Data Analysis and Information Security of an Internet of Things (IoT) Intelligent Transit System

Zachary Yorio*, Raymond Oram*, Samy El-Tawab*, Ahmad Salman*, and M. Hossain Heydari*
*James Madison University, yoriozp@jmu.edu, oramrr@dukes.jmu.edu, {eltawass, salmanaa, heydarmh}@jmu.edu

** B. Brian Park
** University of Virginia, bpark@virginia.edu

***Abstract* - Public transportation around midsize educational cities has become increasingly vital as residential and commuter populations continue to grow every year. Our research team proposes a cyber-physical system that monitors the quality of service of the transit bus system around James Madison University (JMU), located in Harrisonburg, Virginia, USA. By utilizing the power of Internet of Things (IoT) devices, such as Raspberry Pi computing devices, it is possible to create a network of smart nodes that collect data on the bus routing efficiency and ridership. Using the collected data and using big data analysis, improvements can be made to bus route efficiency and traffic congestion in Harrisonburg, as well as similar college towns. This paper presents data analysis for the first deployment of the IoT nodes at seven JMU bus shelters, during the Spring 2017. Data stations compile and submit the Media Access Control (MAC) addresses and timestamps of wireless devices surrounding the bus stations. This information is stored in a Cloud Storage database that allows for big data analysis and convenient access. In this paper we present the usage of the data collected, modifications to improve both the transit system as well as the data collection methods for future node deployments. Our results show the average passenger waiting times at a sample of the seven bus stations. Route optimization and dynamic changes for the route are discussed. Also, we discuss several concerns on security and privacy related to the collection, transmission, and storage of data in the Cloud (e.g., privacy of the ridership MAC addresses, tracking of a specific bus rider in the system ...etc.). Additional security implementation has been suggested to emphasize security and privacy protection.**

Index Terms – Cyber-Physical System, Intelligent Transportation Systems, Transit System, Internet of Things (IoT), Cloud Computing, Security and Privacy

INTRODUCTION

With the rapid technological advancements of sensors, wireless sensors and communications have become the main technology for IoT [1]. Wireless sensors are usually small, communicate wirelessly to a Cloud Storage, and are often deployed without a network topology [2], [3]. The deployment environments of the IoT devices could be over a small or large geographical area in locations that can be public or hostile. Typically, the environments require little human interaction and devices go unattended for months or even years.

According to the United States Department of Transportation, Intelligent Transportation Systems (ITS) is one of the key aspects of improving transportation safety and efficiency. ITS enhances American productivity through the integration of advanced sensing and communication technologies into the transportation infrastructure and in smart vehicles. With the new era of wireless communication, ITS uses these technologies to advance its applications [4], [5].

ITS improves planning, design, efficiency, and management of transportation systems. Topics range from communications, computers, decision systems, sensors, simulation, signal processing, quality assurance and more [6]. Traditional ITS uses mostly legacy technologies such as inductive loop detectors, magnetometers, video detection systems (e.g., cameras), acoustic tracking systems and microwave radar sensors in conjunction with probe vehicles and other means to estimate traffic parameters [7].

According to the United States Department of Transportation, there were 5.3 million crashes and 2.22 million injuries in 2011 alone. These numbers accounted for over 32,000 fatalities. In 2010, the cost of congestion in urban areas was around $101 billion. Furthermore, the total amount of wasted fuel topped $1.9 billion gallons [8]. One of the main goals of intelligent transportation systems is public safety, by sharing information that can prevent potential crashes, keep traffic moving and decrease the negative environmental impact of the transportation sector on the society [9]. Using IoT, ITS systems can collect and analyze important data in public transportation, about bus riders, to put in place an infrastructure to minimize cost and waiting times while also maximizing safety and efficiency.

Despite all these improvements in the world of intelligent transportation, security of all these new systems are still under investigation. In this paper, we examine the security and vulnerabilities of the bus data collection system proposed as an example of an Intelligent Transportation System and suggest remedies to design a more robust system [10]. In our research, Smart units are used to collect

978-1-5386-6343-1/18/$31.00 ©2018 IEEE

data (e.g., Media Access Control (MAC) addresses) to uniquely identify students' smart phone waiting times for the bus at each station. No personal data is collected. The MAC address is used to detect how many bus riders are waiting at the station.

