

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Designing an Android Application for the Wireless Detection of Credit Card Skimmers

Permalink

<https://escholarship.org/uc/item/1zq694cw>

Author

Bland, Maxwell Troy

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Designing an Android Application for the Wireless Detection of Credit Card Skimmers

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Maxwell Troy Bland

Committee in charge:

Kirill Levchenko, Co-Chair
Aaron Shalev, Co-Chair
Deian Stefan

2019

Copyright
Maxwell Troy Bland, 2019
All rights reserved.

The thesis of Maxwell Troy Bland is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California San Diego

2019

DEDICATION

To everyone I've worked with over the past two years and to my family. Each one of you is the epicenter of a unique perspective. I have met no one I didn't cherish, and I am lucky to know you all.

EPIGRAPH

If literature stays away from evil, it rapidly becomes boring.

— Georges Bataille

TABLE OF CONTENTS

Signature Page		iii
Dedication		iv
Epigraph		v
Table of Contents		vi
List of Figures		viii
List of Tables		x
Acknowledgements		xi
Vita		xii
Abstract of the Thesis		xiii
Chapter 1	Introduction	1
	1.1 Background on Gas Pump Skimming	2
	1.2 Methodologies of Detection	3
	1.3 Contributions	4
	1.3.1 Survey of Current Skimmer Detection Methodologies	4
	1.3.2 Design and Implementation of Bluetooth Skimmer Detection	4
	1.4 Research Scope	5
Chapter 2	Understanding Field Agent Behavior	6
	2.1 Survey Background	7
	2.2 Understanding the Documents	8
	2.2.1 Understanding Document Origins	10
	2.2.2 Performing OCR on 11,787 PDFS	11
	2.3 Document Data Extraction	13
	2.3.1 Keyword-based Analysis	14
	2.3.2 Phrase-based Analysis	14
	2.3.3 Random Sampling	15
	2.4 Describing Inspections and Skimmers	16
	2.4.1 Inspection Report Generation Bias	17
	2.4.2 Inspection Time	18
	2.4.3 Skimmer Construction	18
	2.5 The Design of Bluetana	19
	2.6 Related and Future Work	19
	2.7 Summary	20

	2.8 Acknowledgements	21
Chapter 3	A Better Mechanism: Bluetana	22
	3.1 Related Work	23
	3.2 Construction of an Application for Skimmer Detection	23
	3.2.1 Key Limitations	24
	3.2.2 The Benefits and Challenges of Crowdsourcing	24
	3.3 Bluetooth Scanning Implementation	26
	3.4 Collecting Information	26
	3.4.1 Scan Time Reduction	29
	3.4.2 Suspicious Device Flagging	29
	3.5 Future Work and Open Problems	29
	3.6 Summary	31
	3.7 Acknowledgements	31
Chapter 4	Conclusion	32
	4.1 On Crowdsourced Application Design	32
	4.2 On Effective Data Analysis from Heterogeneous Sources	33
Bibliography	34

LIST OF FIGURES

Figure 1.1:	A skimmer taped the same color as the ribbon wire it is attached to.	3
Figure 2.1:	Inconsistent use of the 198 error code. In the top report, the code is included in the inspection notes. In the bottom image, a separate field used for fail codes contains the 198.	9
Figure 2.2:	Section of the inspection reports which demonstrates information on the pre-inspection checklist. In this case, no skimmer was found.	9
Figure 2.3:	Some inspection forms include a checkbox for the exfiltration mechanism of the skimmer. While the selection of SMS may not be accurate, it does indicate not all skimmers use Bluetooth.	10
Figure 2.4:	Section of the inspection reports noting a potential MAC address for a skimmer. Follow up investigation found this MAC address to be uncorrelated with skimmer discovery.	10
Figure 2.5:	Example regular expression used to match document types. In this case, the script finds complaint triggered reports that mention skimmers.	10
Figure 2.6:	Example of a report <code>pdftotext</code> with layout after OCR. On the left hand side is the OCRed version with flawed layout recognition. In the middle is the original document. On the right is the output of <code>pdftotext</code> before OCR. . .	13
Figure 2.7:	Probability of skimmer detection approximated by the phrase-based model per phrase added. The approximation has a large spike upon addition of the phrase “no skimmer”. Otherwise, phrases added have little impact on the reported success rate.	16
Figure 2.8:	The convergence of the estimated skimmer detection probability from random sampling. The filled portion of the plot depicts a 95% confidence interval.	17
Figure 2.9:	Instance of an inspector not checking for skimmers because of security seals. In other cases, inspectors did not check locked or alarmed pumps.	17
Figure 2.10:	Inspection times for randomly sampled skimmer inspections. Some inspections are quick: for instance, the station may not have card readers. Typically inspections that find skimmers take more than an hour.	18
Figure 3.1:	Screenshot of Bluetana’s suspicious device highlighting. Highly suspicious devices are highlighted red. Less suspicious devices are highlighted orange or yellow.	27
Figure 3.2:	Screenshot comparing Bluetana’s localization data between a skimmer and non-skimmer. Skimmers, unlike other devices, have much higher signal strength near the gas pump. Sensitive data has been redacted.	28
Figure 3.3:	Screenshot of Bluetana’s web interface. The site includes a leaderboard as well as query engine for suspicious devices. Notifications of suspicious devices were pushed to team members via email.	28
Figure 3.4:	Plot of the distribution of skimmers in the last stage of flagging. After MAC, Device class, and Bluetooth classic filtering few devices remain.	30

Figure 3.5: Decision tree for determining how Bluetana highlights a device. Orange and red devices suggest the performance of localization or manual inspection. . 30

