.1 Filtering - Data Cleaning

The data were provided in a .csv (comma separated values) file containing 1.2 million rows and 23 columns. The last field of this data is an identifier that gives the name (RSE number) of the collection location. There are time stamps in the first column in epoch time (seconds since January 1, 1970). By examining the data we see multiple records with the same time stamp and the same RSE field. In addition, the records are not in strict chronological order. Our first operation was to sort the records based on their time stamps. However, there were too many records to sort in Excel, which loads only up to a million rows, so a Python program was used to read the original data and split it into six files (in .csv format) one for each intersection. Each of these six .csv files was then sorted in Excel.

After sorting all the files, the smallest start time (epoch) was noted, and set as a constant in the preprocessor program. This was subtracted from all times so the times for all locations were based on the same zero point.

In the data there are two columns for the traffic signal on the Major street, and two for the Minor street that are important. The first column is the signal state, 1 = Red, 3 = Green, and 4 = Yellow. The other column is the number of seconds (given to two decimal places) that the signal has been in the current state. When the data are sorted in Excel to get the correct chronological order, the records with the same time stamp are not necessarily in the correct order. We see a signal change state when the time stamp changes, and then change back a few records later with the same time stamp. This may be because the sorting in Excel is not stable. Another possibility is that records are just not in the right order. We know some time stamps are out of order because we inspected the original file.

To correct this, the preprocessor, when extracting the data, accumulates records with the same time stamp (at the same second) and sorts them to give a continuation of the signal seen in the previous second. A secondary sort on the time field (time in that state) is performed as well. When the data are sorted, it is easy to take the last time of a signal state when a signal changes to the next state. We can see the time spent with the signal showing Red, Green and Yellow. With some novel visual representation of the data, it was later discovered that this cleanup was not sufficient, which will be explained in the paragraph "Preprocessor checks for logic problems" below.

We initially defined a cycle by noting the signal state at the start, and after any data interruption. The records were processed until that state was reached again and the times spent in Red, Green, and Yellow states were saved. For some plots the individual times were output for Red, Green, and Yellow at the end of each of those states. These were used in the multi-line plots to

show the state times independently. For the data used in the scatter plots (plots of Red vs. Green times in the cycle) a data point was created from the cycle data that has the time for each component of the cycle. Calculations were made to normalize this data and output it to a separate file which will be explained later.

Preprocessor checks for logic problems. Checks were made to see if signals achieved all three states. A problem was seen in the data for RSE29. Figure 2 (a) shows state 3 (Green) of RSE29 existing in epoch time 1425409870 (we will call this 870), it then changes to state 4 (Yellow) in the same epoch second. In the next second, 871, we see state 3 again. We know this is a continuation of the previous state 3, because the time in that state (right hand column) is a continuation of the recorded times in the first two records shown. This gives a false cycle, and it is a cycle missing a red light state. In Figure 2 (b) we see the same data before sorting. Here the states and times are in the expected order. The cause of the negative times is unknown, but they are not important, because the program uses the last value in this column before the state changes, and these negative values are not used in the calculation. Because of this, we decided to use the **raw, unsorted** data for further exploration.

(a)

1425409870	3	34.36
1425409870	3	34.5
1425409870	4	-0.87
1425409870	4	-0.74
1425409870	4	-0.6
1425409870	4	-0.48
1425409871	3	34.63
1425409871	3	34.77
1425409871	4	0
1425409871	4	-0.35
1425409871	4	-0.2
1425409871	4	-0.07
1425409871	4	0.06

(b)

1425409870	3	34.36
1425409870	3	34.5
1425409871	3	34.63
1425409871	3	34.77
1425409871	4	0
1425409870	4	-0.87
1425409870	4	-0.74
1425409870	4	-0.6
1425409870	4	-0.48
1425409871	4	-0.35
1425409871	4	-0.2
1425409871	4	-0.07
1425409871	4	0.06

Figure 2: Portions of the sorted RSE 29 data (a) and the unsorted data (b). The highlights in (a) shows the cycle without a Red light state 5, while in (b) the states appear in the correct order despite some negative values.

4.2 Data Enhancements – Data Summarization

An Overview of the data, calculating hourly values. We first want to see if there are large abnormalities. For example, are there specific times of the data when the behavior is consistently different from other times? This could be any of the attacks mechanisms described in Section 1: Denial of Service, Signal Attacks or Light Control. In this examination of the data we accumulate the time in each signal state for a "wall clock" hour. That is, when an hour (3600 seconds) has elapsed the preprocessor outputs the total time spent in each signal state. The epoch time is converted to Hour, and Minute in local (Arizona - Mountain) time. It is reported for the time at which the hour ends. The total times do not equal 3600 seconds because Yellow light times were not included. Even considering variation attributable to this, we see at some times the reported total signal time is smaller than the others. We later learn that this is due to some missing data and the way we define an hour during processing. This will be seen in the subsequent plots.

Total cycle time. Next we look at the total cycle time. The purpose is to see if the total cycle time (Red+Green+Yellow) is a constant, or, has a certain pattern over time, and if there are deviations from the pattern. Any of the attack methods may be detected here if the normal function of the traffic signal is a regular pattern, but we are mostly looking for Signal Attacks.

Greater detail. We next look at the data in a finer resolution. The preprocessor checks for a complete cycle of the traffic light, and then outputs the time spent in Red, Green, or Yellow states. It saves the "state time" (time in that state) at every state transition. This is reported at the time the cycle starts as hours since the start of data recording. We hope to see a regular pattern that is occasionally interrupted. This would be a clear evidence of Signal Attacks or Light Control.

Examination of signal state times for clustering. We plot the time spent in Green state vs the time spent in Red state as a scatter plot of the cycles. It was thought that some clustering might reveal some behavior of the signals not yet seen. Specifically, if other attacks have been ruled out we might find Signal Attacks here.

Normalization to see additional features. In order to more effectively reveal abnormal patterns from these traffic signal data, we designed a new plot to visualize the normalized cycle times. In particular, instead of plotting based on the Red and Green light times for each cycle, we see the data based on the ratios of Red and Green lights in their respective full cycles (i.e. Red+Green+Yellow). Again, the more subtle Signal Attacks may be found with this method. This is a refinement of the previous method that may emphasize timing differences that would be seen in Signal Attacks.

