Physical Layer Attacks on Unlinkability in Wireless LANs

Kevin Bauer[1], Damon McCoy[1], Ben Greenstein[2],
Dirk Grunwald[1], and Douglas Sicker[1]

[1] University of Colorado
{bauerk,mccoyd,grunwald,sicker}@colorado.edu
[2] Intel Research Seattle
benjamin.m.greenstein@intel.com

Abstract. Recent work has focused on hiding explicit network identifiers such as hardware addresses from the link layer to enable anonymous communications in wireless LANs. These protocols encrypt entire wireless packets, thereby providing unlinkability. However, we find that these protocols neglect to hide identifying information that is preserved within the wireless physical layer. We propose a technique using commodity wireless hardware whereby packets can be linked to their respective transmitters using signal strength information, thus degrading users' anonymity. We discuss possible countermeasures, but ultimately we argue that controlling information leakage at the physical layer is inherently difficult.

1 Introduction

The inherent broadcast nature of wireless communications coupled with the widespread availability of commodity receivers poses a significant privacy concern for users of wireless technology. The threat that third parties who eavesdrop on communications may profile users and track their movements is well understood [1,2]. Even when message confidentiality is provided by standards such as WPA for 802.11, only the payload is protected and every user's identifying MAC address is revealed. This enables any third party within signal range to monitor and track other users in the network.

To eliminate the transmission of identifying information at the link layer, recent work has focused on providing identifier-free link layer protocols that encrypt all transmitted bits to increase privacy with respect to third party eavesdroppers [3,4,5]. By obfuscating all bits of the frames including the addresses, these protocols attempt to provide *unlinkability*, since it is difficult for unintended recipients to associate sequences of packets to their source transmitters.

Despite these protocols, we demonstrate that information derived from the physical layer can be applied to classify packets by their respective transmitters, thereby violating this unlinkability property. While we focus our study on a variant of 802.11, we believe that the fundamental problem of information leakage

I. Goldberg and M. Atallah (Eds.): PETS 2009, LNCS 5672, pp. 108–127, 2009.

at the physical layer exists in a wide variety of other wireless protocols including WiMax, 3G, 4G, and future protocols that do not protect the physical layer.

Our approach is based on recording the strength of received signals from devices at several locations and applying a clustering algorithm to perform packet source classification. The method is practical, since it utilizes commodity hardware instead of expensive signal analyzers (as in previous work [6,7]) and requires no training or cooperation from the wireless devices in the network.

While this approach can determine which packets originated at the same source, it won't identify sources by name. However, we demonstrate that the packet source classification is accurate enough to enable complex traffic analysis attacks which use features such as packet size to reveal more about who the user is and what he/she is doing. Examples of the types of information that can be inferred through traffic analysis attacks include videos watched [8], passwords typed [9], web pages viewed [10,11], languages and phrases spoken [12,13], and applications run [14]. These traffic analysis attacks become more dangerous when coupled with additional information such as visual identification of users.

Results. In order to demonstrate the efficacy of this method, we evaluate the technique by conducting experiments in a real indoor office building environment. We apply the packet clustering technique, which uses well-known statistical methods, and the results show that packets are correctly linked to their transmitting devices with 77–85% accuracy, depending on the number of transmitters in the network. As more sophisticated techniques may be applied in the future, we consider these results as a lower bound on attainable accuracy.

Since the clustering method is often imprecise, we evaluate how the reconstructed sequences of packets can be used to perform a previously described website fingerprinting traffic analysis attack [10,11]. While any number of traffic analysis tasks could be performed, we chose website fingerprinting because web browsing is among the most common on-line activities. Our results indicate that a website can be identified 40–55% of the time from source classified packets, depending on the number of devices in the network.

Toward Solutions. Finally, we explore methods to mitigate the effectiveness of source classification using information derived from the physical layer. We evaluate solutions based on transmit power control and directional antennas and show that these techniques make source classification more difficult. However, we observe that altering the properties of the wireless physical layer is fundamentally challenging and we recognize that additional research attention should be focused on addressing information leaks at the physical layer.

Contributions. This paper has three primary contributions:

1. We explore a source of identifying information contained within the wireless physical layer and show that it can be used to violate the unlinkability property of anonymous link layer protocols.
2. We present and experimentally validate an unsupervised statistical technique to perform packet source classification that is robust to the inherent noise

of the RF space and is accurate enough to enable complex traffic analysis tasks to be performed.
3. We experimentally investigate methods to mitigate source classification by altering signal strength properties. While these techniques mitigate the accuracy of packet source classification and subsequent traffic analysis to some extent, we argue that information leakage at the wireless physical layer presents a particularly challenging privacy threat.

2 Background

Traditional anonymity. Anonymous communications have historically been facilitated by mix networks [15] and onion routing networks [16]. Fundamentally, these networks attempt to hide a message's sender and receiver from an adversary residing within the network. This requires that network layer identifiers such as source and destination IP addresses and other transport and application layer identifiers be hidden.

However, due to the inherent broadcast nature of wireless, there is a significant threat that an eavesdropper within range of a wireless signal may use persistent explicit identifiers found at the link layer (such as a MAC address) to uniquely identify users, and subsequently track their movements and profile their activity. This threat presents a serious privacy concern for users of wireless technology such as the ubiquitous 802.11 standard and an even greater threat to users of wide area networking devices, such as WiMax and 4G. These long range protocols allow an attacker potentially up to one mile away from the transmitting device the ability to eavesdrop. While mix network and onion routing techniques hide identifiers at the network layer and above, they were not designed to provide anonymity at the link layer. Thus, additional anonymity mechanisms are necessary to obscure these identifiers found at the link layer.