LIST OF TABLES

Table 2.1:	Overview statistics on the documents from AZDWM gas station inspection reports. Only a portion of the documents are parsable. Some non-skimmer inspections include checks for skimmers.	7
Table 2.2:	Summary of skimmer inspection form data origins and their occurrence rate. Reports are submitted most often for scheduled and skimmer-specific inspections.	11
Table 2.3:	OCR corrected inspection form origins. The ratios of each report type are approximately the same as before OCR. However, vapor recovery and complaint based inspections jumped in number.	12
Table 2.4:	The skimmer detection success rates as reported by different methods for document analysis. Both phrase based and keyword based analysis are inaccurate. However, random sampling is accurate to $\pm 2.6\%$ with 99% confidence. . .	14
Table 2.5:	Subset of the corpus of phrases used to determine whether a skimmer was found. The phrases are in lowercase and without spaces to simplify parsing. Misspellings are accurate to the original text produced by OCR.	15
Table 3.1:	Functions and constants provided by the Android API on device discovery. All the functions return immediately. However, <code>getName</code> can return <code>null</code> until the paging step of discovery is complete.	26

ACKNOWLEDGEMENTS

Thanks to Nishant Bhaskar, without whom this work would have never come to fruition. Thanks to Kirill and Aaron, for their guidance, writing, good humor, and motivation of this research. You helped me to learn more things in a few months than I learned in a lifetime before this project.

Thank you to the foundry, particularly Shelby Thomas, Vector Guo, Lixiang Ao, and Nishant Bhaskar, for your support of my endeavors, encouragement, and advice. Thank you as well to the rest of the Systems and Networking group at UCSD. Being exposed to great minds is a blessing. Thank you to Stewart Grant for looking over this thesis and psuedo-advising me on all matters. There are times when everything is hopeless and your words pull me through. We may have been passing by each-other on this journey, but I am thankful for *even* that.

Thank you to Yvonne Li for the past four years. Though our paths may now diverge, you were a counterbalance to both me and the life I led. Thank you to my family. To my father, whose strange perspective always colored my own in subtle ways. To my mother, whose love and eccentricity remind me that science is not a machine.

Thanks also to the inspectors leveraging Bluetana in their unending hunt for skimmers. You are the real heroes. Thank you for sharing your stories and being open to try the tech of a couple So-Cal punks.

Chapter 3, in part, is a reprint of the material as it appears in USENIX Security 2019. Bhaskar, Nishant; Bland, Maxwell; Levchenko, Kirill; Schulman, Aaron, In proc. USENIX Security, 2019. The dissertation/thesis author was a primary investigator and author of this paper.

VITA

- 2019 M. S. in Computer Science, University of California San Diego
- 2018 B. S. in Computer Science, University of California San Diego

PUBLICATIONS

Nishant Bhaskar, Maxwell Bland, Kirill Levchenko, and Aaron Schulman. "Please Pay Inside: Evaluating Bluetooth-based Detection of Gas Pump Skimmers." In Proceedings of the 28th USENIX Security Symposium, <https://www.usenix.org/conference/usenixsecurity19/presentation/bhaskar>. USENIX, 2019.

Hallahan, William T., Anton Xue, Maxwell Troy Bland, Ranjit Jhala, and Ruzica Piskac. "Lazy counterfactual symbolic execution." In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 411-424. ACM, 2019.

Christian Dameff, M. D., Maxwell Bland, and Jeff Tully. "Pestilential Protocol: How Unsecure HL7 Messages Threaten Patient Lives." Blackhat 2018.

ABSTRACT OF THE THESIS

Designing an Android Application for the Wireless Detection of Credit Card Skimmers

by

Maxwell Troy Bland

Master of Science in Computer Science

University of California San Diego, 2019

Kirill Levchenko, Co-Chair

Aaron Shalev, Co-Chair

Embedded devices installed inside gas pumps allow criminals to steal millions each year. These devices, called skimmers, record unencrypted information transmitted by the pumps' payment circuitry. Despite the monetary impact, no study of law enforcement skimmer recovery methodologies exists. This thesis presents the first such study. I discover that criminals often use Bluetooth to retrieve stolen card information. I also find that current investigative mechanisms are ineffective. These findings motivate the development of an Android application for skimmer detection, Bluetana. Bluetana leverages Bluetooth and crowd-sourced data collection to detect skimmers. This approach is effective: detecting 64 skimmers over two years.

Chapter 1

Introduction

“In conclusion, *Rhagoletis pomonella* may not be that unconventional after all.”
—Jefferey Feder, *The Apple Maggot Fly, Rhagoletis Pomonella*

Criminals use skimmers to steal card data during transactions at Point-of-Sale terminals. The devices are attached either externally or internally to the terminal [Kre16]. External skimmers use a second magnetic read head to steal card information. Criminals hide external skimmers inside a fake face-plate [SPT18]. Internal skimmers are Bluetooth capable and attached to wiring inside the terminal [Kre14]. In gas pumps, this wiring carries unencrypted card information [Kre18]. Skimmer construction is simple, usually consisting of a microcontroller and EEPROM. Internal skimmers oftentimes have Bluetooth module for data exfiltration. Each skimmer captures around 30 cards and makes around \$15,000 per day: the typical card is defrauded for \$500 [Rip18, The17].

There has been no survey of how investigators currently detect internal skimmers. This thesis provides this survey and a discussion of exploratory data analysis. The resulting contribution is an understanding of investigator behavior and skimmer installation. This information was used to develop an Android application, Bluetana, for skimmer detection. Bluetana detected 64 skimming devices, many of which would have otherwise remained undiscovered. The second half of this thesis discusses the construction and operation of Bluetana.

Chapter 1 of this thesis addresses provides background on skimming and skimmer detection. Chapter 2 discusses data analysis and surveys current skimmer investigative techniques. Chapter 3 describes the implementation and design of Bluetana. Chapter 4 concludes and addresses the epistemological background of this thesis.

