DEVICE ASSOCIATION THROUGH PASSIVE WI-FI MONITORING

An Undergraduate Research Scholars Thesis

by

TRAVIS TAGHAVI

Submitted to Honors and Undergraduate Research
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Research Advisor: Dr. Jean-Francois Chamberland

May 2014

Major: Computer Engineering

TABLE OF CONTENTS

ABSTRACT

Device Association Through Passive Wi-Fi Monitoring. (May 2014)

Travis Taghavi
Department of Computer Science and Engineering
Texas A&M University

Research Advisor: Dr. Jean-Francois Chamberland
Department of Electrical and Computer Engineering

This research initiative is focused on identifying associated devices using a network of wireless sensors. These sensors collect network traffic meta data in monitor mode. Monitor mode allows a receiver to see certain aspects of all traffic within a network, including the source of data packets, regardless of their destination; leveraging this information with the radiation patterns of the wireless antennas, devices can be identified as geographically close at a specific instant. Over time, this data shows correlations between certain devices, and can be processed to create a network of relationships. Furthermore, the effect of antenna radiation pattern on the sensors' ability to associate devices is examined by using various antenna designs in the experiments. It is seen that antenna patterns which have smaller footprints are better able to distinguish between related and unrelated devices, at the expense of viewing fewer possible interactions. This relationship is analogous to a quality versus quantity decision.

NOMENCLATURE

Antenna footprint – the area which a specific antenna can monitor

Media Access Control address – MAC address, a unique identifier used by networked devices for communication

Monitor mode – an operating mode for a network interface card which allows the card to view all incoming traffic, regardless of who it is addressed to

Network interface card – a piece of hardware that connects a computer to a network

Radius of interaction – the radius in which, if a device sees a related device, it will stop to interact

CHAPTER I

INTRODUCTION

Wireless devices connected to a Wi-Fi network periodically send out messages with a header that includes, among other things, their media access control (MAC) address. MAC addresses are generally stored in hardware from the time of manufacturing, and are unique: meaning they can be used alone to identify a device. One way to take advantage of this is to passively monitor the wireless traffic in an area using a Wi-Fi receiver operating in monitor mode. This mode allows the network interface card (NIC) to pass all traffic meta data it sees to the processor, rather than discarding anything not addressed to it. The process of viewing and analyzing this information is called packet sniffing. Our experimental setting assumes an enclosed, featureless space with several wireless devices roaming freely. The idea is that by using multiple receivers in monitor mode with known wireless ranges, devices that are in the same general area at the same time can be identified. An assumption is made that related devices will more frequently be in close spatial proximity than unrelated devices. Thus, over time, a network of relationships between the available wireless devices can be inferred from the wireless packet data. A relevant study by Musa and Eriksson has established the possibility of tracking wireless devices through passive Wi-Fi monitoring by tracking smart-phones in cars with several road-side monitoring points. Their problem differs in that Musa and Eriksson tracked devices along essentially one-dimensional roads, and were more focused on observing a devices path along those roads. The current study focuses on establishing relationships by placing a small set of devices in a two-dimensional closed space, and observing the effect of antenna patterns on performance.

The objective with this project is to infer a network of related wireless devices purely through the

information provided by packet sniffing and the antenna patterns used, and study methods of optimizing this process. Meta data is collected from multiple sensor nodes and subsequently analyzed. It is shown that, given the right arrangement of wireless nodes and antennas, meaningful relationships between visible wireless devices can be inferred.

This project consists of multiple stages: preparation, data collection, and analysis. Preparation involves setting up the wireless sensor network and writing simulation code. For the wireless nodes, the ability to access the NIC as well as to change the wireless antenna is required. One reasonable choice for this is a small Linux computer coupled with a wireless adapter. Once the devices for the sensor nodes are selected, they need to be programmed to synchronously collect data in monitor mode and forward this data to a central node for post-processing. The simulation code is used to model the environment to be tested: a featureless area with devices that can move about the area randomly. The code also records when devices enter certain spaces within the area, representing antenna footprints. The data collection stage for the simulation involves running the tests for an extended period of time. For the physical test, a single wireless node is set up in the laboratory and run for close to 4 days, constantly taking 5 minute long “snapshots” of the unique MAC addresses visible to it. With more time, this test would be run with varying antenna strengths and footprints. The analysis stage is performed after the experiments are finished (i.e. not in real-time). The data is run through an algorithm that takes in the “snapshots” of unique MAC addresses, and outputs a weighted graph of relationships between these addresses. Essentially, the algorithm looks for devices that are in the same antenna footprint at the same time, and increases the confidence of a relationship between those devices.