Anonymity in wireless networks. Several strategies have been proposed to address the leakage of identifying information within wireless networks. Gruteser and Grunwald suggest that disposable interface identifiers replace explicit identifiers such as the MAC address to mitigate location tracking and user profiling [17]. Arkko *et al.* propose a generic technique that replaces identifiers such as the MAC address with pseudo-random values drawn from a random number generator seeded with a shared secret [18]. This approach may also be used to obfuscate other identifiers at higher layers of the protocol stack such as IP addresses and TCP sequence numbers. During the session initiation, a mutually agreed-upon seed value is derived by the wireless client and access point. However, it is necessary to share seed values for every potential identifier and this general approach does not hide identifying information revealed by the application layer. A similar approach has been proposed using protocol stack virtualization [19]. This general approach enables the identifiers to change for each packet sent, thereby increasing the size of a wireless client's anonymity set to the number of clients participating in the wireless LAN.

To address the limitations of this general approach, link layer encryption has been proposed to obfuscate all bits transmitted in the wireless frames [3,4,5]. This hides any identifying information contained in the transmission, including explicit identifiers. At the link layer and above, these packets are unlinkable to their senders. However, we show that these protocols that hide explicit identifiers are limited since they do not address the physical layer.

Physical device fingerprinting. Recent advances in physical device fingerprinting technology have introduced the possibility of identifying specific devices. Kohno *et al.* demonstrate that minute, yet distinguishable variations in a device's clock skew persist over time and can be detected remotely without any cooperation from the targeted device [20]. This technique has also been extended for the purpose of locating hidden services within the Tor network. [21,22].

Beyond the identifying characteristics of clock skew, RF-based device identification techniques have been previously proposed. Gerdes *et al.* show that Ethernet interface cards can be uniquely fingerprinted by their varying RF properties [23]. In the wireless context, techniques have emerged for fingerprinting distinct 802.11 interface cards based on the observation that minor flaws in device manufacturing are often manifested as modulation errors [6,7]. Both works propose a machine learning-based identification framework to detect specific modulation errors and empirically demonstrate that the techniques can identify distinct 802.11 cards with over 99% accuracy. While these techniques require expensive signal analyzer hardware, they represent a significant privacy risk to wireless users, especially if the required hardware becomes inexpensive.

Device driver, OS, and user fingerprinting. In addition to physical device fingerprinting, techniques have been developed to remotely identify device drivers of wireless network interface cards, a device's operating system, and even specific users. Probing tools such as Nmap [24] and p0f [25] are widely available to remotely scan ports, determine what operating system (and version) is running, and obtain information about packet filters and firewalls. Such information could potentially be used to aid in identifying and profiling devices. Franklin *et al.* present a passive device driver fingerprinting technique based on the wireless device driver's active probing behavior that can identify specific drivers with high accuracy [26]. Device driver information could also contribute to identifying and profiling wireless devices. Pang *et al.* and Aura *et al.* show that implicitly identifying information can inadvertently leak during wireless communication sessions [1,2]. Examples of such information include service discovery for specific wireless networks, file shares, and networked printers. Even more latent information sources can be uniquely identifying, such as websites viewed or applications used.

Physical device localization. Localization systems such as Place Lab allow wireless devices to passively localize themselves in physical space [27]. A wireless device can identify its location by comparing their beacon observations that identify the nearby stationary wireless infrastructure to a database of prior beacon

observations tagged with physical location information. Widely deployed commercial services such as Skyhook [28] use this technique to help wireless devices perform self-localization.

There also exist a variety of techniques that enable the wireless infrastructure to localize wireless devices based on the physical layer properties of their transmitted signals. The most common approach to infrastructure-based wireless localization applies a supervised learning approach and uses commodity wireless cards. During the training phase, signal strength measurements are collected from several positions throughout a target environment (such as an office space) to train a machine learning algorithm. RADAR uses the k-nearest neighbors classifier to compute the wireless signal's physical position [29]. Other methods use a naïve Bayes classifier for location estimation [30]. While the training procedure can be expensive and time consuming, they are relatively accurate in practice. Other approaches often require specialized non-commodity hardware. Such approaches include estimating a signal's angle of arrival and applying triangulation [31], calculating time of arrival (*i.e.*, the global positioning system) [32], and applying time difference of arrival techniques [33].

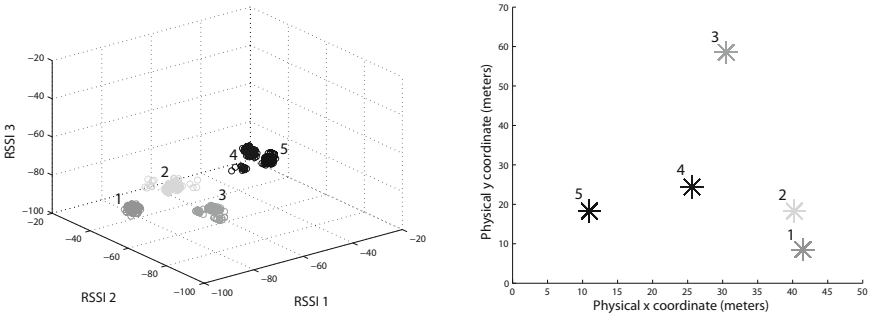
The ease with which a device's location can be estimated from its signal properties presents significant privacy risks. Gruteser and Grunwald present algorithms and middleware that enable anonymous usage of location-based services [34]. Their approach is based on manipulating the resolution of location information along space and time dimensions. However, this solution assumes that the wireless client provides its own location information to a location server that implements the location privacy middleware. It does not address the scenario in which an adversary uses signal strength information to locate and track other users. Jiang *et al.* propose a solution to enhance location privacy based on randomized MAC address pseudonyms and silent periods to help decouple pseudonyms from devices [35]. In addition, this work explores the application of transmit power control to reduce the precision of localization algorithms by reducing devices' transmit power levels such that a only minimal number of listening access points can hear and localize the signals.

Inferring identity from the physical layer. Physical layer information has previously been used to detect identity-based attacks (such as MAC address spoofing) in wireless networks [36]. Since signal strength varies with physical location, a rogue device has distinct signal strength readings from the expected device, assuming that they are transmitting at different locations. Therefore, a device's identity is linked to its physical location. This observation can be useful for determining whether an identity-based attack is taking place. We rely on this fact in the design of our packet source classification technique.