1.1 Background on Gas Pump Skimming

Many devices can host an internal skimmer, but gas pumps are a major target. 972 skimmers were recovered from gas stations in Florida and 148 in Arizona in 2018. Each of these devices is a potential source of hundreds of dollars in fraud per day [Mem11, Cri16, App15, Sen15]. Despite the financial impact, gas station skimmers are sparse and difficult to detect. Arizona inspection reports indicate approximately 1.5% of stations, randomly sampled, have a skimmer (Section 2.3).

Part of the reason why gas pumps are targeted is that it is not difficult to access pump internals. News reports have shown criminals installing internal skimmers in less than 30 seconds [817]. Newer pumps restrict access to the internals, but are only installed in around 30% of gas stations.¹ It can be difficult to catch criminals in the act of collection as well. Skimmers built with a Bluetooth module can use a serial port to allow for remote data collection. This prevents criminals from needing to open up the pump again.

Skimmers are also hidden within the pump infrastructure to avoid physical detection. They are often covered in tape the same color as the ribbon wire they are connected to, such as in Figure 1.1. This can lead investigators to miss skimmers during routine inspections. This occurred once during the evaluation of Bluetana. In this case, Bluetana determined a skimmer was in a pump. However, *two* separate physical inspections were needed for recovery of the skimmer. The device had been hidden in a hard-to-reach crevice of the machine.

¹The result of a survey of 0.5% of the gas stations in the United States via Google Maps.



Figure 1.1: A skimmer taped the same color as the ribbon wire it is attached to.

1.2 Methodologies of Detection

Without wireless scanning, internal skimmer detection requires physical inspection of pump internals. Inspections take an average of thirty minutes and can still fail (Section 2.4.1). Inspections are also not completely random. Credit companies, law enforcement, and gas station owners can provide skimmer location hints (Section 2.2).

Applications and tools for the detection of skimmers have begun to emerge. In 2018, the Skimreaper tool was developed for detecting external skimmers [SPT18]. The tool works by detecting the secondary read head of external skimmers. Additionally, gas station owners can install alarm systems on pump access panels [Tec19]. Android applications for detecting Bluetooth-enabled skimmers also exist [S.18]. However, they have been shown to be ineffective [SBP⁺19]. These applications do not leverage the same sophisticated feature set as Bluetana (Section 3.4.2). Skimmers using other wireless technologies, such as GSM, have also been found

[Kre17b]. Discussions with law enforcement have indicated that these technologies are much less common.

1.3 Contributions

This thesis details the development of an effective skimmer detection mechanism. It demonstrates how a careful survey of current investigative techniques can provide insight. It also describes the challenges with implementing a crowd-sourced data collection platform. This platform continues to see use by government inspectors across the United States. Finally, this work describes how Bluetana led to the detection of 64 skimmers.

1.3.1 Survey of Current Skimmer Detection Methodologies

Chapter 2 surveys 28,135 gas station inspection reports from the Arizona Weights and Measures Department. Classifying and understanding these documents required a sophisticated technical pipeline (Section 2.3). The chapter also provides an analysis of the data recovered. The discoveries of this chapter motivate the development of Bluetana. The challenges and lessons learned suggest fruitful avenues of future work (Section 2.6).

1.3.2 Design and Implementation of Bluetooth Skimmer Detection

Chapter 3 details the design of Bluetana and its data collection and analysis. This reveals how the application detects skimmers using Bluetooth. The chapter also provides insight into the development of crowd-sourced applications for research. The section concludes with open problems relating to skimmer detection.

1.4 Research Scope

Skimming remains an open problem and the source of millions of dollars worth of fraud each year [Rip18]. This work gives insight on skimmer detection and prevention mechanisms. Systems like Bluetana and SkimReaper are a crucial step towards internal skimmer prevention. The next wave of research on skimming can start from these initial findings. For now, Bluetooth-based skimmer detection is an effective improvement over physical inspection alone. However, further work will be needed to solve the problem of skimming.

Chapter 2

Understanding Field Agent Behavior

“It is a capital mistake to theorize before one has data.”
— Sir Arthur Conan Doyle, *Sherlock Holmes*

This chapter surveys gas pump inspection reports from the Arizona Department of Weights and Measures (AZDWM). It begins by attempting to attain descriptive statistics related to skimming from these documents. Challenges in attaining these statistics lead into discussion of data acquisition methodologies. This chapter then highlights discoveries about inspections and skimmers contained within the documents. It concludes with a summary of the motivation for Bluetana’s development provided by the documents and future work.

Analysis of the data provided by AZDWM proved to be difficult, but had two primary benefits. The first was a categorical understanding of current investigative techniques: notes provided by investigators offer insight into how they perform inspections. The second was an understanding of the potential avenues of skimmer detection. Photos and forms included in the reports reveal details of skimmer construction. These details suggest Bluetooth modules are commonly used and that Bluetooth is a potentially effective route for detection.

Finally, This chapter functions as a *case study* of exploratory data analysis. It does not provide universal techniques; instead, it compares multiple approaches to the analysis of gas station inspection documents. This structure allows the chapter to demonstrate some ways *not* to

Table 2.1: Overview statistics on the documents from AZDWM gas station inspection reports. Only a portion of the documents are parsable. Some non-skimmer inspections include checks for skimmers.

Total Number of PDFs	28,135
Number of PDFs with Parsable Origin	22,910
Number of PDFs Originating from a Skimmer Inspection	2,957
Number of PDFs with the word “skimm”	5,109
Success Rate Reported by the AZDWM (Unhinted)	1.5%

do data analysis on this type of data set. It does so by comparing tempting, flawed data analysis techniques to ground-truth data.