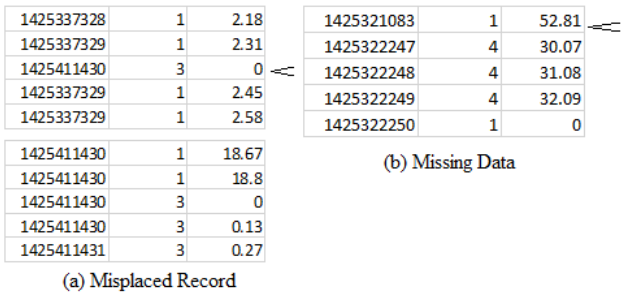
4.3 Visual Representations

After performing the above data cleaning and precalculation, the processed data and the extracted additional information are then visualized using standard plotting techniques, including bar charts, multi-line plots, and scatter plots. Bar charts are useful in showing the summary information of the traffic light cycle to provide an overview of the data (see Figure 5 for an example). Multi-line plots can show the transit between Red, Green, and Yellow light states. Comparing the trends of this transit between intersections and between Major and Minor roads of the same intersection may reveal useful information (Figure 9). We use scatter plots to show the individual traffic light cycles based on their Red light time and Green light time. We also use scatter plots to show the normalized traffic light cycles. As will be shown later, scatter plots are good at revealing the general distribution of the traffic light cycles and help reveal some outliers for further inspection (Figure 10).

Our user interaction supports **details-on-demand**. This is specifically useful when exploring the above scatter plots, in which the user can select a point and the coordinates of the selected point (corresponding to the Red light time and Green light time) are shown in a pop-up dialog. In addition, the meta data extracted above can be associated with points in the scatter plots to display the epoch time the point was created, as well as the Red, Green, and Yellow times as a character string. The raw data values are included in the metadata to identify the cycle that made up the calculation. See Figure 4 for an example. This has been

Table 1: Statistics for counting cycles as they are found or only counting a cycle starting in a certain state.

	Cycles Counted as Found						Cycles Start on Certain States					
	RSE25		RSE26		RSE29		RSE25		RSE26		RSE29	
Major/Minor	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min
Num Cycles	320	320	521	520	294	293	310	316	519	517	293	292
Avg Red Time	55.70	58.80	23.10	53.44	29.83	99.36	55.46	58.76	32.11	53.18	29.87	99.58
Avg Green Time	34.60	31.87	43.21	13.86	86.58	17.73	34.41	31.92	42.97	13.87	86.63	17.77
Avg Yellow Time	3.75	3.41	3.83	2.90	3.83	3.24	3.75	3.42	3.83	2.84	3.83	3.14
Max Red Time	100.3	138.27	49.15	200.41	73.43	385.54	100.30	138.27	49.15	200.41	73.43	385.54
Min Red Time	24.12	24.85	16.80	24.51	17.41	26.72	24.12	24.85	16.80	24.51	17.41	26.72
Max Green Time	71.69	66.78	190.15	39.92	375.17	45.07	67.02	66.78	190.15	39.92	375.17	45.07
Min Green Time	14.39	14.20	14.65	7.40	16.37	7.77	14.39	14.20	14.65	7.40	16.37	7.81
Std Dev Red	16.34	15.24	5.93	27.92	14.74	61.51	16.36	15.10	5.93	27.74	14.75	61.52
Num > StDev	56	50	67		67	51 37	54	50	67	66	50	37
Num < Stdev	62	55	23	36	0	1	60	53	22	36	0	1
Std Dev Green	11.22	11.50	27.94	5.88	62.01	12.36	10.76	11.54	27.70	5.89	62.11	12.37
Num > StDev	47	40	68	69	37	52	45	39	67	69	37	52
NUm < Stdev	55	61	36	23	2	0	53	60	36	23	3	0
Records	196141	196141	257030	257030	250344	250344						



1425337328	1	2.18
1425337329	1	2.31
1425411430	3	0
1425337329	1	2.45
1425337329	1	2.58

1425411430	1	18.67
1425411430	1	18.8
1425411430	3	0
1425411430	3	0.13
1425411431	3	0.27

(a) Misplaced Record

1425321083	1	52.81
1425322247	4	30.07
1425322248	4	31.08
1425322249	4	32.09
1425322250	1	0

(b) Missing Data

Figure 3: We see a data record is misplaced then corrected (left) and on the right we see that there is a gap in the data.

a valuable tool in exploring the scatter plots and understand the causes of the outliers.

Further data cleaning with the preprocessing information Upon examination we find in almost all cases there are records out of order, which makes the extraction of traffic light cylces unstable. Figure 3(a) provides an example of out-of-order records. This record, and other instances were fixed by editing the .csv files, so that the data is in the correct order.

After the above errors in the data are fixed, only one point appears to be apart from the others. This is shown in Figure 4. The point selected is at (0.51,0.13). We examine this point using the details-on-demand feature for showing data, also shown in Figure 4. The metadata is a string in the format of [epoch time:r:Red light seconds:g:Green light seconds:y:Yellow light seconds]. In this example, the metadata of this point is: **"1425322307:r:46.06:g:11.53:y:32.09"**. The light is in state 4, Yellow, for 32.09 seconds, then it transitions to state 1, Red, followed by state 3 Green. The time in red is 46.06 seconds, and in green 11.53 seconds. The calculated Red/total is 0.51, the Green/total is 0.13.

The Yellow state time of 32 seconds is much longer than other cycles and out of proportion to the Red and Green state times when compared with other cycles. It was determined that the cycle ends in the Green state. Other cycles examined have ended in the Red state. Tracing back up the data file until the cycle starts we see the situation shown in Figure 3(b). That is, there is a gap in the data whereby we hypothesize that the Yellow light may be thought of as part of another cycle for which we have no data for the Red and Green times; we see only three records for this Yellow state.