3 Packet Source Classification

In this section, we first provide the necessary background and intuition behind the packet source classification techniques. Next, we describe the design of the

RSS-Localization based technique that can be used to perform packet source classification. However, it requires an expensive training process to learn the relationship between signal strength and physical location. To address this limitation, we present RSS-Clustering, a packet source classification method that does not require training.



(a) RSSI values (in dB) from three sensors for five different devices

(b) The corresponding physical locations of the five devices

Fig. 1. A visualization of the RSSI values from transmitters at five different locations

3.1 Background and Intuition

When a commodity 802.11 wireless card receives a packet, it records the signal strength of the received packet as a received signal strength indication (RSSI) value. The RSSI value reported by standard 802.11 hardware is measured only during the reception of a message's preamble, which is transmitted at the lowest rate (1 Mb/s). In a simplified signal propagation model, wireless signals fade with distance as they propagate over physical space. Thus, the RSSI values are (roughly) inversely proportional with the distance between the transmitter and receiver. This means that the same transmission will be received at different RSSI values depending on the distance between the transmitter and receiver. Using these RSSI values, we show that it is possible to passively associate a set of packets to their source device.

However, several factors affect a packet's RSSI value in real world environments, which makes accurately associating packets to their transmitting devices using physical layer information a very challenging task. At one receiver, the RSSI values of different packets from the same transmitter often vary over time due to noise factors such as multipath interference and unpredictable fading [37]. Figure 1(a) shows the RSSI values recorded from multiple packets sent over time from five distinct transmitting devices whose corresponding physical locations are given in Figure 1(b). While the values are similar for each device, there is some unpredictable, but small fluctuation due to the inherent noise in the physical environment.

3.2 RSS-Based Localization

RF-based localization is a well-studied problem in which wireless devices are physically located using the signal strengths of their transmitted packets. Therefore, it is reasonable to try this localization strategy to perform packet source classification, since these methods have been shown to provide accurate device localization to within about three meters of the device's true location [29].

The localization technique uses the k-nearest neighbors supervised learning framework, as in previous work [29] to perform packet source classification.[1] Beyond source classification, this approach has the ability to add semantic location information, which could be used to associate packets to a particular device or user and thereby reconstruct persistent identities.

However, localization requires that the adversary collect training data for every environment in which they wish to perform this attack. Furthermore, the training process must be repeated if environmental changes occur. This training data collection is very expensive and even unnecessary, since our goal is not to localize packets, but instead is to perform packet source classification.

3.3 RSS-Based Clustering

To address the limitations of the localization approach, we propose RSS-Clustering, an unsupervised technique to perform packet source classification. Since the RSSI values are inherently noisy, we use the k-means clustering algorithm [38] to group packets by their respective transmitting devices. In order to perform source classification, k-means requires the RSSI feature vectors and the number of devices (k), which we assume is known (or can be closely estimated) by the attacker using visual information or one of many techniques to determine the number of clusters in a data set [39,40,41,42]. While k-means is a computationally efficient linear-time algorithm, it is stochastic and therefore, not guaranteed to produce a globally optimal clustering solution. For this reason, it is common to execute k-means several times on a data set to arrive at a stable clustering result.

There exist several classes of cluster analysis algorithms, including hierarchical, partitional, and spectral techniques [38]. We chose k-means for its simplicity and strong performance on our clustering task. However, it is possible that other clustering algorithms may offer better performance or relax the requirement that the number of clusters be known in advance. Consequently, we consider the results obtained with k-means to be a lower bound on attainable performance.

4 Threat Model

In this section, we enumerate our assumptions about the attack, the adversary, and the victims.

[1] Since these localization techniques have a certain amount of error, it is necessary to cluster the imprecisely localized packets by estimated location.

Attack. An eavesdropper first performs packet source classification and subsequently uses the sequences of encrypted packets associated with their respective transmitters to perform complex traffic analysis tasks. The attack is completely passive, so users can be subjected to it without their knowledge. In addition, this technique requires only commodity 802.11 hardware.

Adversary. We consider the adversary to be a person or group of people with limited resources and access to only commodity 802.11 hardware. The adversary has the ability to place n passive commodity 802.11 wireless sensors in chosen positions around a target location (such as a building). For each received packet p_i, the RSSI values across all sensors are combined into a feature vector $(RSSI_{i1}, RSSI_{i2}, ..., RSSI_{in})$. Also, the attacker has the ability to estimate how many devices are present in the area.

Victims. It is trivial to classify packets when it is known that only a single device is active at any particular time, *e.g.,* at a public hotspot. However, we assume a more common situation in which several devices may transmit at arbitrary times, possibly with interspersed transmissions. A prior analysis of wireless traces has shown that there are often many simultaneously active devices at tight time scales [4].

The victims use a standard 802.11 wireless device to communicate using an identifier-free link layer protocol and transmit at a constant power level. Also, the victims use a common application such as a web browser. They remain stationary while they transmit, but are free to move when their transmitters are silent.

5 Experimental Validation

To demonstrate the efficacy of the physical layer source classification technique, we present a series of experiments conducted with 802.11 devices in a real indoor office building environment. In this section, we describe the methodology used to collect real RSSI values. To understand how the packet source classification techniques performs in practice, we present metrics with which to evaluate their ability to accurately associate packets to wireless devices. We characterize the clustering technique's performance with respect to how the number of devices effects clustering accuracy and how the number of listening sensors effects accuracy. Our results show that this method is highly accurate even when 25 devices are active at the same time and requires few sensors.

5.1 Experimental Setup

In order to understand how our physical layer packet clustering technique works in practice, we deployed five 802.11 wireless devices to act as sensors in the "Center for Innovation and Creativity" building located on the University of Colorado's Boulder campus. Deploying five sensors ensures that signals can be received when transmitted from nearly any position in the building, and multiple

overlapping sensors also increases the accuracy of our method. This single-storey office building measures $75\,m \times 50\,m$. Each sensor, a commodity Linux desktop machine, passively listens for packets on a fixed 802.11 channel. This allows the sensors to record RSSI values from all audible packets on that particular channel. To collect RSSI measurements, we used a laptop computer to transmit 500 packets at a constant power level of 16 dBm at 58 distinct physical locations throughout the office space (see Appendix A for detailed hardware specifications).