2.1 Survey Background

Inspection report PDFs from the AZDWM are accessible through a queryable web interface (<https://ctutools.azda.gov/PdfOriginals/>). The data set of reports was retrieved via an html parser and batched `wget` requests. The PDFs downloaded from this source include documents unrelated to inspections. PDFs relating to inspection reports follow a naming scheme of $\{\text{BMF \#}\}-\{\text{Inspection \#}\}.(\text{pdf|PDF})$. A BMF, or Business Master File, number is a unique identifier given to each business by the AZDWM [oWM19b]. It is not clear from the filenames which reports pertain to skimming. Thus, classifying and analyzing these reports requires parsing the PDFs themselves. Luckily, reports contain a header that indicates the origin of the inspection.

There are 83,895 inspection report PDFs available on the AZDWM website. The oldest report available is from July 2010. This survey will focus on reports filed from the beginning of 2016 to the end of 2018. This amounts to a total of 28,135 reports. By restricting analysis to the past three years of data, we preserve the relevance of the study to the present day. Table 2.1 provides overview statistics on the documents.

A primary goal of this survey is to understand skimmer inspections. For instance, how

rigorously inspectors investigate pumps and the time it takes to do so (Section 2.4.1). Another goal was to quantify the hit rate of skimmer inspections (Section 2.3); this provides a measure of how common skimming is. Answers to these questions can drive improvements in detection methods (Section 2.5).

2.2 Understanding the Documents

Upon examining the PDFs, we find the information contained within is highly unstructured. For example, the documents have an inconsistent use of the 198 failure code. This code indicates a skimmer was found. Figure 2.1 depicts two different uses of this code. There also exist successful skimmer inspections with no 198 code in the document at all [Wet16]. Unfortunately, this indicates a single regular expression search for successful investigations is nontrivial. Additionally, there are latent ambiguities in the data collected. Many documents do not state whether a found skimmer was internal or external. They also do not state whether the AZDWM was responsible for the skimmer’s detection. For instance, a station owner or service tech can discover a skimmer and “call it in”. Hereafter, reports triggered by individuals outside the AZDWM are referred to as “hinted”.

The exact skimmer inspection methodology is not noted in any of the reports. There are, however, hints within the inspection notes and the document structure. For instance, vapor recovery ¹ reports contain a pre-inspection checklist. This checklist contains a check for skimmers (Figure 2.2). The reports do not contain information on the whether skimmer investigations occur alongside other inspections. However, the AZDWM site notes skimmer inspections are always performed during pump inspections [oWM19a]. Thus, any report that includes a pump inspection includes a skimmer check.

The inspection reports also contain information on skimmer construction. Some report

¹Checking the pumps to make sure they do not leak gasoline vapors into the atmosphere.

called the Department. I came out and verified a skimmer in pump
 added to the call and came out Code 198

Pump	Grade	Fail Codes
16	all	198 possible card reader type
		skimmer found waiting on PhX
		PD to identify

Figure 2.1: Inconsistent use of the 198 error code. In the top report, the code is included in the inspection notes. In the bottom image, a separate field used for fail codes contains the 198.

Pre Test Completion Checklist:		Pass/Fail
AL	<input checked="" type="checkbox"/>	Results P
PV Tested, Dry & Clear	<input checked="" type="checkbox"/>	Results P
Test Dry Brakes	<input checked="" type="checkbox"/>	Results P
gpm checked	<input checked="" type="checkbox"/>	Results P
Checked Vapor Pot	<input type="checkbox"/>	Results N/A
Checked for Skimmers	<input checked="" type="checkbox"/>	Results P
Communication	<input type="checkbox"/>	(indicate results in test section)
Liquid Blockage	<input type="checkbox"/>	
Pressure Decay	<input checked="" type="checkbox"/>	

Figure 2.2: Section of the inspection reports which demonstrates information on the pre-inspection checklist. In this case, no skimmer was found.

SKIMMER DESIGN:	<input type="radio"/> Suspected to have Bluetooth capability	<input checked="" type="radio"/> Suspected to have SMS (text) capability
OTHER COMMENTS/NOTES:		

Figure 2.3: Some inspection forms include a checkbox for the exfiltration mechanism of the skimmer. While the selection of SMS may not be accurate, it does indicate not all skimmers use Bluetooth.

bluetooth address 00:03:19:85:51:75 after removing the skimmer the address is showing

Figure 2.4: Section of the inspection reports noting a potential MAC address for a skimmer. Follow up investigation found this MAC address to be uncorrelated with skimmer discovery.

forms have a checkbox for tagging whether a skimmer is SMS or Bluetooth enabled (Figure 2.3). However, inspectors may not have a perfect sense of this. AZDWM has informed us investigators do not to remove any tape used to encase or hide the skimmer in the pump. This is done to preserve fingerprints [Wea16]. Some documents do attempt to correlate Bluetooth features to discovered skimmers. For example, MAC addresses and skimmer names (Figure 2.4). During the Bluetana study, these hints were found to be inaccurate. However, they do illustrate some investigator awareness of skimmers' Bluetooth capabilities.

2.2.1 Understanding Document Origins

Inspection reports have nine different origin headers. Table 2.2 contains statistics and descriptions of the different report types. Regular, scheduled inspections are the most common, followed by skimmer inspections. Complaint-triggered inspections are also common. Unfortunately, determining whether a complaint was about a skimmer installation is difficult. Complaints unrelated to skimmers can include skimmer checks and the word `skimm` [Bro18]. Additionally, note that the total number of documents is greater than number of parsable PDFs. This is because some documents contained multiple inspection reports concatenated.

```
grep -li "skimm" $(grep -li "origin[ :]*\s*com" *.txt)
```

Figure 2.5: Example regular expression used to match document types. In this case, the script finds complaint triggered reports that mention skimmers.

Table 2.2: Summary of skimmer inspection form data origins and their occurrence rate. Reports are submitted most often for scheduled and skimmer-specific inspections.