RELATED WORK

Recently, several researchers have been looking into the idea of using Wireless Communication as part of Intelligent Transportation System. For example, Tubaishat et al. proposed the idea of using Wireless Sensor Networks (WSNs) as a significant improvement over the traditional wired sensors for several applications (e.g., Smart Parking and Traffic Monitoring) [11]. Although most of the researchers focused on how to use the wireless technology in conjunction with the rapid increase of Cloud Computing (e.g., Vehicular Cloud [12]) and Internet of Things (IoT) used in the Intelligent Transportation, few researchers highlight on the threat of wireless attacks [13]. Moreover, most of the later researchers focused on the Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication security [9]. Recently, WiFi, Bluetooth and Wireless technology have gained attention among researchers for more intelligent transportation applications without depending on vehicle communication [14], [15]. In fact, WiFi readers installed inside buses were used to estimate origin-destination of bus passengers [16].

Real-time sensing gains its efficiency from the evolution of sensing capabilities through smartphones, smart vehicles, and Internet of Things (IoT). The usage of smartphones in the U.S. was doubled from 2010 to 2014 and expected to double again in 2018 [17]. Almost every commuter has a smartphone, loaded with sensors that can be easily utilized in real-time sensing. Smart vehicles with all their sensing and communication capabilities (e.g. Bluetooth) are operating in some cities already, in 2013 showing 48% sales increase in California [18], [19].

OVERVIEW OF THE CYBER-PHYSICAL SYSTEM

El-Tawab et al. introduced a cyber-physical system (CPS) to monitor the efficiency and the quality of service for transit buses around an academic institute [10], [20]. This cyber-physical system detects the number of riders on each bus. The CPS included several components that were used to scan and analyze the WiFi data (e.g., Raspberry Pi 3 Model B equipped with a micro SD card, 5V/2A portable battery, and a wireless adapter). In this work, we use smart nodes; which are self-contained hardware packages used to collect data. Smart nodes are located at fixed locations in the bus stations. Each smart node is set into monitor mode sniffing wireless network traffic in the surrounding area of the bus station (with radius of 7m) [21].

Without compromising the privacy of any of our passengers, we use TShark, a network protocol analyzer to capture these packets of data including the arrival time, MAC address of device, strength of the WiFi signal, etc. from various WiFi-enabled devices [22]. It is worth noting that no personal data is collected, except the MAC address to count the number of people waiting at the station. The packets of data have useful information that can then be parsed. Once the parsing is complete, it would be exported into a data file, then sent to a cloud-based database as shown in Figure I.

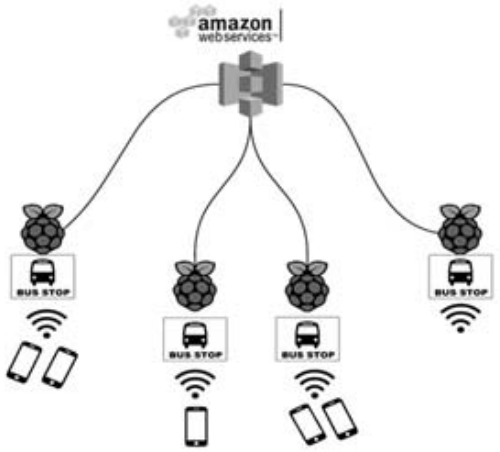

FIGURE I
THE NETWORK ARCHITECTURE OF SMART NODES INSTALLED AT EACH BUS STATION

A total of seven nodes were configured and deployed to collect data from around James Madison University (JMU) as shown in Figure II. Data sent to the Cloud Storage consists of time stamps, MAC addresses, and Received Signal Strength Indicator (RSSI(s)). Using the minimum and maximum waiting times for a specific device in a bus station, we can determine the waiting time for each rider. These MAC addresses are further filtered by re-matching them with the MAC addresses obtained at the next bus stop(s) to eliminate false positives and false negatives.