In addition, to evaluate the localization approach, we collect RSSI readings from 179 additional training locations at a constant 16 dBm transmit power level. The k-nearest neighbors algorithm is used for localization and we verify that the median localization error is approximately 3.5 meters, which is consistent with prior work [29]. The layout of the office space marked with the positions of the passive sensors, training locations, and device locations is provided in Appendix B.

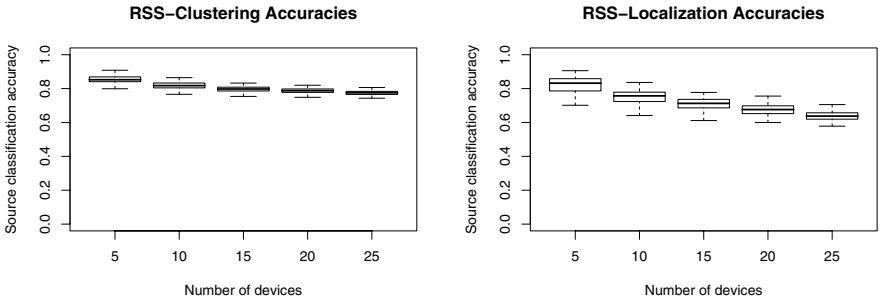
To evaluate how the number of devices effects the accuracy, we vary the network size from 5, 10, 15, 20, to 25 devices. Since we only used a single wireless device to transmit packets at multiple locations, to construct scenarios with multiple devices we generated traces of packets transmitted at multiple locations. However, during the data collection, there were other wireless devices transmitting which added interference to the RF space. In order to ensure that there is no bias in the selection of the devices' locations that may influence performance, we generate 100 randomly chosen device location configurations for each network size[2]. Next, we perform clustering on these device location configurations. Recall that since k-means is not guaranteed to provide a globally optimal solution, it is necessary to perform the clustering several times to arrive at a stable clustering solution. We observed that the algorithm stabilized after approximately 100 runs, which takes approximately one minute to complete on a 3.6 GHz Pentium computer. Therefore, we perform k-means clustering 100 times on each device location configuration.

To measure clustering accuracy, we apply the standard *F-Measure* metric from information retrieval. The F-Measure is a weighted harmonic mean *precision* and *recall* in which both are weighted equally [43]. Within the context of our clustering problem, precision captures the homogeneity of each cluster. Recall measures the extent to which packets from a given device are clustered together.

5.2 Packet Source Classification Results

We next present the results of the physical layer packet clustering technique in terms of its ability to accurately associate packets with their respective transmitting devices. In particular, we examine two factors that we believe to be significant with respect to clustering accuracy: (1) the number of devices in the observation space, and (2) the number of sensors in the observation space.

[2] Although we collected RSSI measurements at 58 distinct positions, we chose to limit the number of devices to 25 in any experiment to allow for variety in the randomly chosen locations of the devices included in the experiments.



(a) RSS-Clustering method (b) RSS-Localization method

Fig. 2. Packet source classification accuracies as the number of devices increases

Effect of number of devices on accuracy. The average packet source classification accuracy ranged from 85–77% as the number of devices in the network was varied from 5–25, as shown in Figure 2. In general, the accuracies decrease as the number of devices increases. In other words, the clustering algorithm performs better on a smaller number of devices and produces additional clustering errors as more devices are introduced. However, the 20 and 25 device experiments produced similar clustering accuracies, so there is evidence that the clustering accuracy may, in fact, level off as the number of devices reaches a critical threshold. Additionally, within all device configurations, the RSS-Clustering method provided slightly better accuracy than the RSS-Localization approach.

Effect of number of sensors on accuracy. As shown in Figure 3, the clustering accuracy is surprisingly high, ranging from 75–47%, when just one sensor is used for clustering. However, as more sensors are added, the accuracy for each configuration increases gradually, with diminishing returns: as the number of sensors increases from three to five, the accuracy only improves by at most 3%. This indicates that the resources required—in terms of number of sensors to deploy—are very minimal, making the packet clustering technique practical for a low resource adversary.

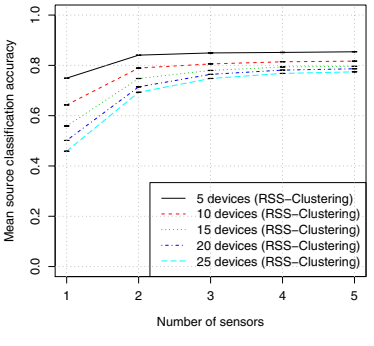


Fig. 3. Mean source classification accuracies (with 95% confidence intervals) for each device configuration as the number of sensors varies

6 Traffic Analysis Application: Website Fingerprinting

Having evaluated the packet source classification techniques in isolation, we now explore how they can be used to perform complex traffic analysis attacks. In

particular, we demonstrate that the ability to achieve short-term linking with relatively high accuracy provides sufficient information to perform a sophisticated website fingerprinting traffic analysis attack in which the source of an encrypted HTTP session is discovered using only packet count and size information [10,11]. While we could have chosen to demonstrate the utility of our packet clustering technique with a variety of other classes of traffic analysis attacks, website fingerprinting is a sufficiently complex problem which can be practically implemented by an attacker. In addition, through such traffic analysis, it may be possible to uniquely identify users based on their browsing habits.

In this section, we first present the traffic analysis methodology. Next, using our real RSSI data in combination with encrypted HTTP traces, we demonstrate the efficacy of a website fingerprinting attack using packets that have been classified by their source.