Origin Header	Meaning	Number of Documents
SCH	Regular Scheduled Inspection	10,673
SKI	Skimmer Inspection Report	2,288
COM	Complaint Triggered Inspection	1,693
REI	Price Posting Inspection	768
VDA	Meter Licensing Inspection	377
VAN	Vapor Recovery Inspection	352
COL	Licensing Fee Collection	168
SPC	Fuel Quality Inspection	44

A regular expression was built to extract the origin header the documents (Figure 2.5). This excluded 11,787 from Table 2.2. Most of these documents have missing or incorrect ToUnicode CMaps. Documents with embedded fonts include this mapping for the extraction of Unicode content [Inc03]. High-quality optical character recognition (OCR) has potential to correct these documents. This analysis was attempted and is detailed in Section 2.2.2. 2,743 of the documents did not contain the first, primary report page. These correspond to Pressure Decay Test forms, used when testing the pumps for leaks.

2.2.2 Preforming OCR on 11,787 PDFS

The OCR to correct the missing ToUnicode CMaps was somewhat effective. After performing this analysis, the number of unrecognized origin headers dropped to 22,910. The OCR was performed using `ocrmypdf` which operatings using the Tesseract library [Smi07]. The `-clean` and `--rotate-pages` flags were used for document cleaning and orientation detection. The final results table is included in Figure 2.3. Additionally, 5,109 reports contained the word “skimm”.

All the 28,135 PDFs underwent OCR. The analysis could have isolated the 11,787 unparsable PDFs only. However, it was unclear whether the entire set of PDFs were non-corrupted. Additionally, OCR allowed for parsing the cleanly handwritten reports. With EC2, this analysis

Table 2.3: OCR corrected inspection form origins. The ratios of each report type are approximately the same as before OCR. However, vapor recovery and complaint based inspections jumped in number.

Origin Header	Number of Documents
SCH	15,265
SKI	2,957
COM	2,521
REI	848
VDA	425
VAN	633
COL	213
SPC	47

took 59.23 hours on a 72 core machine. The cost was \$188.76.

A downside of OCR is that it has trouble inferring the layout of text on the page. Figure 2.6 compares the layout of the original report to its OCR'd duplicate. In this case, the OCR version does not maintain the document's original layout. However, it does contain the investigator's name, which was originally unparseable. Incorrect layout inference made the generation of regular expressions for page content nontrivial. To solve this problem, I developed a more sophisticated tool for text extraction. This tool, data-extract-PDF [Bla18], allows the automated extraction of PDF sections. Unfortunately, this tool was not used for the final analysis in this project. However, another project with a large number of PDFs uses a variant of data-extract-PDF for content extraction.

Open Problems in Optical Character Recognition

Additional work is needed on document content extraction. It may be possible to build an inference system which can correct the CMap issues seen in this group of PDFs. This would be a benefit to search engine document indexing and empirical research. Better document layout recognition would also benefit this research. Scene text recognition has been a longstanding open problem. The first public data set, the ICDAR Robust Reading challenge, was created in 2003

Page 2 of 3
 12/10/16 12:10:16 9886 297392

ARIZONA DEPARTMENT OF AGRICULTURE
 WELFARE AND BUSINESS SERVICES DIVISION
 Large Scale Inspection Form
 AZ-LS-1464

DATE: 12/10/16 BMT#: 9886 INSPECTION #: 297392

INSPECTION AREA	INSPECTION DATE	INSPECTION TIME	INSPECTION TYPE	INSPECTION RESULT
1. GENERAL	12/10/16	12:10:16	9886	297392
2. GENERAL	12/10/16	12:10:16	9886	297392
3. GENERAL	12/10/16	12:10:16	9886	297392
4. GENERAL	12/10/16	12:10:16	9886	297392
5. GENERAL	12/10/16	12:10:16	9886	297392
6. GENERAL	12/10/16	12:10:16	9886	297392
7. GENERAL	12/10/16	12:10:16	9886	297392
8. GENERAL	12/10/16	12:10:16	9886	297392
9. GENERAL	12/10/16	12:10:16	9886	297392
10. GENERAL	12/10/16	12:10:16	9886	297392
11. GENERAL	12/10/16	12:10:16	9886	297392
12. GENERAL	12/10/16	12:10:16	9886	297392
13. GENERAL	12/10/16	12:10:16	9886	297392
14. GENERAL	12/10/16	12:10:16	9886	297392
15. GENERAL	12/10/16	12:10:16	9886	297392
16. GENERAL	12/10/16	12:10:16	9886	297392
17. GENERAL	12/10/16	12:10:16	9886	297392
18. GENERAL	12/10/16	12:10:16	9886	297392
19. GENERAL	12/10/16	12:10:16	9886	297392
20. GENERAL	12/10/16	12:10:16	9886	297392
21. GENERAL	12/10/16	12:10:16	9886	297392
22. GENERAL	12/10/16	12:10:16	9886	297392
23. GENERAL	12/10/16	12:10:16	9886	297392
24. GENERAL	12/10/16	12:10:16	9886	297392
25. GENERAL	12/10/16	12:10:16	9886	297392
26. GENERAL	12/10/16	12:10:16	9886	297392
27. GENERAL	12/10/16	12:10:16	9886	297392
28. GENERAL	12/10/16	12:10:16	9886	297392
29. GENERAL	12/10/16	12:10:16	9886	297392
30. GENERAL	12/10/16	12:10:16	9886	297392
31. GENERAL	12/10/16	12:10:16	9886	297392
32. GENERAL	12/10/16	12:10:16	9886	297392
33. GENERAL	12/10/16	12:10:16	9886	297392
34. GENERAL	12/10/16	12:10:16	9886	297392
35. GENERAL	12/10/16	12:10:16	9886	297392
36. GENERAL	12/10/16	12:10:16	9886	297392
37. GENERAL	12/10/16	12:10:16	9886	297392
38. GENERAL	12/10/16	12:10:16	9886	297392
39. GENERAL	12/10/16	12:10:16	9886	297392
40. GENERAL	12/10/16	12:10:16	9886	297392
41. GENERAL	12/10/16	12:10:16	9886	297392
42. GENERAL	12/10/16	12:10:16	9886	297392
43. GENERAL	12/10/16	12:10:16	9886	297392
44. GENERAL	12/10/16	12:10:16	9886	297392
45. GENERAL	12/10/16	12:10:16	9886	297392
46. GENERAL	12/10/16	12:10:16	9886	297392
47. GENERAL	12/10/16	12:10:16	9886	297392
48. GENERAL	12/10/16	12:10:16	9886	297392
49. GENERAL	12/10/16	12:10:16	9886	297392
50. GENERAL	12/10/16	12:10:16	9886	297392