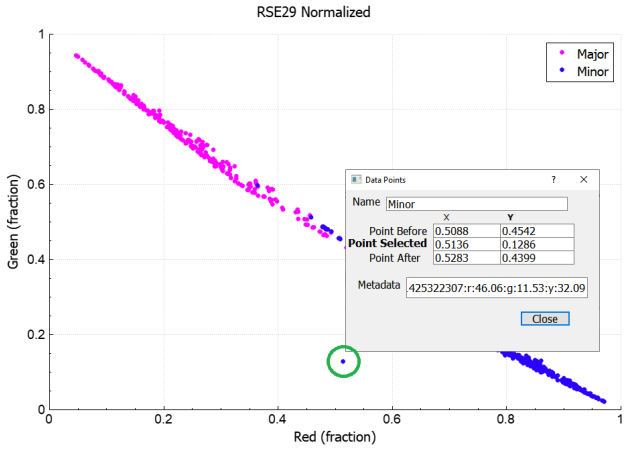


Figure 4: The metadata in the dialog indicates the epoch time for the outlier point, circled, as well as Red, Green and Yellow state times.

This explains, and mostly fixes, all points seen outside of the main cluster. This is an interactive data cleaning, which will be difficult to do otherwise.

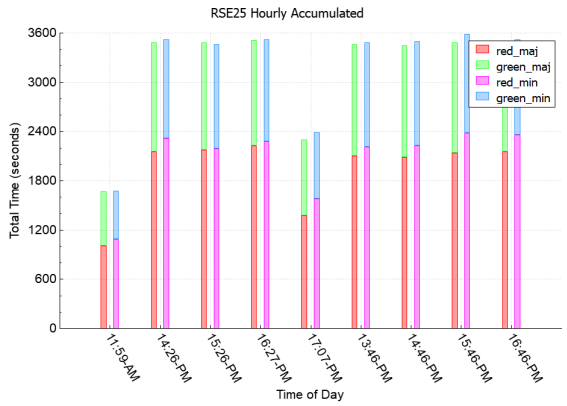
After seeing the apparent anomaly that seems to be dependent on how the cycles are chosen, we make the decision to only consider cycles starting in a particular state. We arbitrarily decide to use only cycles that start in Red for the Major road, and only cycles that start with Green for the Minor road. We had the preprocessor calculate various statistics before and after this change was made. The results are shown in Table 1. We see that few data were lost and the mean time for Red, Green, and Yellow states is nearly the same, as are the standard deviations. The last row of the table, Records, is the number of records read as raw data.

5 Results

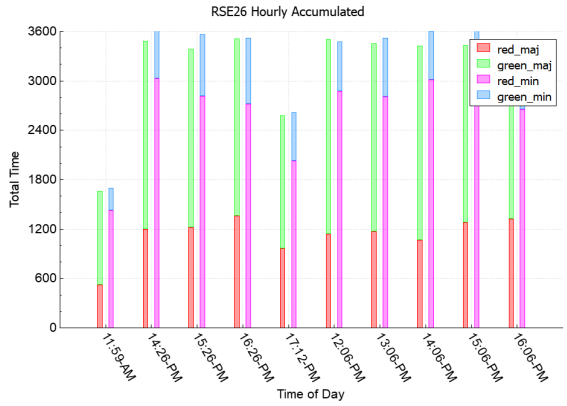
After the data are clean, we use the more promising visual representations to attempt to locate abnormal behaviors of the signal system. These are only some of the data visualizations evaluated, there are others that don't seem to contribute new information.

5.1 Overview

For RSE25 (Figure 5(a)), we see that the Red light time is longer than the Green light time for **both the Major and Minor streets**. This is unexpected, and cannot be explained without

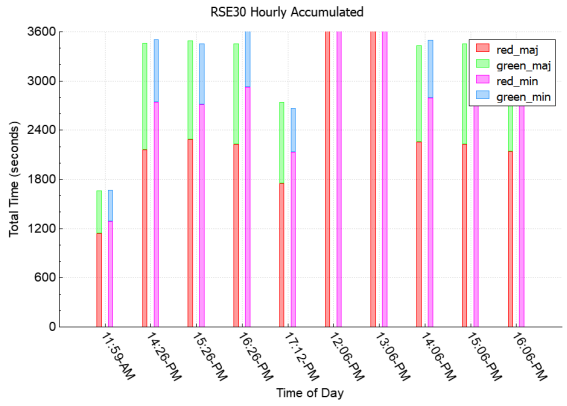


(a) RSE25 Major street (left bar) and Minor street (right bar)

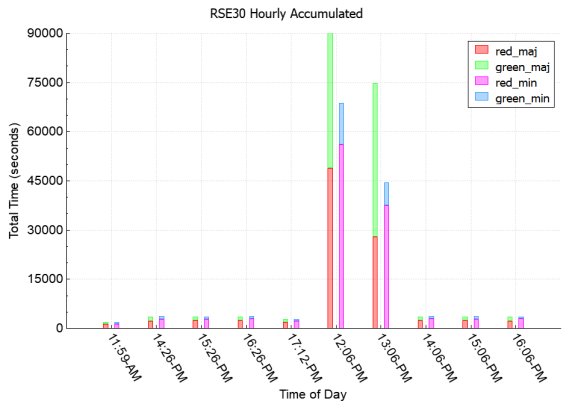


(b) RSE26 Major street (left bar) and Minor street (right bar)

Figure 5: Hourly accumulated signal time for RSE25(a) and RSE26(b). Note that Red light times are longer than Green for both Major and Minor streets in RSE25, but only for the Minor street in RSE29.



(a) RSE29 Major (left bar) and Minor (right bar)



(b) RSE30 Major street (left bar) and Minor street (right bar).

Figure 6: Hourly accumulated signal time for RSE29 and RSE30, Note two hourly totals for RSE30 are off scale.

other information. Also note that the Red light times and Green light times are nearly the same for this intersection. For RSE26 the plot shows much longer Red times in the Minor street than in the Major street, which is expected. Note that in these summary plots the signal times are reported at specific day and time. The intersections RSE27 and RSE28 do not have data for the Minor street, but the Major street shows longer Green than Red times, which is expected.

Figure 6 shows the data for the intersection RSE30. What is not apparent on the plot, is an obvious error in the data, i.e. two of the bars are going off scale. We expanded the scale by a factor of 15 to see the actual size of the bars in Figure 6(b). It was determined by examining the data file that for a segment of the time series, data in two columns were swapped, which causes the error.