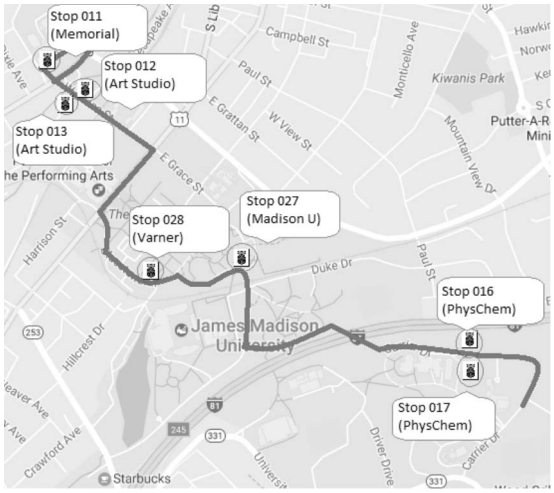

FIGURE II
THE NETWORK ARCHITECTURE OF SMART NODES INSTALLED AT EACH BUS STATION

False positives can occur with people passing by the bus stop, cars driving by, and/or WiFi signals coming from inside the buildings nearby the station. The system can eliminate these false positives using the MAC address identifier of the captured data (parsed later to be removed) and/or by removing inconsistent cases: such as someone sitting at the bus stations (e.g., waiting for someone) for longer than normal. False negatives can also occur, where students do not have active smart cell phones. With the large number of students riding the bus at each station, an approximated estimate can be accepted in these cases.

MEMORIAL BUS STOP (STATION #011)

We choose one of the seven bus stations to highlight our results and analysis of the data. We pick "Memorial Hall Bus Station" known as station #011 at James Madison University. This station is one of the furthest stations at JMU that usually students would prefer to take the bus to/from rather than walking.

I. Daily Bus Traffic Frequency

Using the Pandas graphing extension in Python, automated graphs were generated from the .CSV files pulled from the MySQL database via our tailored queries [23]. Figure III is an example graph created for the data collected on April 12, 2017. It shows the number of students waiting for a bus at the Memorial stop in ten-minute intervals. From Figure III, we notice that traffic pattern at the Memorial stop is at the peak around noon time, when the most number of classes are scheduled in Memorial Hall, on Wednesdays, in the Spring semester.

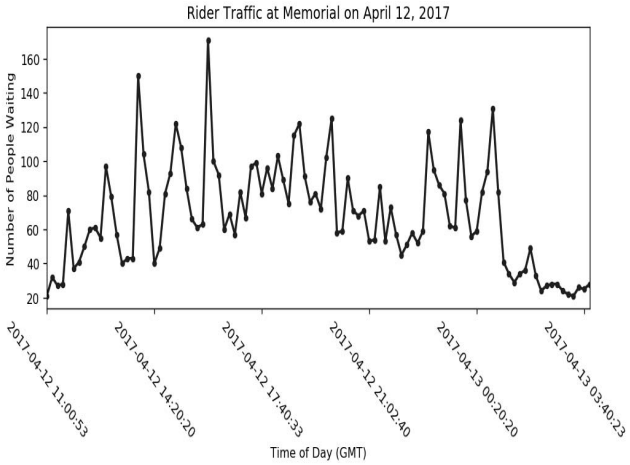

FIGURE III
BUS TRAFFIC FREQUENCY AT MEMORIAL ON APRIL 12, 2017

Graphs like these are used to help visualize traffic numbers. However, extra work needs to be done to determine the daily and weekly patterns to improve bus route efficiency.

II. Bus Traffic Frequency Analyzed by Day of Week

From the big data collected, and with graph generation automated with the help of Python and Pandas, a graph per day of the week was easily created for the rest of the month during which data was collected. Figure IV depicts the three Tuesdays for which data was collected.