ARIZONA DEPARTMENT OF AGRICULTURE
 WELFARE AND BUSINESS SERVICES DIVISION
 Large Scale Inspection Form
 AZ-LS-1464

DATE: 12/10/16 BMT#: 9886 INSPECTION #: 297392

INSPECTION AREA	INSPECTION DATE	INSPECTION TIME	INSPECTION TYPE	INSPECTION RESULT
1. GENERAL	12/10/16	12:10:16	9886	297392
2. GENERAL	12/10/16	12:10:16	9886	297392
3. GENERAL	12/10/16	12:10:16	9886	297392
4. GENERAL	12/10/16	12:10:16	9886	297392
5. GENERAL	12/10/16	12:10:16	9886	297392
6. GENERAL	12/10/16	12:10:16	9886	297392
7. GENERAL	12/10/16	12:10:16	9886	297392
8. GENERAL	12/10/16	12:10:16	9886	297392
9. GENERAL	12/10/16	12:10:16	9886	297392
10. GENERAL	12/10/16	12:10:16	9886	297392
11. GENERAL	12/10/16	12:10:16	9886	297392
12. GENERAL	12/10/16	12:10:16	9886	297392
13. GENERAL	12/10/16	12:10:16	9886	297392
14. GENERAL	12/10/16	12:10:16	9886	297392
15. GENERAL	12/10/16	12:10:16	9886	297392
16. GENERAL	12/10/16	12:10:16	9886	297392
17. GENERAL	12/10/16	12:10:16	9886	297392
18. GENERAL	12/10/16	12:10:16	9886	297392
19. GENERAL	12/10/16	12:10:16	9886	297392
20. GENERAL	12/10/16	12:10:16	9886	297392
21. GENERAL	12/10/16	12:10:16	9886	297392
22. GENERAL	12/10/16	12:10:16	9886	297392
23. GENERAL	12/10/16	12:10:16	9886	297392
24. GENERAL	12/10/16	12:10:16	9886	297392
25. GENERAL	12/10/16	12:10:16	9886	297392
26. GENERAL	12/10/16	12:10:16	9886	297392
27. GENERAL	12/10/16	12:10:16	9886	297392
28. GENERAL	12/10/16	12:10:16	9886	297392
29. GENERAL	12/10/16	12:10:16	9886	297392
30. GENERAL	12/10/16	12:10:16	9886	297392
31. GENERAL	12/10/16	12:10:16	9886	297392
32. GENERAL	12/10/16	12:10:16	9886	297392
33. GENERAL	12/10/16	12:10:16	9886	297392
34. GENERAL	12/10/16	12:10:16	9886	297392
35. GENERAL	12/10/16	12:10:16	9886	297392
36. GENERAL	12/10/16	12:10:16	9886	297392
37. GENERAL	12/10/16	12:10:16	9886	297392
38. GENERAL	12/10/16	12:10:16	9886	297392
39. GENERAL	12/10/16	12:10:16	9886	297392
40. GENERAL	12/10/16	12:10:16	9886	297392
41. GENERAL	12/10/16	12:10:16	9886	297392
42. GENERAL	12/10/16	12:10:16	9886	297392
43. GENERAL	12/10/16	12:10:16	9886	297392
44. GENERAL	12/10/16	12:10:16	9886	297392
45. GENERAL	12/10/16	12:10:16	9886	297392
46. GENERAL	12/10/16	12:10:16	9886	297392
47. GENERAL	12/10/16	12:10:16	9886	297392
48. GENERAL	12/10/16	12:10:16	9886	297392
49. GENERAL	12/10/16	12:10:16	9886	297392
50. GENERAL	12/10/16	12:10:16	9886	297392

ARIZONA DEPARTMENT OF AGRICULTURE
 WELFARE AND BUSINESS SERVICES DIVISION
 Large Scale Inspection Form
 AZ-LS-1464