CHAPTER II

DESIGN CONSIDERATIONS

As stated in the introduction, the procedure for testing and evaluating these hypotheses is two-fold. First, a simulation program is written from scratch to model the environment to be tested, as well as to perform preliminary tests on efficiency of various antenna patterns and graph inference techniques. The other part of the procedure is to assemble physical system of one or more wireless nodes, and test it in a real, active environment.

The simulation program is written entirely in C++, using a library called the Simple and Fast Multimedia Library (SFML) in order to connect to OpenGL to provide a visualization of the running simulation. C++ is a valuable language to use for simulations as it combines little overhead with ease of use. The visualization is particularly useful in this simulation because there are many factors that, although they can be described completely mathematically, are very intuitively understood with a graphical interface. These include the modeled antenna patterns, device density, and most importantly the movement of devices in an open, confined space. The simulation program uses a simplified model of the environment that is tested physically. The finite, enclosed space is modeled as a two-dimensional grid of finite resolution. The size of the "spaces" on the grid are representative of about the average footprint of a human being carrying a wireless device: around 1 square foot. Thus, a 30 by 30 unit grid in this simulation is representative of a 30 foot by 30 foot closed room. Each space on the grid may or may not be occupied by a device, and also may be visible to one or more antenna patterns. The antenna patterns, then, are modeled as a group of (usually, but not necessarily) continuous spaces on the grid. Visually, they are shaded in different light colors. While this is a simplification of how antenna patterns

actually work, it is a reasonable simplification in that it should not make a noticeable functional difference in this model. The most important simplification made for the simulation model is in the movement of devices. A Gaussian random walk is implemented as a model of human movement within a confined space. The random walk involves, for each time step, first deciding whether or not to move (based on a set probability), and then in which direction to move (based on a different set probability). Clearly, this is not perfectly indicative of how humans move. There is no velocity associated, so a given device is just as likely to turn at any moment as it is to proceed forward. However, the random walk model is sufficient to propagate the devices around the entire grid randomly, which is the main consideration in this case. The graph inference algorithm for the simulation is the same as for the physical system. In other words, data is processed in the same way in both situations.

For the physical system, there are several different considerations. First and foremost, the wireless adapter to be used for this experiment needs to be chosen. This decision is narrowed by the requirement that the adapter have the ability to operate in monitor mode. Monitor mode is a powerful tool that can be used for many types of analysis, including some uses that are not legitimate (i.e. hacking, spying, etc.). As such, many manufacturers choose to not make monitor mode available on their adapters. One manufacturer that does allow for their NIC chipsets to be placed in monitor mode is Atheros. The TP-LINK TL-WN722N Wireless N150 High Gain USB Adapter is selected for use in this project, as it utilizes the Atheros chipset. The next major consideration for building the physical system is the actual computers to use for each sensing node. Any reasonably powered computer, from a full desktop to a simple single-board computer, is powerful enough to perform the tasks necessary for this project. For ease of use and convenience, we employ Intel's Next Unit of Computing (NUC) for use in data

collection. The NUC has sufficient power to perform the necessary operations, and has the advantage of a small form-factor. Since smartphones and laptops are ubiquitous and are commonly Wi-Fi enabled, they are used as the wireless devices for the physical test. The final major design consideration for the physical system is the antennas to use for data collection. Since only one physical test is run in this project, a cantenna is selected. A cantenna is a simple, fairly directional, antenna constructed from a metal can. This cantenna is tuned to 2.4 gigahertz, the most popular radio band for Wi-Fi.

CHAPTER III

METHODOLOGY

As mentioned previously, there are two stages of experiments for this project. First, we test the performance of the idealized system using a simulation, which is written in C++. Data collected from this simulation model is then processed using the same graph inference algorithm as the actual physical tests. After simulations verify the concept and model, an extended physical test is run.

To test the simulation model, it is only necessary to test with two devices: once when they are related, and once when they are not. In other words, the unrelated simulation has two devices randomly walking around the area and ignoring each other completely. In the related test, the two particles are also randomly walking about the area, but when they cross paths, they stop for a random amount of time to simulate an interaction. The final output relationship graphs for these two tests only contain two nodes and one weighted edge. If the weight of the edge in the related test is consistently higher than that in the unrelated test, then the model is sufficiently able to distinguish between related and unrelated devices. Adding more devices and relationships is interesting to view, but essentially amounts to a linear combination of these related and unrelated tests. So, using these two scenarios, the data is evaluated using varying antenna footprint sizes, showing the relationship between the antenna footprint size and the strength ratio of the related and non-related edges in their respective graphs. In the random walk, the probability that any given device will move, provided it is not interacting with another device, is 75 percent.