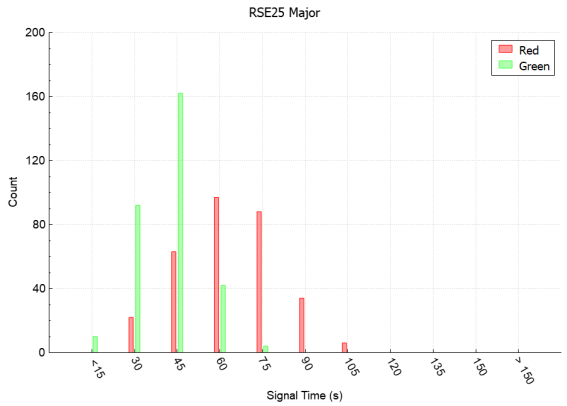
Next, we considered the distribution of the times for the Red and Green lights. Figure 7 shows the counts in various bins for RSE25 and Figure 8 shows the distribution for RSE29. When comparing RSE25 and RSE29 we see very different distributions of state times for the two intersections. In Figure 7 we see more short Green lights than Red lights for both the Major and Minor roads. For RSE29, shown in Figure 8 we see many short Red lights for the Major road and many short Green lights for the Minor road. The results for this intersection are expected and were confirmed by other plots not shown here but are similar to those for RSE25 in Figure 9. This trend can also be seen in Figure 11.

5.2 Detect Subtle Patterns in Signal States

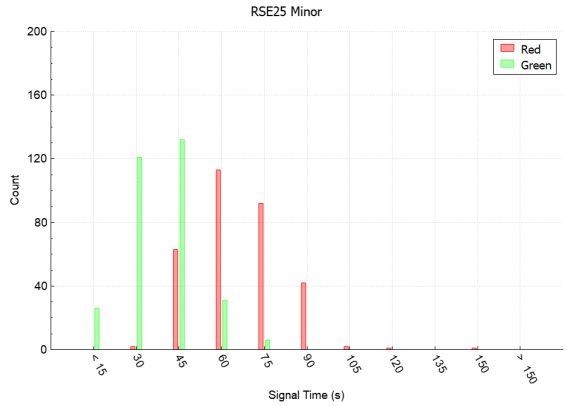
Figure 9 plots the times for Red, Green and Yellow lights for the Major road and the Minor road for the first intersection, RSE25. One thing we immediately notice is that there is a large amount of data missing, which is the case for the data of all intersections. This was seen in the previous plot as well, but the figure displayed is zoomed in to show some detail. There is no explanation in the document associated with the data on why this occurred. This could be a sign of a DDoS (Distributed Denial of Service, described earlier) attack, or Control attack that would shut down the system. We note that the signal for the major street spends more time on Red than on Green, which is confirmed by Figure 5.

5.3 Reveal Unusual Distributions of Times Spent in States

Next, we examine the Red and Green times in a scatter plot to look for any clustering that would indicate frequent occurrence of the same time spent in state. We see in Figure 10 that the behavior of the signals at two intersections is very different. If we look at Figures 5(a) and 5(b) in which the bar charts show the total

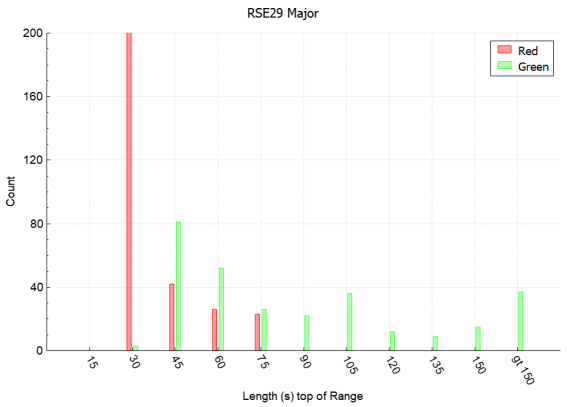


(a) Distribution of Major road

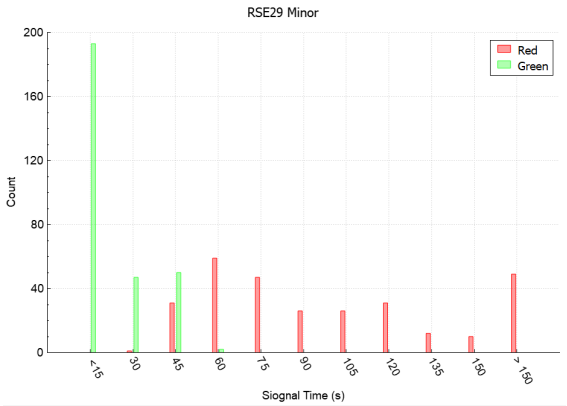


(b) Distribution for Minor Road

Figure 7: RSE25 distribution of signal lengths for the Major and Minor roads.



(a) Distribution of Major road



(b) Distribution for Minor Road

Figure 8: RSE29 distribution of signal lengths for the Major and Minor roads.

time per hour spent in Red and Green, we see confirmation of this difference. If we compare Figure 5(a) to Figure 5(b) we see that the time spent in Red and Green states for the Major and Minor roads is nearly equal. For other intersections, e.g., RSE25, and RSE26, shown in Figure 10, we see a marked difference between the times spent in Red and Green states for the two roads. In particular, we do not see clear clustering in the times for Major and Minor roads for RSE25. However, we do see a minimum Red time and a minimum Green time, but they are not distinctly different for the Major road and the Minor road. For RSE26 we clearly see that the Major road has a larger minimum Green time and smaller minimum Red time than the Minor road. We also see that there is a general clustering of times different for the Major and Minor roads.