6.1 Traffic Analysis Methodology

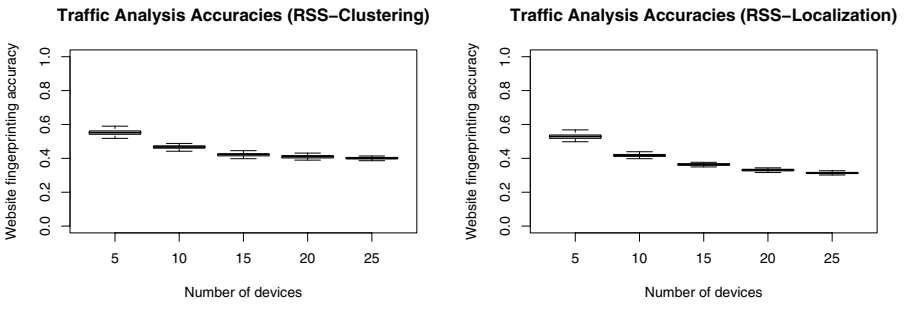
In order to apply our real RSSI data to the problem of website fingerprinting, it is necessary to combine the RSSI data with an encrypted HTTP data set. Liberatore and Levine [10] provide a data set consisting of several instances of encrypted connections to many distinct real websites over the course of several months. A website instance consists of the number of packets and their respective sizes.

To perform a simplified website fingerprinting traffic analysis attack after packet source classification, we extract multiple instances of 25 distinct websites from this data set. In general, to perform a website fingerprinting attack it is necessary to partition the website trace data into two disjoint sets, a training set, and a validation (or testing) set, and consider the task of website identification as a classification problem. We construct the website training set by collecting precisely 20 instances of each of the 25 websites that we wish to identify. The validation set is constructed by affixing an RSSI vector onto a packet that is taken from a new instance (*i.e.*, not in the training set) of one of the 25 websites. For the website classification, we apply the naïve Bayes classifier provided by `Weka` [44], as in Liberatore and Levine [10].

Similar to the experiments presented in Section 5, we construct realistic scenarios by varying the number of wireless devices from 5, 10, 15, 20, to 25 and fix the number of sensors at 5. However, instead of including an equal number of generic packets, we make the assumption that every device downloads a single randomly selected webpage and include all packets with affixed RSSI vectors from a randomly selected position.

6.2 Traffic Analysis Results

We first explore the performance of the clustering algorithm on the website data. A key distinguishing feature of the website data is that each website has an arbitrary number of packets. For some websites, the device transmits several hundred packets, while for others the device transmits less than ten packets.



(a) RSS-Clustering method (b) RSS-Localization method

Fig. 4. Website fingerprinting accuracies as the number of devices increases

Clustering devices that transmit an unequal number of packets does not appear to be a significant factor. The accuracy for the website data is only marginally lower (72–82% accuracy) than for the equal packet data (given in Figure 2).

Given the clustering algorithm's ability to accurately classify encrypted website data, we next perform a website fingerprinting attack on packets that are grouped by wireless device. The website fingerprinting accuracies for each experiment are shown in Figure 4. Using the naïve Bayes classifier, the attack is able to correctly identify the encrypted web page between 40–55% of the time. This accuracy is significantly greater than random chance, in which an adversary guesses the website. In this case, the expected accuracy is $1/25 = 4\%$. For comparison, if packets are perfectly clustered, the website fingerprinting attack achieves 92% accuracy for each device configuration. The accuracy of the website identification is strongly linked to the accuracy of the clustering result. For example, in the 5 device network, both the clustering and website identification accuracies are the highest, and each respective accuracy degrades as the number of devices increases. The website fingerprinting accuracy when the localization approach is applied is slightly worse than the clustering approach.

7 Discussion

In this section, we discuss techniques for reconstructing persistent identifiers, mitigating source classification, the benefits of large crowds for anonymity in the wireless context, and the potential for using jamming and frequency hopping to protect privacy.

7.1 Reconstructing Persistent Identifiers

The packet source classification technique as presented enables short-term linking, but cannot directly reconstruct the persistent identifiers that are necessary to enable user tracking or profiling across sessions. Once short-term linking has been accomplished, it becomes possible to perform a variety of traffic analysis

tasks to identify such information about the device including its wireless NIC driver, operating system, firewall settings, and/or more specific user behavior. This information can sometimes be used to uniquely identify devices across session, and thus could be used to reconstruct persistent identifiers. For instance, a device with an obscure OS/NIC driver combination may be easy to uniquely identify.

In addition, semantic location information can augment the packet source classification with a physical location binding. Such information could also be used to link source classified packets back to a specific source. The RSS localization-based source classification technique ostensibly provides the device's location, but it comes at the cost of collecting training data for the target environment.

7.2 Mitigating Packet Source Classification

We next explore techniques using transmit power control and directional antennas to reduce the effectiveness of packet source classification.

Intuition. For a given transmitter's location, the expected received signal strength at each sensor is predictable within some variance. However, if the transmitter's signal strength is reduced or amplified, then it becomes more likely that the received signal strengths observed at each sensor may overlap with those from other wireless devices. The result of a single transmitter varying its power levels often results in a cluster that encompasses a different portion of the signal space. In addition, directional antennas attenuate the wireless signal in certain directions while amplifying the signal in other directions, enabling the packets sent in each direction to form their own distinct clusters.[3] This phenomenon, as we will demonstrate, has an adverse effect on clustering accuracy and therefore reduces an adversary's ability to perform traffic analysis attacks on the source classified packets.

Transmit Power Control. We conduct experiments to understand the extent to which variable transmission power levels can be used to protect devices from short-term linking at the physical layer (see Appendix A for detailed hardware specifications). All other devices in the network transmit their packets at a fixed 16 dBm. Experiments are conducted with 15 total devices in which 1, 3, 6, 9, and 12 devices transmit their packets at a randomly chosen power level. As the number of devices with variable transmit power levels increases, the source classification accuracy using the clustering method varies between 61–72%.[4] The accuracy decreases by 10–15% from the results in Section 5.2. The reduction in clustering accuracy has a negative impact on the website fingerprinting traffic analysis. The traffic analysis accuracy is approximately 30%, an improvement

[3] We also conducted informal experiments in which the throughput is measured while manipulating a single transmitter's power levels. We found that the impact on throughput was insignificant. Similarly, pointing a directional antenna in different orientations also had an insignificant impact on throughput.

[4] The localization-based source classification method performed similarly.