For the physical test, as mentioned in the design considerations, smartphones and laptops are used as

the wireless devices. For this project, the physical tests of the system are not run as extensively as the simulations. This is due to the fact that a simulation can run the equivalent of 10 days of physical tests in under a minute. Beyond this, varying antenna footprint shapes and sizes is done almost instantaneously in code, whereas in the real world this takes a large amount of analysis and testing. So, the physical test run in this research project is more to validate that this physical system functions properly, rather than to gather a large amount of valuable data to analyze. The physical system is set up in the laboratory, where there is a lot of foot traffic, and many of the MAC addresses of devices are known. The test aims to see if certain pairs of devices that are known to be related result in strong edges in the final relationship graph.

CHAPTER IV

RESULTS

As described in the methodology, the main results of this project come from the computer simulations. The goal of the simulations is to find a correlation between the size of an antenna footprint and the performance of the system. Figure 1 in the appendix shows the results of a set of 10 million time-step simulations with increasing antenna footprint size on a 20 by 20 unit grid, with a probability of moving for the random walk of 75 percent. Figure 2 shows the same simulations run in a larger environment, specifically a 30 by 30 unit grid. As a general trend, it is observed that smaller antenna footprints yield a better result, in other words a higher ratio of related to unrelated edge strength in the relationship graph. Viewing the charts of these results, particularly for the larger environment, we can see that there is a point at which the footprint can be too small, and detrimental to performance. This is partially due to the ratio of environment size to antenna footprint size making the likelihood of interaction within the footprint's area small. It is also partially due to the antenna footprint size becoming close to or smaller than the radius of interaction of two devices. It is logical that both charts seem to tend towards a ratio of one as the antenna footprint size is increased, and no longer indicates relatively close physical proximity.

The results of the physical test are in the form of a relationship graph, output in a text file ordered by edge strength. The test covers 1050 time steps of 5 minutes each, for a total time of 87.5 hours. The graph shows a few MAC addresses that are connected to other MAC addresses with a strength of 1050, indicating that these devices were present for the entire duration of the test, and do not provide valuable

data. After many edges with a strength at or near 1050, the strength drops dramatically. Although the data is difficult to analyze due to these edges, it is likely that edges which are strong, but below the initial group of very strong edges, indicate a true relationship between devices.

CHAPTER V

DISCUSSION

The results from the simulations are promising. A general trend has been observed that correlates antenna footprint size with the quality of the data for graph inference. The smaller footprints yield a large ratio of around or above 5 between the related and unrelated tests, indicating that with the correct antenna pattern, this system is able to distinguish between related and unrelated devices in an area over time. As stated previously, the quality of data drops as the antenna footprint grows larger. This is seen in both the 20 by 20 and 30 by 30 cases, with the right side of the graph tending towards one. This is consistent with the intuition that a large enough antenna pattern will see related and unrelated devices equally, and can no longer be said to indicate a likely interaction between devices. In the 30 by 30 case, there seems to be a sweet-spot that balances the quality of the data with the amount of data that is gathered, yielding optimum performance. This is to be expected, and is a promising avenue for further research on this topic. Further work on this simulation potentially includes improving the random walk to more closely simulate human motion, as well as testing differently shaped antenna footprints.

The physical test shows that the system functions as expected, and test results highlight an important issue with the implementation of the system. It is apparently necessary to identify the MAC addresses of stationary devices that are within range of the antenna. These stationary devices can include desktop computers and wireless access points. Since these devices are not identified and ignored by the data collecting unit, their presence clouds the resulting relationship graph by showing a strong relationship between these stationary devices and all other devices that are observed. Further research may include identifying stationary devices in real time and excluding them from results, as well as varying antenna

footprint patterns.

APPENDIX

Pattern Size(n by n)	Unrelated	Related	Ratio
2	490	2711	5.533
3	3981	20065	5.040
4	14695	55393	3.770
5	38284	122572	3.202
6	77937	227532	2.919
7	149289	355596	2.382
8	257472	539431	2.095
9	427344	740534	1.733
10	651736	1139270	1.748
11	891929	1418080	1.590
12	1112430	1781360	1.601
13	1812390	2259760	1.247
14	2283560	2906070	1.273