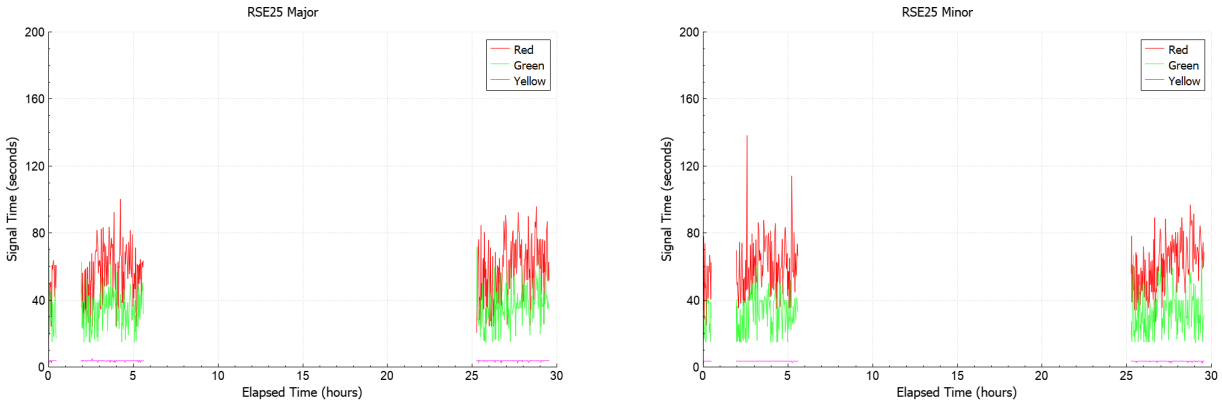
5.4 Detect Unusual Behaviors with Normalized Cycle Times

We revisit the normalized data using the now cleaned files. In Figures 11(a) for RSE25 and in 11(b) for RSE26 there are no points outside of the main point cloud. In Figure 12 (lower inset), there are two points from the Major road signal that are among the Minor road group, and one point from the Minor road signal among the Major road points. These are in the circled areas that are shown with a "zoomed in" scale in the insets of Figure 12.

We can take a closer look at the marked areas. There is one point for the Major road outside of the cluster of points. We first examine the point circled in the upper inset. The associated metadata for that point indicates that it is generated for the cycle ending at 25.7289 hours. We plot a trace of the light cycles, assigning a value of .5 when the light is Red, a value of 1.0 when Green, and 1.5 when yellow. In Figure 13 the Major and Minor roads are plotted together. We see in this figure that the Green light for the Minor road is longer than normal. The mean is 17.7 seconds and this instance is 43.85 seconds (from the metadata). It is interesting to note that what follows is both lights showing Red for a period of time after the next cycle, and the following cycle of the Minor road.

We next examine the point (0.73,0.22) circled in the lower inset. After investigating the raw data for these cycles it is seen that they just have a very short Green light for the Major road. The metadata shows that the cycle ends at epoch time 1425413688 (at the end of the yellow light, Major road cycles start on Red). We can see in Figure 14 that the Red light (state 1) time is 53.4 seconds, and the Green light (state 3) time is very short at 16.37 seconds. While we might think a short Green light on the Minor road is due to the test vehicles the short green light on the Major road could possibly be the result of signal tampering.

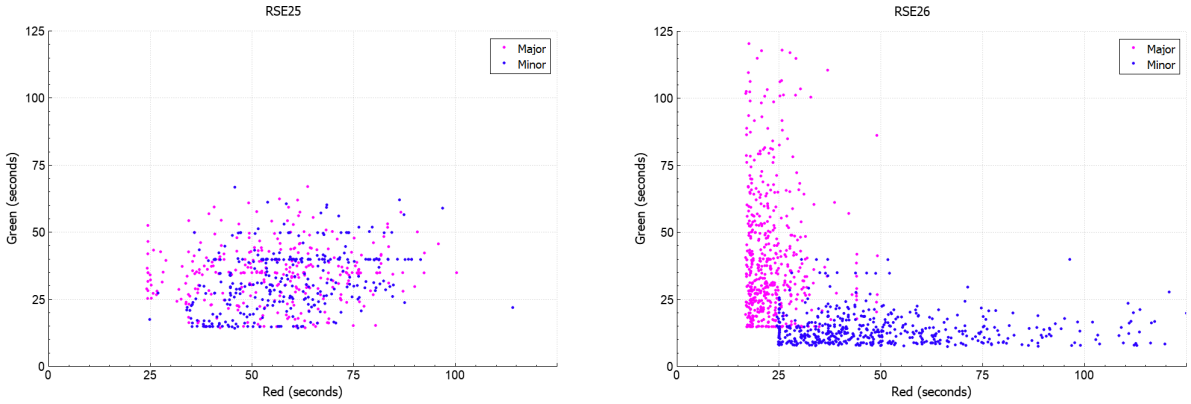
We look at an extreme point, not just aberrant ones. For example, we examine the metadata for the point at the bottom right of Figure 11 and see that this cycle ends at 24.4564 hours.



(a) Major Street

(b) Minor Street

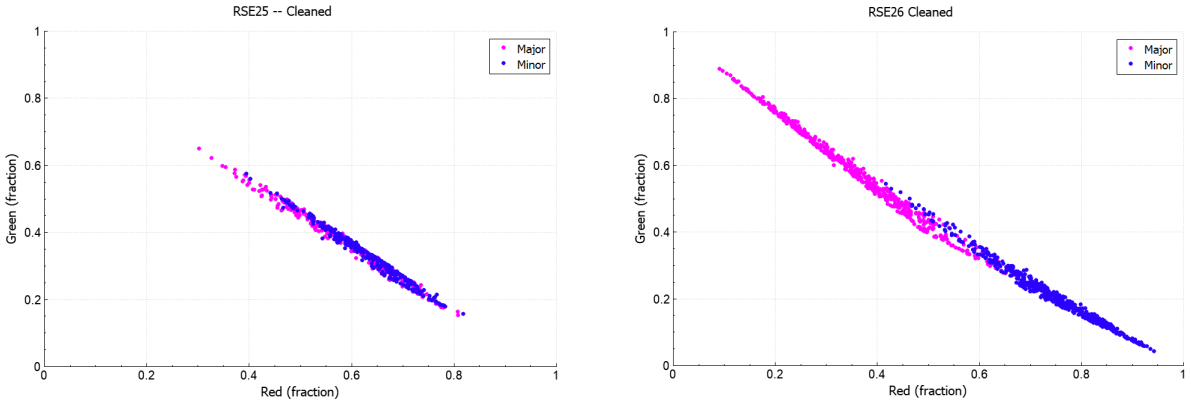
Figure 9: Comparison of Major and Minor Street Signals for RSE25. The red line is time spent in the Red state for each cycle, the green line is time for the Green light, and the purple line time for the Yellow light.



(a) RSE25 data

(b) RSE26 data

Figure 10: Scatter plots of Green vs Red actual times in seconds for RSE25 and RSE26. Note the difference in clustering for RSE25 and RSE26.