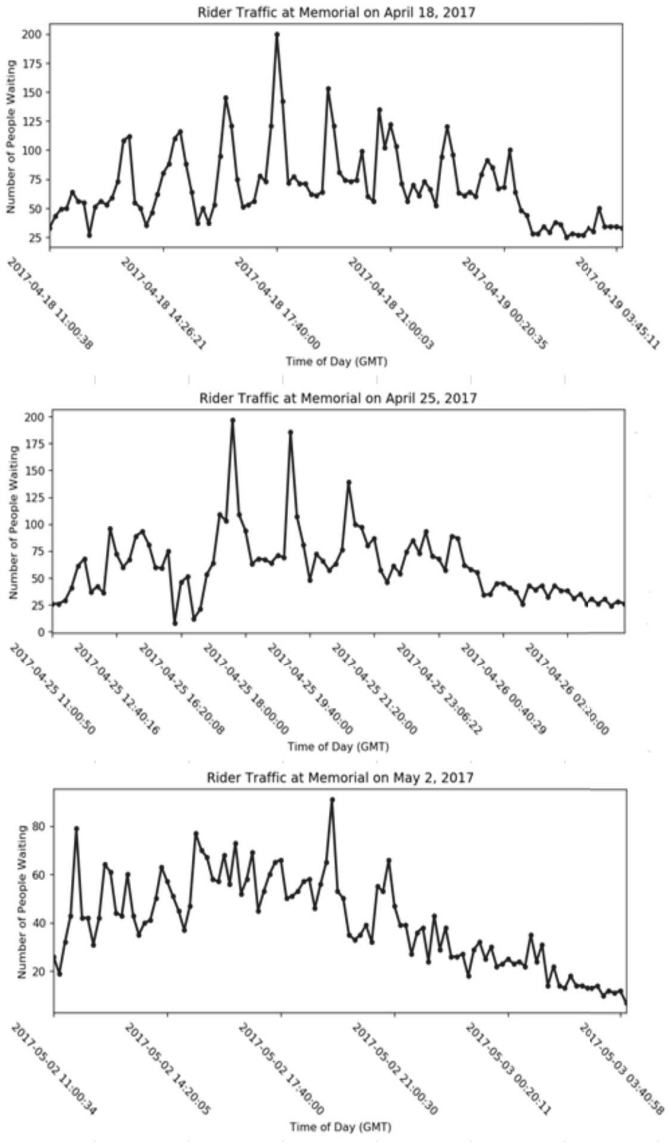

FIGURE IV
TUESDAY BUS RIDER TRAFFIC AT MEMORIAL ON THREE DIFFERENT WEEKS, TWO NORMAL CLASS WEEKS AND ONE EXAM WEEK.

From Figure IV, the numbers of ridership on Tuesday, April 18 and 25 are very similar, and even the patterns of peaks are fairly close, while on the exam week (May 2nd), the number of ridership is much less, and the pattern is different. One possible recommendation concluded is that the number of buses needed on the exam week is less than any other week as students take the exams and usually go home afterwards.

Figure V shows traffic patterns for four Wednesdays, April 2017 through May 2017. It is important to note that the last week of data collection was done during final exams week. It is for that reason that the graphs in May show a significantly reduced bus ridership traffic, compared to ridership during the semester.