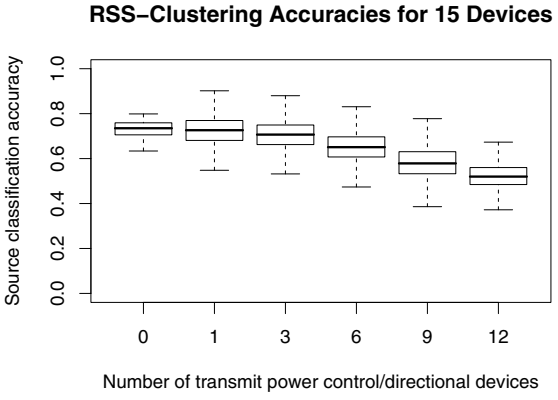


Fig. 5. For a 15 device network, the effect of introducing 0–12 devices using transmit power control in combination with directional antennas is shown

over the results from Section 6. However, devices that do not use this strategy show the same vulnerability to traffic analysis.

Directional Antennas. Low-cost directional antennas, such as sectored or MIMO antennas, are becoming widely deployed with the 802.11n standard. We next explore how directional antennas can be used to alter physical layer information, by repeating the same experiments as above except using directional antenna transmitters in place of variable transmit power level devices. The directional antenna was oriented in four different directions as the device transmitted packets. We found that the clustering accuracy decreases in a similar fashion as the experiments with the variable transmit power levels. The website fingerprinting traffic analysis attack also achieves about 30% accuracy for directional devices while non-directional devices obtain no protection from traffic analysis.

Combined Effect. The most significant reduction in source classification accuracy occurs when devices utilize transmit power control in combination with directional antennas. Figure 5 shows that the mean clustering accuracy decreases to nearly 50% as more devices use the combined strategy. The traffic analysis attack's accuracy also decreases to 26% for devices that utilize this strategy.

Hiding Signal Strength Information is Hard. The relative success of the source classification and subsequent traffic analysis despite these defensive techniques highlights the inherent difficulty of manipulating the properties of the physical layer. Ultimately, intentionally changing RSSI values is a hard problem, since there are many unobservable and environmental factors including multipath fading and attenuation that are difficult to isolate and predict. Furthermore, it is necessary to transmit at a level that is sufficient to reach an access point. Thus, these observations are consistent with prior findings that there are fundamental limitations to the extent to which the signal strength properties of the physical layer can be altered [45,46].

7.3 Anonymity *Still* Loves Company

Anonymity mechanisms for wireless networks (discussed in Section 2) such as link layer encryption achieve sender anonymity for wireless clients by effectively randomizing explicit identifiers. At the link layer and above, wireless packets are unlinkable to their senders. However, in order for this condition to hold, there is an implicit assumption that there are significantly many wireless clients in the network. For instance, if only one client uses the network, it is trivial to link their traffic to a user.

Since signal strength varies with physical distance, devices that are closer to one another typically have similar signal strengths. A group of devices within close physical proximity may be more difficult to distinguish using their signal strengths. Thus, as with traditional anonymity, a larger user base enables stronger anonymity properties than a smaller one [47]. In the wireless case, the caveat is that these users should physically arrange themselves close to each other so their signal strengths are less distinguishable to the source classification method.

7.4 Wireless Cover Traffic

Cover traffic is a well-known strategy to frustrate traffic analysis [15]. In wireless networks, cover traffic may be another tool to mitigate traffic analysis, but there are additional challenges posed by the wireless medium. First, the wireless medium is a shared resource and adding additional traffic may degrade everyone's performance. In addition, wireless devices are often battery powered and, thus try to conserve energy. Contributing cover traffic could have serious implications for power consumption and may reduce a device's lifetime. Cover traffic increases the number of packets on which an adversary could perform source classification, but the subsequent traffic analysis tasks may become more difficult. A complete study of cover traffic in the wireless context is beyond the scope of this work.

7.5 Physical Space Security, Jamming, and Frequency Hopping

Beyond hiding the contents of a communication session with cryptography, other radical approaches have been proposed that aim to reduce the number of packets that can be overhead by an eavesdropper. Lakshmanan *et al.* and Sheth *et al.* demonstrate this by using directional antennas to focus transmissions within a secure physical space that is free of eavesdroppers [48,49].

In addition, jamming has been suggested as another method to mitigate an eavesdropper's ability to overhear wireless packets [50]. An intelligent jamming strategy aimed at the locations of potential eavesdroppers can effectively raise the noise floor at their positions, which makes it difficult to distinguish between wireless signals and normal background noise on the wireless medium. While jamming may be an effective way to neutralize eavesdroppers, it may also interfere with legitimate communications and degrade the network's performance.

Another potential technique to evade eavesdroppers is to use frequency agility to transmit on different channels in a certain pattern [51]. However, the 802.11 standard limits transmissions to the 2.4 GHz and 5 GHz frequency bands, which have a limited number of channels; thus, an eavesdropper could feasibly monitor all channels simultaneously. To mitigate harmful interference among devices, most governments in developed nations regulate the allocation and usage of wireless spectrum for specific wireless devices. Consequently, spectrum is a scarce resource, which impedes the effectiveness of frequency hopping to evade eavesdroppers.

8 Conclusion

In this paper, we demonstrate that even when explicit identifiers are removed from wireless packets at the link layer, a significant amount of information remains preserved within the wireless physical layer. We provide a packet source classification technique that uses this information to achieve short-term linking. The proposed packet source classification approach is unsupervised and requires no specialized hardware.

Through experiments, we show that this approach provides sufficient accuracy to enable complex traffic analysis tasks. As an example, we conduct a website fingerprinting attack on source-classified packets with reasonably high success. To mitigate the effectiveness of the packet source classification, we evaluate methods to alter the transmitted signal strength of packets, thereby introducing additional noise which degrades the accuracy of both the packet source classification and the subsequent traffic analysis. We hope that this work will bring more awareness to the privacy problems that are present at the wireless physical layer and encourage further exploration of methods to mitigate these types of attacks.