(a) RSE25 data

(b) RSE26 data

Figure 11: Normalized scatter plots of RSE25 and RSE26. We now see no points outside of the main point cloud. These are labeled "Cleaned" to distinguish them from similar plots shown previously.

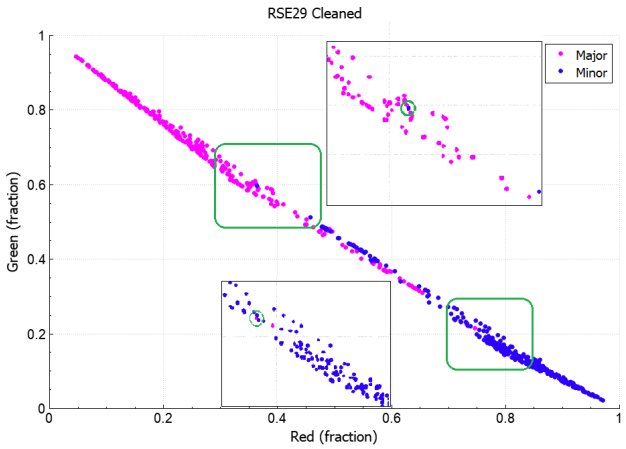


Figure 12: Normalized scatter plots of RSE29 with closeup views (insets) showing a point from the Major road (left) and two points from the Minor road.

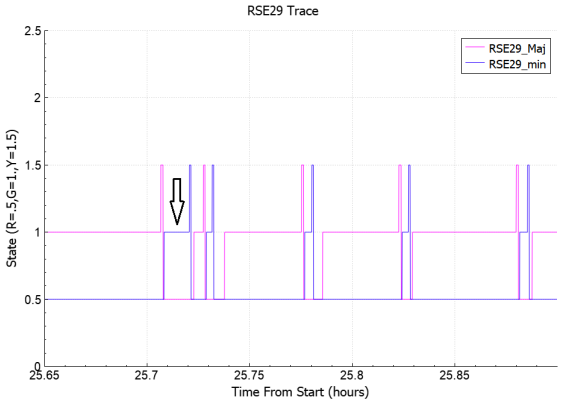


Figure 13: The states of the traffic signals in time on Major and Minor roads. The arrow shows the long Green light that gives the point shown in Figure 12 (upper inset).

We see in Figure 15, which shows the states of the signals as time progresses, and it is clear we have a very long Red light on the Minor road corresponding to a very long Green light for the Major road. This is almost certainly due to vehicles taking priority on this Major road. This is, in effect, a Signal Data attack.

The apparent slope of the line formed by the changed Red light data may be explained as follows. The total time (T) is Red + Green + Yellow and the points plotted are Red/T, Green/T. Neglecting the Yellow time Red/(Red+Green) + Green/(Red+Green) = 1. This is a reminder of the equation of a line $ax + by = c$. Now we restrict the Red to values that are nearly equal to yellow so we have approximately $2Rx + Gy = 1$, or, a line with a slope of -2.

We explore the ability to detect a particular Signal Data attack by simulating the data pattern expected. In reading the data we change the state from Red to Green after 4 seconds, limiting the Red time and lengthening the Green time. In this experiment we tampered with every third cycle. We call this **mutated data**.

In Figure 16 we see a detailed examination of points from the mutated data and from the original data. Using the metadata we can see the resulting calculations for the mutated point T = 4.00+70.99+3.9 = 78.89. Red fraction, $R_f = 4./78.89 = .05$, and Green, $G_f = 70.99/78.89 = 0.899$. For the point from the unmodified data we see a very long green light time, and

1425413660	1	45.26	#	3	43.4
1425413660	1	45.7	#	3	43.85
1425413660	1	46.16	#	4	0
1425413667	1	53.1	#	1	3.97
1425413668	1	53.24	#	1	4.1
1425413668	1	53.4	#	1	4.26
1425413668	3	0	#	1	4.4
1425413684	3	16.37	#	1	20.77
1425413685	4	0	#	1	21.63
1425413688	4	3.04	#	1	24.67
1425413688	4	3.18	#	1	24.8
1425413688	1	0	#	1	24.97

Figure 14: Picture of the raw data file with rows hidden to show the data that gives point (0.73,0.22) in the Major road data. Left column is epoch time, the next column is state, 1 = Red, 3 = Green, 4 = Yellow, and then the time in that state. The two right most columns are state and time for the Minor road.

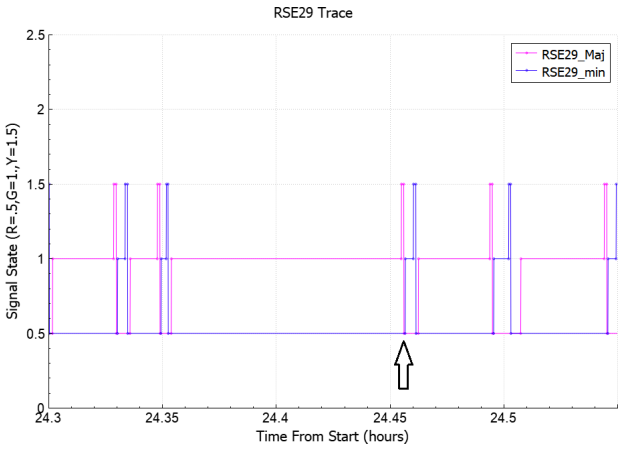


Figure 15: The states of the traffic signals in time on Major and Minor roads. The arrow shows the end of the cycle that gives the right most point in Figure 12 (lower inset).