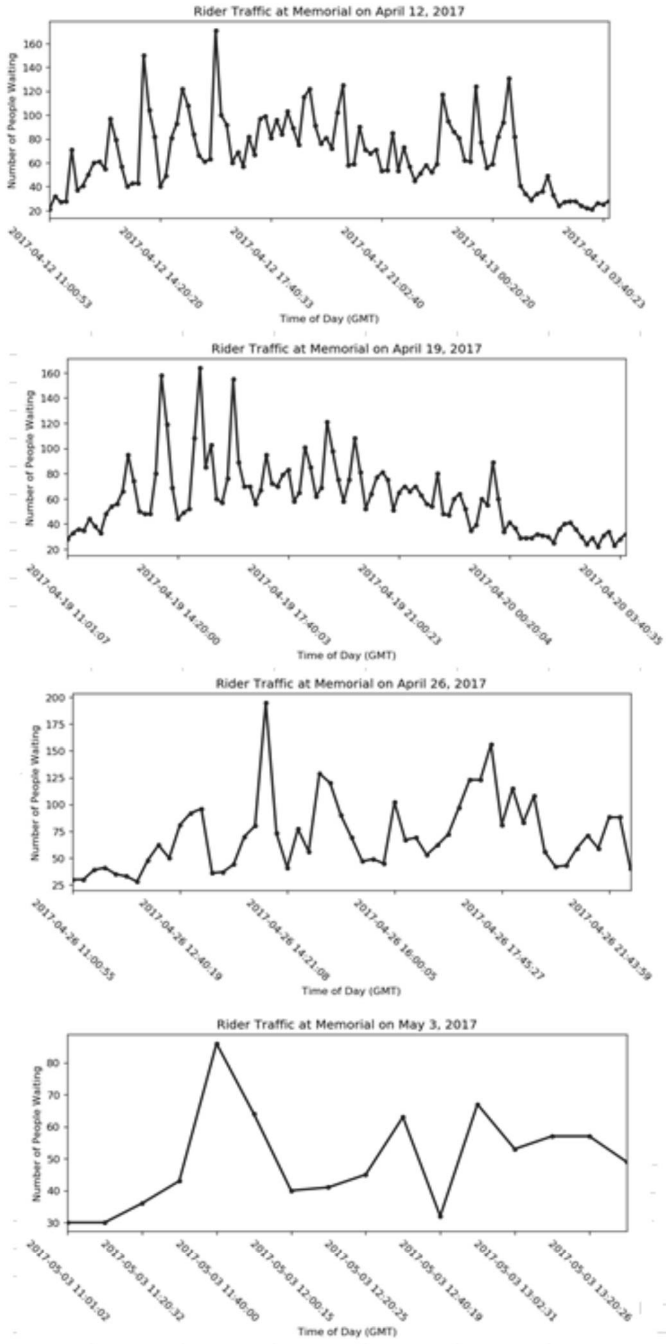

FIGURE V
WEDNESDAY'S BUS RIDER TRAFFIC AT MEMORIAL HALL STATION #011

It is worth mentioning that the bus ridership remains the same when compared with the same day of the week throughout the span of multiple weeks, or the entire semester. It is within this scope that we hope to find improvements to bus schedules and routes to increase public transport efficiency and reduce traffic around JMU and Harrisonburg.

III. Wait Time Frequency

The second type of data analysis is wait time frequency. We use the arrival and departure times of unique wireless devices within the range of the smart node to determine the number of minutes each rider waits at the Memorial Hall bus stop. Histograms show how often bus riders were waiting for lengthy periods of time for a bus as shown in Figure VI. We suspect many of the wait times recorded