DATE: 12/10/16 BMT#: 9886 INSPECTION #: 297392

INSPECTION AREA	INSPECTION DATE	INSPECTION TIME	INSPECTION TYPE	INSPECTION RESULT
1. GENERAL	12/10/16	12:10:16	9886	297392
2. GENERAL	12/10/16	12:10:16	9886	297392
3. GENERAL	12/10/16	12:10:16	9886	297392
4. GENERAL	12/10/16	12:10:16	9886	297392
5. GENERAL	12/10/16	12:10:16	9886	297392
6. GENERAL	12/10/16	12:10:16	9886	297392
7. GENERAL	12/10/16	12:10:16	9886	297392
8. GENERAL	12/10/16	12:10:16	9886	297392
9. GENERAL	12/10/16	12:10:16	9886	297392
10. GENERAL	12/10/16	12:10:16	9886	297392
11. GENERAL	12/10/16	12:10:16	9886	297392
12. GENERAL	12/10/16	12:10:16	9886	297392
13. GENERAL	12/10/16	12:10:16	9886	297392
14. GENERAL	12/10/16	12:10:16	9886	297392
15. GENERAL	12/10/16	12:10:16	9886	297392
16. GENERAL	12/10/16	12:10:16	9886	297392
17. GENERAL	12/10/16	12:10:16	9886	297392
18. GENERAL	12/10/16	12:10:16	9886	297392
19. GENERAL	12/10/16	12:10:16	9886	297392
20. GENERAL	12/10/16	12:10:16	9886	297392
21. GENERAL	12/10/16	12:10:16	9886	297392
22. GENERAL	12/10/16	12:10:16	9886	297392
23. GENERAL	12/10/16	12:10:16	9886	297392
24. GENERAL	12/10/16	12:10:16	9886	297392
25. GENERAL	12/10/16	12:10:16	9886	297392
26. GENERAL	12/10/16	12:10:16	9886	297392
27. GENERAL	12/10/16	12:10:16	9886	297392
28. GENERAL	12/10/16	12:10:16	9886	297392
29. GENERAL	12/10/16	12:10:16	9886	297392
30. GENERAL	12/10/16	12:10:16	9886	297392
31. GENERAL	12/10/16	12:10:16	9886	297392
32. GENERAL	12/10/16	12:10:16	9886	297392
33. GENERAL	12/10/16	12:10:16	9886	297392
34. GENERAL	12/10/16	12:10:16	9886	297392
35. GENERAL	12/10/16	12:10:16	9886	297392
36. GENERAL	12/10/16	12:10:16	9886	297392
37. GENERAL	12/10/16	12:10:16	9886	297392
38. GENERAL	12/10/16	12:10:16	9886	297392
39. GENERAL	12/10/16	12:10:16	9886	297392
40. GENERAL	12/10/16	12:10:16	9886	297392
41. GENERAL	12/10/16	12:10:16	9886	297392
42. GENERAL	12/10/16	12:10:16	9886	297392
43. GENERAL	12/10/16	12:10:16	9886	297392
44. GENERAL	12/10/16	12:10:16	9886	297392
45. GENERAL	12/10/16	12:10:16	9886	297392
46. GENERAL	12/10/16	12:10:16	9886	297392
47. GENERAL	12/10/16	12:10:16	9886	297392
48. GENERAL	12/10/16	12:10:16	9886	297392
49. GENERAL	12/10/16	12:10:16	9886	297392
50. GENERAL	12/10/16	12:10:16	9886	297392

Figure 2.6: Example of a report pdftotext with layout after OCR. On the left hand side is the OCR'd version with flawed layout recognition. In the middle is the original document. On the right is the output of pdftotext before OCR.

[LPS+03]. However, these data sets focus on images and not on documents. Recent advances in Visual Question Answering (VQA) could be used for this purpose. Last year, Anderson presented Bottom-Up and Top-Down attention models. [AHB+18]. These models operate at the level of objects and image regions, rather than uniform grids. This technology may be adaptable to document region detection.

2.3 Document Data Extraction

Despite parsing difficulties, the documents contain useful information about skimmer investigations. Of particular interest is the number of skimmer inspections which discover skimmers. A manual analysis of all the documents would be required to answer this question with perfect precision. Thus, several alternative approximations were attempted. Each alternative method and the resulting approximation are detailed in Table 2.4. Details about these methods are included in the next three subsections. The AZDWM was also contacted directly for the ground

Table 2.4: The skimmer detection success rates as reported by different methods for document analysis. Both phrase based and keyword based analysis are inaccurate. However, random sampling is accurate to $\pm 2.6\%$ with 99% confidence.

Method	Inspections Considered	Unsuccessful Inspections
Keyword Analysis	5,109	2,689
Phrase-based Analysis	5,109	3,175
Random Sampling	256	7

truth data. Their reported “hit” rate for unhinted inspections was 1.5%.

2.3.1 Keyword-based Analysis

The first attempt at skimmer detection rate estimation was keyword based. The intuition was to subtract the reports with the words “no skimm” from the reports with the word “skimm”. This number deviates from the ground truth for trivial reasons. For one, the pre-inspection checklist includes the term “skimm”. The reported skimmer detection success rate based upon this metric was 46.6%. The actual detection rate for unhinted inspections is much lower. Using a keyword search misses semantic details of the content. For instance, the skimmer may have been discovered before AZDWM was called.

2.3.2 Phrase-based Analysis

A phrase-based approach was also evaluated. This looked for phrases to classify the document rather than keywords. Documents with these phrases were subtracted from those with the word “skimm”. 59 phrases were used in total. These more complex signifiers allow for the preservation of more semantic content. The corpus of phrases is included in Table 2.5. To build this table, the documents were randomly sampled and manually analyzed. Negative phrasing was used rather than positive phrasing. This is because positive phrases are often subsets of negative phrases, e.g. “[no] skimmer found” In this case, the percentage discovery rate reported by this technique was 37.8%. This is closer to the the ground truth reported by the AZDWM. However,

Table 2.5: Subset of the corpus of phrases used to determine whether a skimmer was found. The phrases are in lowercase and without spaces to simplify parsing. Misspellings are accurate to the original text produced by OCR.

```
pumpsinspectedforskimmers,externalandinternal;nonefound  
skimmerswithnegativeresults  
noskimmerfund  
checkedthesiteforskimmersanddidnotlocateany  
skimmers;nonefound  
skimmersbutnonewerefound  
checkforskimmers0found  
checkforaskimmingdevicewithnegativeresults  
noskimmerinspectionperformed  
noinspectionforinteriorskimmer  
skimmernonewalkedalone  
...
```

it still suffers many of the same drawbacks as the keyword based approach.

This method was effective because a small number of investigators generated the reports. The phrases captured the individual idiosyncrasies of investigators. Each inspector will, for a time, have *their way* of signifying whether they found a skimmer or not. However, phrase based analysis remains inaccurate. There are many subtle indicator phrases that an inspection was not random. Some of these, such as “owner found skimmer,” can be captured. Others are report-specific and cannot be included without manual analysis of every document. This makes it difficult to determine whether the phrase corpus is unbiased.