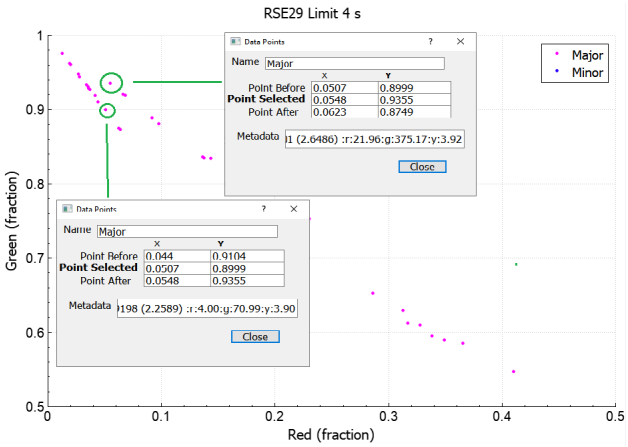
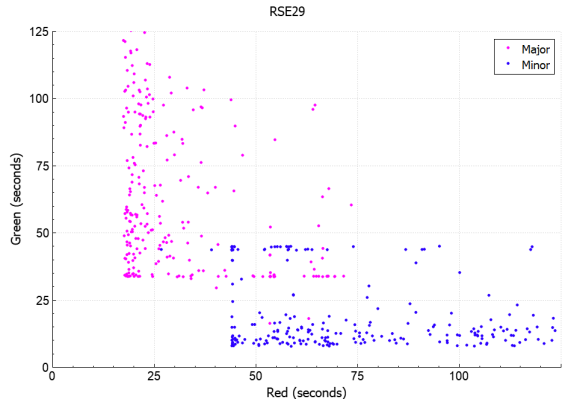
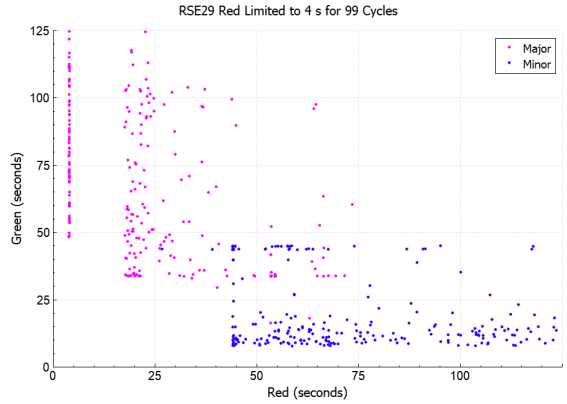


Figure 16: Detailed examination of a point resulting from mutated data (smaller circle), and a point from the unmodified data (larger circle).

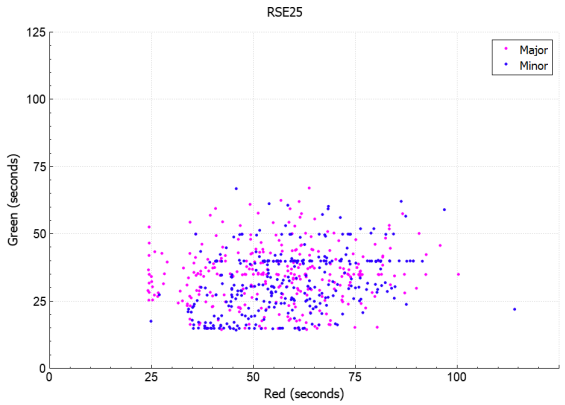


(a) Scatter plot without modifying the data

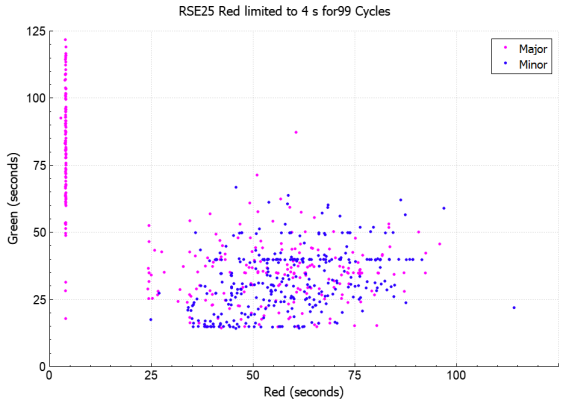


(b) Red limited to 4 seconds for 90 cycles of Major road.

Figure 17: This is the result of mutating the data to shorten the Red light time for the Major road to appear to give priority.



(a) Scatter plot without modifying the data



(b) Red limited to 4 seconds for 99 cycles of Major road

Figure 18: This is the result of mutating the data to shorten the Red light time for the Major road to appear to give priority. The Green light is given the additional time taken from the Red light time.

$T = 21.96 + 375.17 + 3.92 = 401.05$, and $R_f = 21.96/401.05 = 0.055$, and $G_f = 385.17/401.05 = 0.94$

We return to the scatter plots showing the raw time in seconds for the Red and Green states. In Figure 17 we see the result of forcing the shorter Red light times.

In a similar manner we alter the data in RSE25. This intersection displays somewhat different characteristics from the others. The modified cycles are still apparent. We see that the obvious very short Red lights in both cases are easily distinguishable from the cluster of points that make up the bulk of the cycles seen.

6 Conclusions and Future Work

Intrusion detection is crucial for maintaining a reliable and safe traffic control system for our daily life. Different from previous methods, in this work we demonstrate that traffic light data may be applied to detect intrusion to the traffic control system. To achieve that, we devised a first visual analytics system to help process and analyze a large-scale traffic light data that is made available to the public for the first time. Our system integrates a number of data cleaning operations and summary information computations specifically developed for the traffic light data. It also provides a few simple but effective visualization and user interactions to support a level-of-details data exploration of the traffic light data. We have applied the developed system to successfully identify a number of errors hidden in the given data that are difficult to find otherwise. Our system also effectively revealed the abnormal behaviors in the given traffic light data (e.g., very short or long Red and Green light times) that indicate the potential attacks to the corresponding traffic signal system.

6.1 Summary of Results

Various preprocessing calculations and plotting techniques have revealed subtle errors in the data. We identified two root causes of errors: missing data, and records out of order. Both of these could reasonably be attributed to Light Control (described in Section 1) hacking. These more likely have other explanations. The data recording may have been turned off overnight. The records out of order may have been caused by the data logging software not maintaining the correct sequence when merging data from various intersections. We just don't know the source of the errors. When the data errors were removed, we were able to find such anomalies such as very short and very long Green lights. These may have been caused by manipulation of the system beyond our research.

6.2 Limitations and Future work

There are a number of limitations of our current approach that we plan to improve in the future. First, while we believe a

short Green light on the Minor road is caused by the test vehicles, the short Green light on the Major road could possibly be the result of signal tampering. We don't know the system logic; it may just rush the whole cycle to sync with an approaching vehicle. It's probably reasonable to assume that the points at the extreme edges of the normalized plots are not due to tampering, but to normal operations. Further investigation is needed to clarify this.