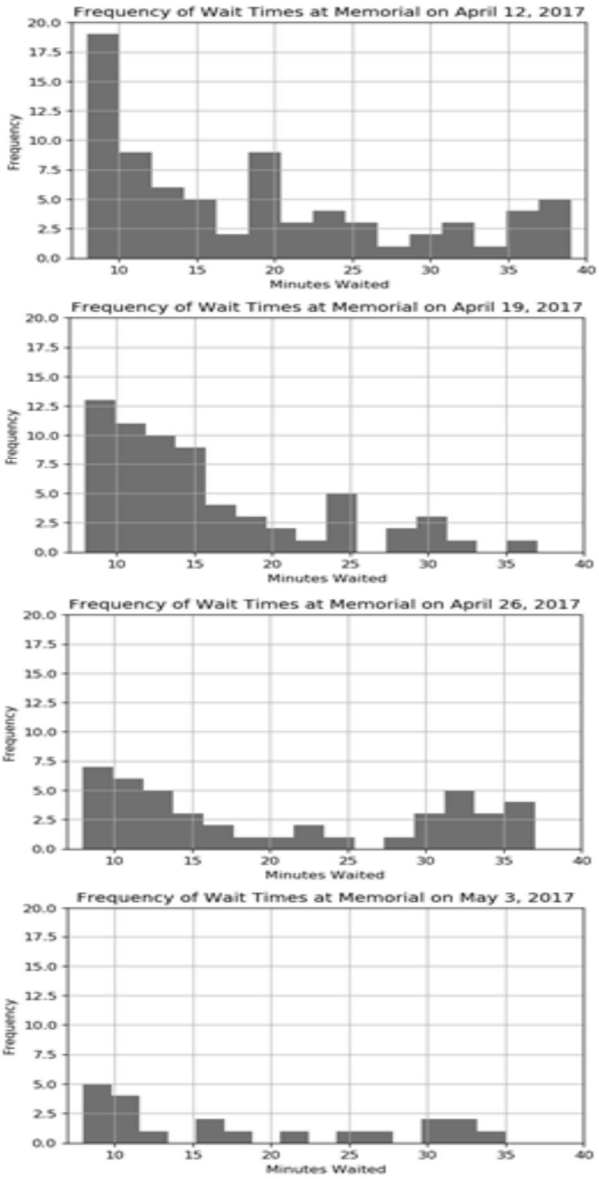

FIGURE VI
FREQUENCY OF WAIT TIMES AT MEMORIAL ON WEDNESDAYS IN APRIL 2017 AND MAY 2017

over 30 minutes to be instances of false positives which will be filtered out in future node deployments and data collections.

Figure VI displays the wait times recorded for the four Wednesdays in April 2017 through May 2017. As we would hope, the trends show similar wait times for each Wednesday, with the third of May showing lower frequency of wait times. This shows that the final exam week's patterns are different.

SECURITY MEASURES: PROTECTING PRIVACY

As explained in the overview section, we identify patrons at bus stations through the collected MAC addresses from their WiFi enabled devices. Collecting MAC addresses is necessary to uniquely identify patrons, but it might also raise a concern regarding privacy invasion [24]. For this reason, we decided instead of collecting MAC addresses, to use a hash function to generate a hash value from the MAC address and then store this value instead. A hash function takes an arbitrary message size as an input and produces a fixed message digest, known as hash value, as shown in Figure VII. One of the important security properties in a hash function is that given a hash value of a message, h(m), it is computationally infeasible to find the original message m. This property ensures that storing the hash value of the MAC addresses will not violate the privacy of the patrons.

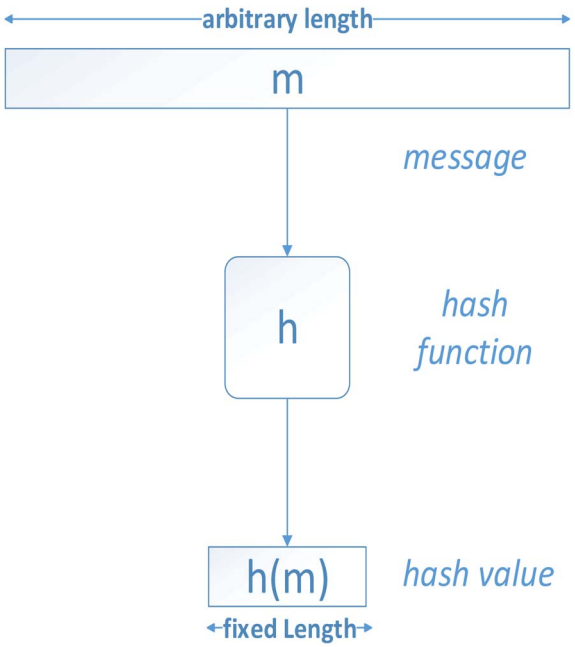

FIGURE VII
DERIVATION OF A HASH VALUE FROM A MESSAGE

We use the Secure Hash Algorithm (SHA-256) function to derive hash values. SHA-256 not only provides all the security requirements needed in a hash function as summarized in Table I, but also has a fast software implementation which ensures the hashing of the detected MAC addresses without the need to use a large buffer.

