Project Proposal

**Analysis of Basic Safety Message in Vehicle to Vehicle Communication using Principal Component Analysis and Pearson Co-relation Matrix**

CIS: 602-01 Data Visualization

Spring 2018

Dated: 03/05/2018

By:

Prinkle Sharma

01474931

Ph.D., Electrical and Computer Engineering

University of Massachusetts, Dartmouth

Table of Contents

[Problem Statement 3](#_Toc507960201)

[Rationale 3](#_Toc507960202)

[Research Question(s) 3](#_Toc507960203)

[Methodology/Approach 4](#_Toc507960204)

[Deadline/Plan/Schedule/Timeline 5](#_Toc507960205)

[References 6](#_Toc507960206)

# 

# Problem Statement

With the evolution of the Internet of Things (IoT), a new quantifiable and measurable world are created which tends to make people’s life more secure, comfortable and productive. Sensor devices like smart homes, automobiles, smart watches, etc. which produces a massive amount of data every second can help us to derive knowledge in making timely-decisions that can improve our lives. But with the rise of connected real-time data sources, cybersecurity has become a consumer protection issue. Due to the rapid rise in such data sources, the detection of anomalies injected intentionally or due to malfunction in critical infrastructures, such as vehicular ad-hoc networks (VANET), challenges conventional methodologies. Due to the ephemeral nature of VANET, the existing intrusion or malware detection schemes for ad-hoc networks render ineffectively to investigate misbehaviors. To ensure security and privacy inside VANET, vehicles broadcast information via anonymous pseudonym IDs in Basic Safety Message (BSM) which makes existing misbehavior detection schemes inapplicable in VANET. So far, misbehavior detection in VANET using Machine Learning has not been thoroughly investigated. In this project, we are aiming to visualize the Principle Component Analysis (PCA) reduced BSM for every vehicle journey and then compare it with other vehicles using Pearson correlation for a misbehavior detection by using d3 visualizations (e.g. heat map, matrix reordering).

# Rationale

Vehicular networks are emerging majorly to enhance the safety and traffic on road. VANET is a cyber-physical system which tends to make the decision based on the information communicated via V2V and V2I (vehicle to vehicle and vehicle to infrastructure). The vehicles tend to communicate via broadcasting certain types of messages up to the range of 300 meters. The issue arises when the legitimate vehicles start manipulating the mobility information for personnel benefits which might lead to a catastrophic failure. Injecting false information associated with their movement (GPS location, speed, heading, vehicle size, etc.) to create a misapprehension for other vehicles thereby obliging them to take wrong decisions. Many techniques based on the cryptography, subjective logic, game theory, probability theory and reliability models have been studied but unfortunately these prevention techniques, due to highly dynamic nature of the VANET, are insufficient to thwart the inside attackers to spread incorrect information. This work focuses on studying the behavior of the vehicles traveling together on the same path and finding the correlation between them using machine learning techniques to detect anomalies. Currently, US Department of Transportation is running active V2V enabled vehicle projects in three different locations: Wyoming, Tampa, and New York. For this research, the dataset (Wyoming) available at <https://catalog.data.gov/dataset/wyoming-cv-pilot-basic-safety-message-one-day-sample>, where the complete dataset for 2 days will be considered. The dataset is updated every day and made available for public on the website.

# Research Question(s)

The anticipated results from this project will answer the below-stated questions (not limited to):

1) How to keep track of vehicles that falls under 300-meters BSM broadcast range?

2) Is it a good idea to reduce the dimensionality of BSM? Visualize the principal components of the dataset.

3) Is there a (co)-relation between the vehicles traveling on the same road at the same time? Visualize it.

4) How does the speed, elevation, heading, and GPS correlate to each other and can we track back the vehicle identity through its attributes excluding the pseudo ID (in case it starts broadcasting wrong info)?

5) Limitations, if any?

# Methodology/Approach

For vindicating the research, the first and foremost challenge is to capture, analyze, filter and visualize the huge and complex dataset that is inadequate to process at the first stage. The aim to build a powerful visualization that could allow the access to huge amount of data in straightforwardly digestible visuals. The project will be divided into certain tasks that will be accomplished to justify the reasoning about the quantitative information provided. This section will outline our expectations of each of the three phases of the project. The deliverables and expected outcomes will also be pointed out.

**Task-1: Examine the dataset.**

This phase will involve the extensive background research of the attributes of the dataset and then label them in “Attribute Types.” Then the attributes will be visually encoded using marks and channels.

**Task-2: Selecting the attributes.**

This phase will focus to reduce the dimensionality of the dataset by selecting the key/significant attributes that contribute the most to detecting the misbehavior or that are highly co-related to each other. Here I would also like enlighten why these attributes are important to study and how their correlation with respect to other attributes is dependent.

**Task-3: Correlation and PCA**

The correlation matrix of the selected attributes for an individual vehicle will be generated using Pearson co-relation and will be visualized. Then the other vehicle’s co-relation with the same time and date stamp will be compared. The reason behind performing this experiment is to figure out does two vehicles share the same attributes co-relation at a certain location and then to make the judgment by looking at co-relation if the vehicle, which was in the vicinity couple of seconds, sends wrong information about its location. Principle Component Analysis will also be visualized.