Acknowledgments

We thank Jeffrey Pang and the anonymous reviewers for their insightful suggestions and comments, James Martin for granting access to our office building testbed, and Eric Anderson for assisting with the data collection. This research was partially funded by NSF Awards ITR-0430593 and CRI-0454404.

References

1. Pang, J., Greenstein, B., Gummadi, R., Seshan, S., Wetherall, D.: 802.11 user fingerprinting. In: MobiCom (2007)
2. Aura, T., Lindqvist, J., Roe, M., Mohammed, A.: Chattering laptops. In: Borisov, N., Goldberg, I. (eds.) PETS 2008. LNCS, vol. 5134, pp. 167–186. Springer, Heidelberg (2008)

3. Armknecht, F., Girão, J., Matos, A., Aguiar, R.L.: Who said that? Privacy at link layer. In: INFOCOM. IEEE, Los Alamitos (2007)
4. Greenstein, B., McCoy, D., Pang, J., Kohno, T., Seshan, S., Wetherall, D.: Improving wireless privacy with an identifier-free link layer protocol. In: Mobisys (2008)
5. Singelée, D., Preneel, B.: Location privacy in wireless personal area networks. In: WiSe (2006)
6. Brik, V., Banerjee, S., Gruteser, M., Oh, S.: Wireless device identification with radiometric signatures. In: MobiCom (2008)
7. Danev, B., Capkun, S.: Physical-layer identification of wireless sensor nodes. In: Technical Report ETH Zurich System Security Group D-INFK 604 (August 2008)
8. Saponas, T.S., Lester, J., Hartung, C., Agarwal, S., Kohno, T.: Devices that tell on you: Privacy trends in consumer ubiquitous computing. In: Proc. 16th USENIX Security Symposium (2007)
9. Song, D.X., Wagner, D., Tian, X.: Timing analysis of keystrokes and timing attacks on ssh. In: 10th USENIX Security Symposium (2001)
10. Liberatore, M., Levine, B.N.: Inferring the source of encrypted HTTP connections. In: CCS 2006: Proceedings of the 13th ACM conference on Computer and communications security. ACM, New York (2006)
11. Sun, Q., Simon, D.R., Wang, Y.M., Russell, W., Padmanabhan, V.N., Qiu, L.: Statistical identification of encrypted web browsing traffic. In: IEEE Symposium on Security and Privacy (2002)
12. Wright, C., Ballard, L., Monrose, F., Masson, G.: Language identification of encrypted VoIP traffic: Alejandra y roberto or Alice and Bob? In: Proceedings of the 16th USENIX Security Symposium (2007)
13. Wright, C.V., Ballard, L., Coull, S.E., Monrose, F., Masson, G.M.: Spot me if you can: Uncovering spoken phrases in encrypted VoIP conversations (2008)
14. Wright, C., Monrose, F., Masson, G.: On inferring application protocol behaviors in encrypted network traffic. Journal of Machine Learning Research (2006)
15. Chaum, D.: Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM (February 1981)
16. Goldschlag, D.M., Reed, M.G., Syverson, P.F.: Hiding routing information. In: Anderson, R. (ed.) IH 1996. LNCS, vol. 1174, pp. 137–150. Springer, Heidelberg (1996)
17. Gruteser, M., Grunwald, D.: Enhancing location privacy in wireless LAN through disposable interface identifiers: A quantitative analysis. ACM MONET 10 (2005)
18. Arkko, J., Nikander, P., Nslund, M.: Enhancing privacy with shared pseudo random sequences. In: Christianson, B., Crispo, B., Malcolm, J.A., Roe, M. (eds.) Security Protocols 2005. LNCS, vol. 4631, pp. 197–203. Springer, Heidelberg (2007)
19. Lindqvist, J., Tapio, J.M.: Protecting privacy with protocol stack virtualization. In: WPES 2008: Proceedings of the 7th ACM workshop on Privacy in the electronic society, pp. 65–74. ACM, New York (2008)
20. Kohno, T., Broido, A., Claffy, K.: Remote physical device fingerprinting. In: IEEE Symposium on Security and Privacy, pp. 211–225. IEEE Computer Society, Los Alamitos (2005)
21. Murdoch, S.J.: Hot or not: Revealing hidden services by their clock skew. In: Proceedings of CCS 2006 (October 2006)