TABLE I
SECURITY REQUIREMENTS OF A HASH FUNCTION

Security Requirement	Given	Computationally Infeasible to Find
Preimage resistance	y	x, such that $h(x) = y$
Second preimage resistance	x & $y = h(x)$	$x \neq x$, such that $h(x) = h(x) = y$
Collision resistance	$x \neq x$,	such that $h(x) = h(x)$

The block size for SHA-256 is 512-bits and the size of a single MAC address is 48-bits. To bring the MAC address to the block size, it needs to be padded according to the padding scheme described by SHA-256 which can be summarized as follows:

- Start with original message of size n-bit
- Concatenate with a 1-bit of value '1'
- Concatenate with m-bit of value '0' such that m+n+1 = 448-bit
- Concatenate with the binary equivalent of n in 64-bit representation

Since the message used as an input to the SHA-256 function is always the MAC address, the padding size and value are fixed as shown in Figure VIII. This value is stored and appended with every MAC address detected then input to the hash function. The hash value is then stored along with the WiFi signal strength and the time stamp to be transferred to the database later.

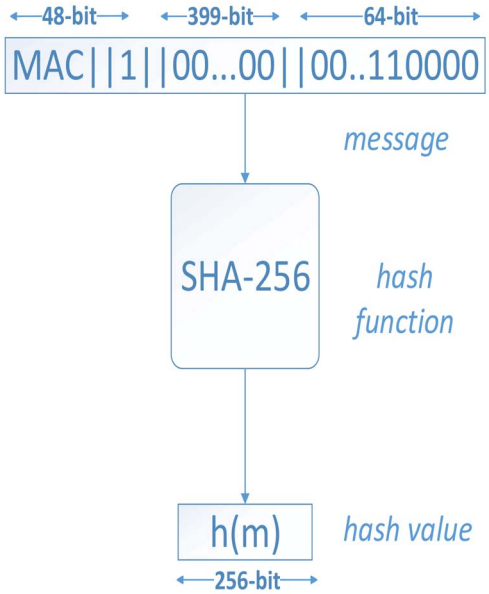

FIGURE VIII
PADDING OF MAC ADDRESS TO MATCH SHA-256 BLOCK SIZE

CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we show a sample of data analysis done on a cyber-physical system utilizing IoT for monitoring the quality of service of the transit bus system around an academic institution where the bus system is the primary mode of transportation. Results shows a pattern in ridership and waiting times around one of the seven stations where smart nodes were deployed. We also highlight how arrival and departure times can be used to calculate how long patrons are waiting for a public transit bus. Instead of assuming security of the smart nodes' data, we have assessed the possible actions to be taken by malicious actors as well as possible mitigation techniques, deciding on SHA hashing for the protection of riders' privacy. Future goals include origin and destination analysis, a mechanism of notification for the system administrator of security alert or possible attacks on a specific node and securing the location of the smart nodes.

ACKNOWLEDGMENT

This work was supported by the 4-VA Collaborative at James Madison University 4-va.org Fall 2017, as well as Global Research Laboratory Program through the NRF funded by the Ministry of Science, ICT & Future Planning (2013K1A1A2A02078326).

The authors would like to thank JMU Public Safety, and transit bus manager (Mr. Lee Eshelman) for his feedback on our experiments.

REFERENCES

[1] K. O'Flaherty, "Securing the Internet of Things, "SC Magazine UK, 2015.

[2] Y. W. Law, J. Doumen, and P. Hartel, "Survey and benchmark of block ciphers for wireless sensor networks," ACM Transactions on Sensor Networks (TOSN), vol. 2, no. 1, pp. 65–93, 2006.

[3] M. Tellez, S. El-Tawab, and H. M. Heydari, "Improving the security of wireless sensor networks in an IoT environmental monitoring system," in 2016 IEEE Systems and Information Engineering Design Symposium (SIEDS), April 2016, pp. 72–77.

[4] United States Department of Transportation. (2015, Nov.) Intelligent Transportation Systems. [Online]. Available: http://www.its.dot.gov/

[5] M. Garcia, P. Rose, R. Sung and S. El-Tawab, "Secure Smart Parking at James Madison University via the Cloud Environment (SPACE)," 2016 IEEE Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2016, pp. 271-276.

[6] A. D. May, Traffic Flow Fundamentals. Prentice Hall, 1990.

[7] S. El-Tawab and S. Olariu, "Communication protocols in FRIEND: A cyber-physical system for traffic Flow Related Information Aggregation and Dissemination," in IEEE International Conference on Pervasive Computing and Communications Workshops, March 2013, pp. 447–452.