So far, I have believed d3 based visualization like a stacked bar chart (for PCA) and graph like heatmap, bubble chart and motion chart could be approached to answer the questions defined above. While comparing co-relation charts of two different vehicles, the third map generated will show the difference between the attribute values when the cursor is placed on it or when the attribute is selected from the dropdown list placed at the legend position (as shown in picture). The visualization is not restricted to one shown in the Figure 1.

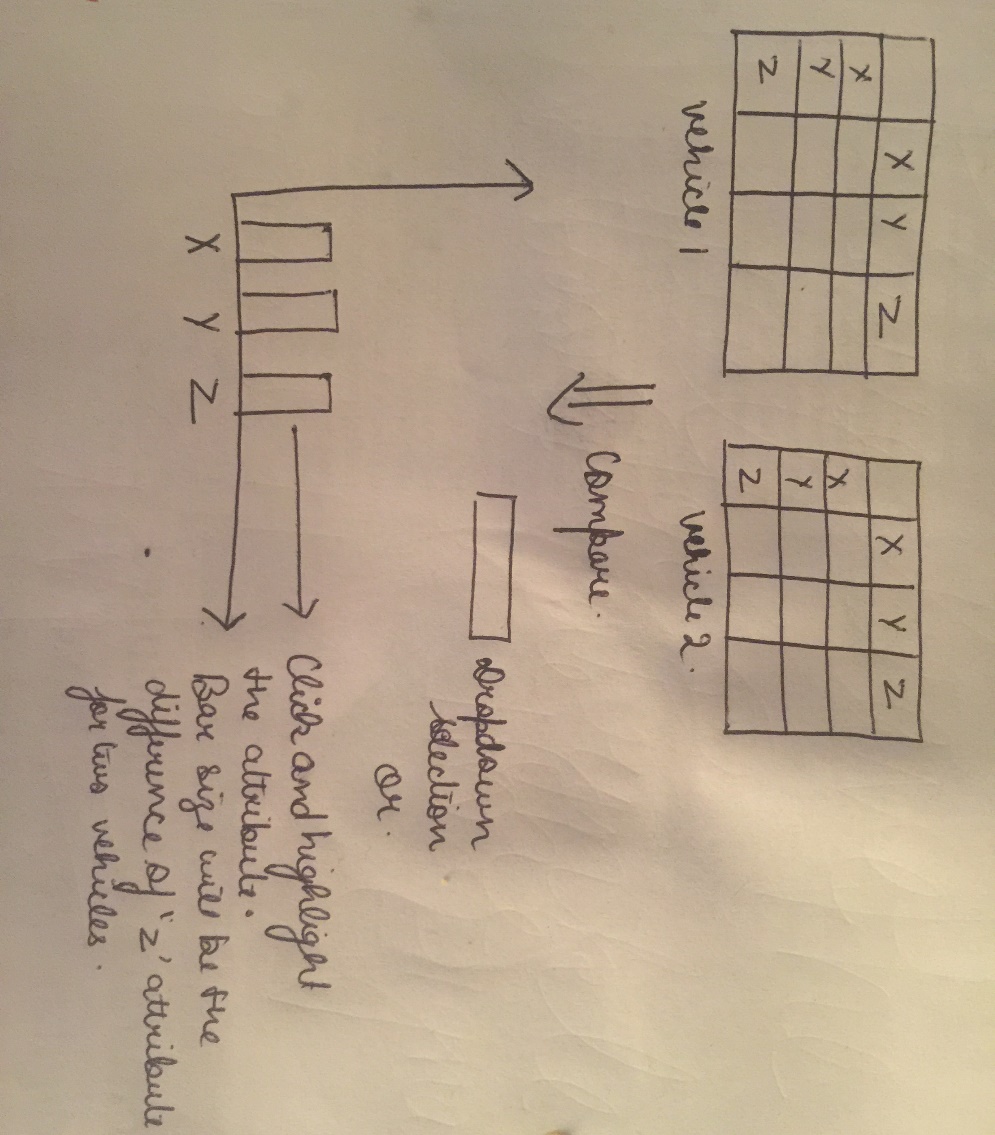


Figure : Preliminary Visualization

# Deadline/Plan/Schedule/Timeline

March 2nd week (11th- 17th): Collect the dataset for two dates and filter it accordingly.

March 3rd week (18th-24th): Generate the correlation matrix table. (using XLSTAT tool)

March 4th week (25th-31st): Code for the building the graph depicting the correlation between different attributes.

April 1st – 2nd week (1st-14th): Code for building the graph which compares the correlation values between two vehicles and visualizes the comparison on the different graph.

April 3rd-4th week (15th- 30th): Code for graph calculating the PCA values of the correlation graphs generated.

May 1st week: Presentation and report preparation.

# References

* Munzner, Tamara. *Visualization analysis and design*. CRC press, 2014.
* Najafzadeh, Sara, et al. "BSM: Broadcasting of safety messages in vehicular ad hoc networks." *Arabian Journal for Science and Engineering* 39.2 (2014): 777-782.
* <https://github.com/d3/d3/wiki/Gallery>
* <https://www.xlstat.com/en/>
* <https://wydotcvp.wyoroad.info/>