Second, the intersections in this study were simple, i.e., one Major road and one Minor road, and there is no "turn lane" signal. A more complex intersection may also be considered. The additional part of the complete cycle when a turn lane is introduced may be observed for strange behaviors. For example, to facilitate setting the light Green in time for a particular vehicle's arrival the Green light for the turn lane may be skipped altogether.

Third, the current system is developed specifically for a traffic light data, which we would like to apply to other data of this type. It is conceivable that for the purpose of security monitoring a record could be generated at every state transition for the signal. This would reduce the volume of data. With fewer records longer times could be studied. Once patterns emerge machine learning or other automated methods could be developed to monitor signals in real time. A "sliding window" of time may be used to monitor the signal and an alert issued before there is major disruption in the system.

References

[1] Ahn, K.; Rakha, H.;Hale, D. (2015), Multi-Modal Intelligent Traffic Signal Systems (MMITSS) Impacts Assessment [Online]. Available: https://rosap.ntl.bts.gov/view/dot/3557

[2] Andrienko, G. ;Andrienko, N.;Hurter, C.;Rinzivillo, S.;Wrobel, S. (2011) From movement tracks through events to places: Extracting and characterizing significant places from mobility data," 2011 *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Providence, RI 161-170 2011.

[3] Exposing Congestion Attack on Emerging Connected Vehicle based Traffic Signal Control
Chen, Q.;Yin, Y.;Feng, Y.;Mao, Z.;Liu, H. (2018)
Network and Distributed Systems Security (NDSS) Symposium, San Diego, CA, [Online] Available: https://pdfs.semanticscholar.org/2f41/0fd80f5468307647432d95ec5f268ac1db11.pdf

[4] Milnik, j.(2019) How to build a Google Data Studio Dashboard [online] Available: https://www.socialmediaexaminer.com/how-to-build-google-data-studio-dashboard

[5] Ernst, J.; Michaels, A. (2017) Framework for Evaluating the Severity of Cybervulnerability of a Traffic Cabinet *Transportation Research Record* 2619(1), 55–63 [Online] Available: https://doi.org/10.3141/2619-06

[6] Etoty, R; Erbacher, r (2014) A Survey of Visualization Tools Assessed for Anomaly-Based Intrusion Detection Analysis *Army Research Laboratory* [Online] Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a601590.pdf

[7] N. Ferreira, J. Poco, H. T. Vo, J. Freire and C. T. Silva, (2013) Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips *IEEE Transactions on Visualization and Computer Graphics* (19)12, 2149-2158, 2013.

[8] Ghena, B.; Beye, W.; Hillaker, A.; Pevarnek, J.; Halderman, J. Green lights forever: Analyzing the Security of Traffic Infrastructure *WOOT'14 Proceedings of the 8th USENIX conference on Offensive Technologies*

[9] Häussler, J.; Stein, M.; Seebacher, D.; Janetzko, H.; Schreck, T.; Keim, D. (2018) Visual Analysis of Urban Traffic Data based on High-Resolution and High-Dimensional Environmental Sensor Data *Visualization in Environmental Sciences (EnvirVis 2018)* [Online] Available: https://pdfs.semanticscholar.org/5002/f43e0e85625e3c7afe03b0a3f428b1ded5d9.pdf

[10] Kaur, J. (2016) Data Visualization with GGpplot2 [online] Available: https://datascience899.wordpress.com/category/ggpplot2/

[11] McLendon, M.;Shhead, t.; Wilson, A.;Wylie, B.; Baumes, j. (2010), Network algorithms for information analysis using the Titan Toolkit *44th Annual 2010 IEEE International Carnahan Conference on Security Technology* (1-10) San Diego, CA

[12] McMillian, R. (2007), Two Charged with hacking LA traffic lights [online], Available: https://www.computerworld.com/article/2549204/two-charged-with-hacking-la-traffic-lights.html 103-111 (2010)

[13] Pack, M;Wongsuphasawat, K (2009) ICE – Visual Analytics for Transportation Incident Datasets *2009 IEEE International Conference on Information Reuse & Integration* 200-205 (2009) Las Vegas, NV

[14] Sisense (2019) Power to the Analytics Builders [online] Available: https://www.sisense.com

[15] Tableau (2019) Changing the Way You Think About Data [online] Available: https://www.tableau.com

[16] Tela, A. (2008), *Data Visualization Principles and Practice* A. K. Peters Ltd

[17] Teoh, S.;Ma, K; Wu, S; Jankun-Kelly, T (2004) Detecting flaws and intruders with visual data analysis *IEEE Computer Graphics and Applications*

[18] Wang, Z.;Lu, M.;Yuan, X.;Zhang, j.;Wetering, H. (2013) Visual Traffic Jam Analysis Based on Trajectory Data *IEEE Transactions on Visualization and Computer Graphics* (19)12, 2159-2168, 2013.

[19] Zeng, W.; Fu, C.;Arisona, S.;Erath, A.;Qu, H. Visualizing Mobility of Public Transportation System *IEEE Transactions on Visualization and Computer Graphics* (20) 12, 1833-1842, 2014.

Author Biography

Please submit a brief biographical sketch of no more than 75 words. Include relevant professional and educational information as shown in the example below.

Glenn Turner is a PhD student at the University of Houston Department of Computer Science. After a career writing engineering software, earned a masters degree in Computer Science from the University of Houston (December 219) with emphasis on simple system for data visualization.

Guoning Chen is an Associate Professor at the Department of Computer Science at the University of Houston. He earned his Ph.D. in Computer Science from Oregon State University in 2009. Before joining the University of Houston, he was a post-doctoral research fellow at the Scientific Computing and Imaging (SCI) Institute at the University of Utah. His research interests are in Data Visualization, Geometric Modeling, Geometry Processing, and Physically-based Simulations.

Yunpeng Zhang holds a Ph.D. degree in computer science. He is working as an Assistant Professor at University of Houston. Dr. Zhang has worked for Boise State University and Dakota State University (U.S.), University of Melbourne (Australia), and Imperial College London (U.K.) as a Cybersecurity and Software Engineering expert for more than 15

years. Dr. Zhang's research interests include cybersecurity and software engineering. He has invented more than 40 high-performance/security new algorithms/methods and developed 8 software systems. He has published 80 papers in peer-reviewed journals and conferences.