[8] U.S. Department of Transportation. (2016, August) 2015 motor vehicle crashes: Overview. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318

[9] O. Popescu, S. Sha-Mohammad, H. Abdel-Wahab, D. C. Popescu and S. El-Tawab, "Automatic Incident Detection in Intelligent Transportation Systems Using Aggregation of Traffic Parameters Collected Through V2I Communications," in IEEE Intelligent Transportation Systems Magazine, vol. 9, no. 2, pp. 64-75, Summer 2017.

[10] S. El-Tawab, R. Oram, M. Garcia, C. Johns, and B. B. Park, "Data Analysis of Transit Systems Using low-cost IoT Technology," in First International Workshop on Mobile and Pervasive Internet of Things'17- 2017 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)., Mar 2017.

[11] M. Tubaishat, P. Zhuang, Q. Qi, and Y. Shang, "Wireless sensor networks in intelligent transportation systems, "Wireless communications and mobile computing, vol. 9, no. 3, pp. 287–302, 2009.

[12] R. Florin, P. Ghazizadeh, A. G. Zadeh, S. El-Tawab, and S. Olariu, "Reasoning about job completion time in vehicular clouds," IEEE Transactions on Intelligent Transportation Systems, vol. PP, no. 99, pp.1–10, 2016.

[13] J. Blum and A. Eskandarian, "The threat of intelligent collisions, "IT Professional, vol. 6, no. 1, pp. 24–29, Jan 2004.

[14] M. N. Mejri, J. Ben-Othman, and M. Hamdi, "Survey on vanet security challenges and possible cryptographic solutions, "Vehicular Communications, vol. 1, no. 2, pp. 53–66, 2014.

[15] M. Elhamshary, M. Youssef, A. Uchiyama, H. Yamaguchi, and T. Hi-gashino, "Transitlabel: A crowd-sensing system for automatic labeling of transit stations semantics," in Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services.ACM, 2016, pp. 193–206.

[16] M. Dunlap, Z. Li, K. Henrickson, and Y. Wang, "Estimation of Origin and Destination Information from Bluetooth and Wi-Fi Sensing for Transit," in Transportation Research Board 95th Annual Meeting, no.16-6837, 2016.

[17] Statista, "Forecast: number of smartphone users in the U.S. 2010-2018," Tech. Rep., 2015. [Online]. Available: http://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/

[18] "United states vehicle registration data, automobile statistics and trends,"Tech. Rep., 2015. [Online]. Available: http://hedgescompany.com/automotive-market-research-statistics/auto-mailing-lists-and-marketing

[19] A. Salem, T. Nadeem, M. Cetin, and S. El-Tawab, "Driveblue: Traffic incident prediction through single site Bluetooth," in 18th IEEE International Conference on Intelligent Transportation Systems, September 15-18, 2015.

[20] S. El-Tawab, R. Oram, M. Garcia, C. Johns, and B. B. Park, "Poster: Monitoring Transit Systems using Low cost WIFI Technology," in 2016 IEEE Vehicular Networking Conference (VNC), Dec 2016, pp. 1–2.

[21] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through wifi networks," in 2011 IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems, Oct 2011, pp. 192–201.

[22] B. Merino, How-to Instant Traffic Analysis with Tshark. Packt Publishing Ltd, 2013.

[23] Python Data Analysis Library [Online] https://pandas.pydata.org

[24] A. Salman, W. Diehl and J. P. Kaps, "A light-weight hardware/software co-design for pairing-based cryptography with low power and energy consumption," 2017 International Conference on Field Programmable Technology (ICFPT), Melbourne, VIC, 2017, pp. 235-238.

AUTHOR INFORMATION

Zachary Yorio, Graduate Student, James Madison University
Raymond Oram, Alumnus, James Madison University
Samy El-Tawab, Professor, College of Integrated Science and Engineering, James Madison University
Ahmad Salman, Professor, College of Integrated Science and Engineering, James Madison University
M. Hossain Heydari, Professor, College of Integrated Science and Engineering, James Madison University
B. Brian Park, Professor, Civil and Environmental Engineering Department, University of Virginia