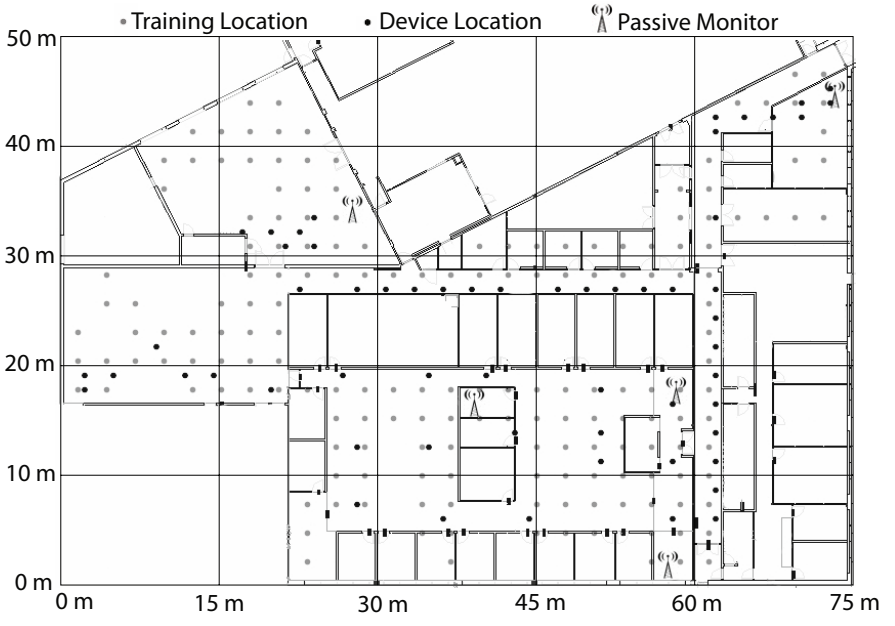
22. Zander, S., Murdoch, S.J.: An improved clock-skew measurement technique for revealing hidden services. In: Proceedings of the 17th USENIX Security Symposium, San Jose, CA, US (July 2008)
23. Gerdes, R., Daniels, T., Mina, M., Russell, S.: Device identification via analog signal fingerprinting: A matched filter approach. In: NDSS (2006)
24. Fyodor: Nmap network security scanner, `http://insecure.org/nmap`
25. p0f, `http://lcamtuf.coredump.cx/p0f.shtml`
26. Franklin, J., McCoy, D., Tabriz, P., Neagoe, V., Randwyk, J.V., Sicker, D.: Passive data link layer 802.11 wireless device driver fingerprinting. In: USENIX Security Symposium, Vancouver, Canada, July-August 2006, pp. 167–178 (2006)
27. Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit, B.: Place lab: Device positioning using radio beacons in the wild. In: Gellersen, H.-W., Want, R., Schmidt, A. (eds.) PERVASIVE 2005. LNCS, vol. 3468, pp. 116–133. Springer, Heidelberg (2005)
28. Skyhook Wireless, `http://www.skyhookwireless.com`
29. Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based user location and tracking system. In: INFOCOM (2), pp. 775–784 (2000)
30. Haeberlen, A., Flannery, E., Ladd, A.M., Rudys, A., Wallach, D.S., Kavraki, L.E.: Practical robust localization over large-scale 802.11 wireless networks. In: Proceedings of the Tenth ACM International Conference on Mobile Computing and Networking (MOBICOM), Philadelphia, PA (September 2002) (to appear)
31. Niculescu, D., Nath, B.: VOR base stations for indoor 802.11 positioning. In: MobiCom 2004: Proceedings of the 10th annual international conference on Mobile computing and networking, pp. 58–69. ACM, New York (2004)
32. Hofmann-Wellenhof, B., Lichtenegger, H., Collins, J.: Global Positioning System: Theory and Practice. Springer, Heidelberg (1997)
33. Yamasaki, R., Ogino, A., Tamaki, T., Uta, T., Matsuzawa, N., Kato, T.: TDOA location system for IEEE 802.11b WLAN. In: IEEE WCNC (2005)
34. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: MobiSys 2003: Proc. 1st international conference on Mobile systems, applications and services, pp. 31–42. ACM Press, New York (2003)
35. Jiang, T., Wang, H., Hu, Y.C.: Preserving location privacy in wireless LANs. In: MobiSys (2007)
36. Faria, D.B., Cheriton, D.R.: Detecting identity-based attacks in wireless networks using signalprints. In: WiSe 2006: Proceedings of the 5th ACM workshop on Wireless security, pp. 43–52. ACM, New York (2006)
37. Reis, C., Mahajan, R., Rodrig, M., Wetherall, D., Zahorjan, J.: Measurement-based models of delivery and interference in static wireless networks. SIGCOMM Comput. Commun. Rev. 36(4) (2006)
38. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, Heidelberg (2001)
39. Hamerly, G., Elkan, C.: Learning the k in k-means. In: Proc. 17th NIPS (2003)
40. Dan Pelleg, A.M.: X-means: Extending k-means with efficient estimation of the number of clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727–734. Morgan Kaufmann, San Francisco (2000)
41. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the gap statistic. Technical report (2000)

42. Fallah, S., Tritchler, D., Beyene, J.: Estimating number of clusters based on a general similarity matrix with application to microarray data. Statistical applications in genetics and molecular biology 7 (2008)
43. Van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Dept. of Computer Science, University of Glasgow (1979)
44. Witten, I.H., Frank, E.: Data mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
45. Shrivastava, V., Agrawal, D., Mishra, A., Banerjee, S., Nadeem, T.: Understanding the limitations of transmit power control for indoor WLANs. In: IMC 2007: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, pp. 351–364. ACM, New York (2007)
46. Blanco, M., Kokku, R., Ramachandran, K., Rangarajan, S., Sundaresan, K.: On the effectiveness of switched beam antennas in indoor environments. In: Claypool, M., Uhlig, S. (eds.) PAM 2008. LNCS, vol. 4979, pp. 122–131. Springer, Heidelberg (2008)
47. Dingledine, R., Mathewson, N.: Anonymity loves company: Usability and the network effect. In: Anderson, R. (ed.) Proceedings of the Fifth Workshop on the Economics of Information Security (WEIS 2006), Cambridge, UK (June 2006)
48. Lakshmanan, S., Tsao, C.L., Sivakumar, R., Sundaresan, K.: Securing wireless data networks against eavesdropping using smart antennas. In: ICDCS 2006: Proceedings of the 2008 The 28th International Conference on Distributed Computing Systems, Washington, DC, USA, pp. 19–27. IEEE Computer Society, Los Alamitos (2008)
49. Sheth, A., Seshan, S., Wetherall, D.: Geo-fencing: Confining Wi-Fi coverage to physical boundaries. In: Seventh International Conference on Pervasive Computing (2009)
50. Martinovic, I., Pichota, P., Schmitt, J.B.: Jamming for good: Design and analysis of a crypto-less protection for WSNs. In: Proceedings of the Second Conference on Wireless Network Security (WiSec) (March 2009)
51. Xu, W., Wood, T., Trappe, W., Zhang, Y.: Channel surfing and spatial retreats: defenses against wireless denial of service. In: WiSe 2004: Proceedings of the 3rd ACM workshop on Wireless security, pp. 80–89. ACM, New York (2004)

A Hardware Used in Experiments

Device type	Wireless NIC type	Antenna type
Sensors	D-Link DWL-AG530	Omni directional dipole antenna 2-4 dBi
Transmitters	WNC WLAN Cardbus Adaptor CB9	Omni directional dipole antenna 2-4 dBi
Directional Transmitters	WNC WLAN Cardbus Adaptor CB9	"Super Cantenna" 12 dBi 30 degree beam width directional antenna

B Building Floorplan for Experiments



Wireless devices are placed at 58 distinct physical locations in an office building. The training locations for the localization approach are also shown.