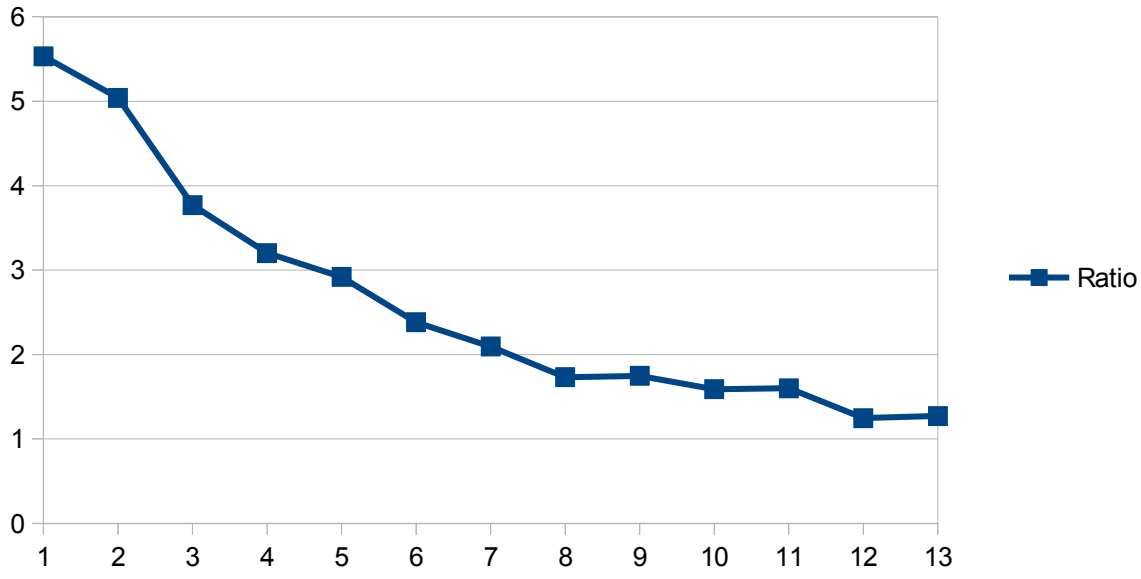

Figure 1

Pattern Size(n by n)	Unrelated	Related	Ratio
2	105	645	6.143
3	823	6228	7.567
4	2818	14713	5.221
5	7245	30351	4.189
6	15070	46775	3.104
7	28166	82621	2.933
8	48375	123606	2.555
9	78447	176915	2.255
10	120746	245383	2.032
11	177735	333566	1.877
12	250730	445312	1.776
13	349225	545878	1.563
14	465579	728646	1.565

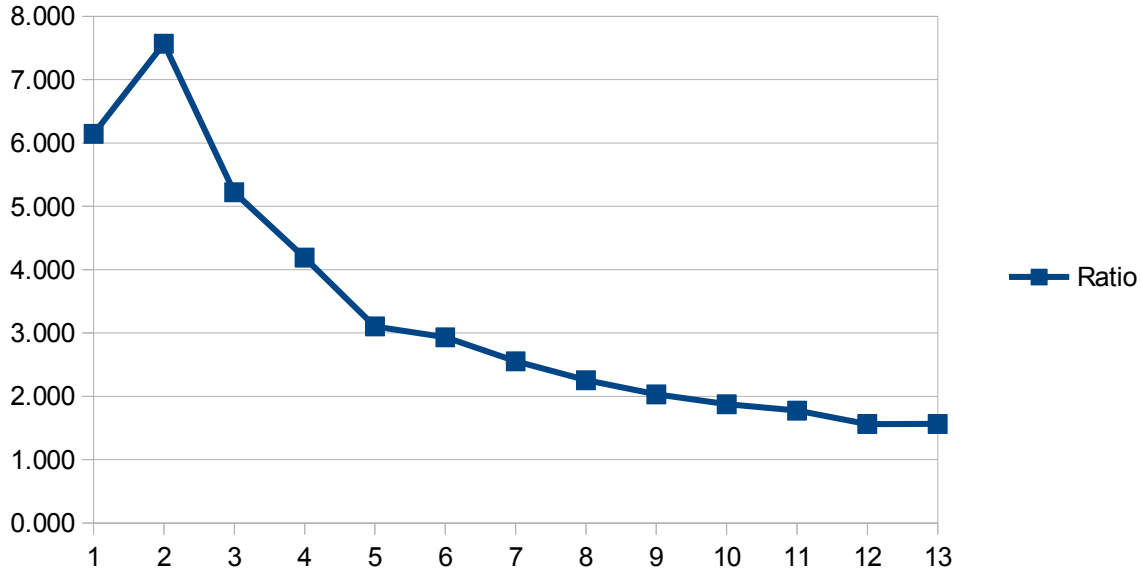

Figure 2

REFERENCES

A. Musa and J. Eriksson “Tracking Unmodified Smartphones Using Wi-Fi Monitors”, SenSys’12, November 6–9, 2012, Toronto

150Mbps High Gain Wireless USB AdapterTL-WN722N." *TL-WN722N*. N.p., n.d. Web. 26 Jan. 2014.

"A Pint-sized Powerhouse." *Intel*. N.p., n.d. Web. 26 Jan. 2014.