2.3.3 Random Sampling

The next method attempted was random sampling of documents containing the word “skimm”. This approximated a skimmer detection success rate of 2.73%. ($\pm 2.6\%$ with 99% confidence). Manual analysis of the sampled documents allowed the removal of hinted reports. Police, gas station owners, and credit companies hinted five of the seven reports. This approach also won out in the time taken to perform the analysis. Classifying this small number of reports, however, took only two hours. Finding a “reasonable” phrase corpus took more time. These

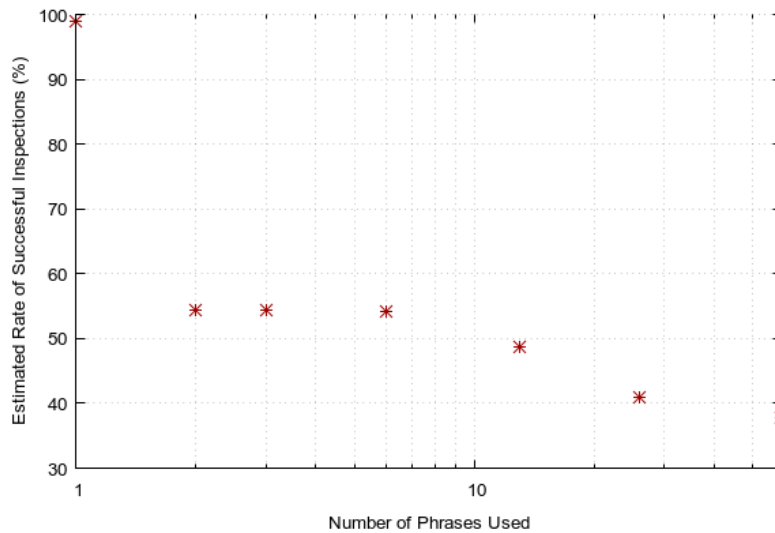


Figure 2.7: Probability of skimmer detection approximated by the phrase-based model per phrase added. The approximation has a large spike upon addition of the phrase “no skimmer”. Otherwise, phrases added have little impact on the reported success rate.

approaches show different prediction convergence behavior as time progresses. Random sampling zeroes in on a mean estimate, while phrase-based analysis moves step-wise. The latter of the two is only rigorous under restricted circumstances. Figures 2.7 and 2.8 provide this comparison.

The findings show naive automated analysis is inadequate in some empirical measurement domains. Automated approaches can be tempting in their formality. However, they can also result in severe statistical bias. Manual inspection *can* contain bias as well. For some studies, such as this one, the accuracy of random sampling suggests bias is not an issue.

2.4 Describing Inspections and Skimmers

By looking at the sample of documents, we get a broader understanding of skimming. First, complaints and hints from customers are somewhat helpful when detecting skimmers. Additionally, internal, Bluetooth enabled skimmers are common. Finally, manual skimmer inspections are slow: often taking 30 minutes or more.

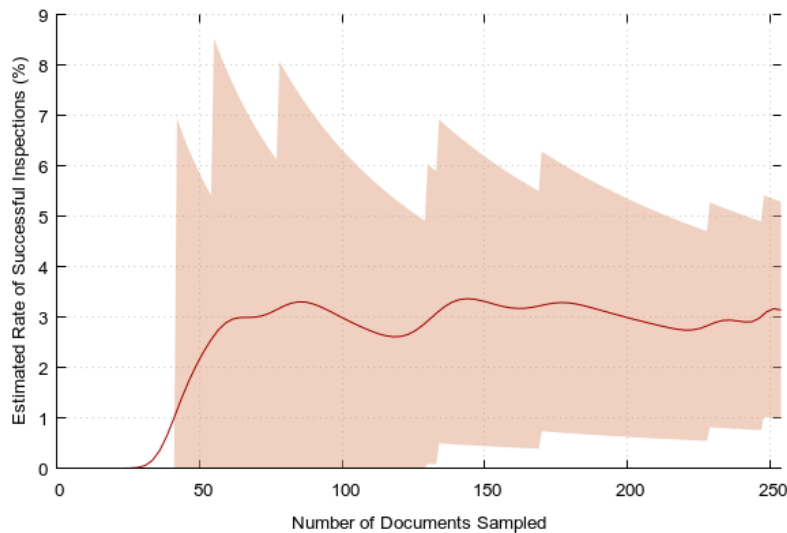


Figure 2.8: The convergence of the estimated skimmer detection probability from random sampling. The filled portion of the plot depicts a 95% confidence interval.

Site has security tamper seals installed on each pump. These seals are checked daily and immediately reported to maintenance if one is broken or appears to have been tampered with. All pumps have a site/company specific lock installed in addition to the manufacturer's lock. No evidence was found during this inspection to indicate that any pump had been tampered with.

Figure 2.9: Instance of an inspector not checking for skimmers because of security seals. In other cases, inspectors did not check locked or alarmed pumps.

2.4.1 Inspection Report Generation Bias

Random sampling was also able to determine the effectiveness of hints. I considered a hint to be anything that suggests to the inspector that there is a skimmer. This includes owners calling in AZDWM to report a found skimmer. It also includes complaints filed about skimming which trigger an inspection. Finally, AZDWM also receive hints from credit companies and local PD. 5 of the inspections in the 256 sampled documents were hinted. Five of these hints were correct, making the efficacy of hints 29.4% for this sample.

Additionally, some documents show inspectors are not following best practices. This includes checking the pumps even if there are security seals in place (Figure 2.9). Criminals can buy some pump security seals online [Tyd19]. Therefore inspectors should check pumps even if seals are in place.

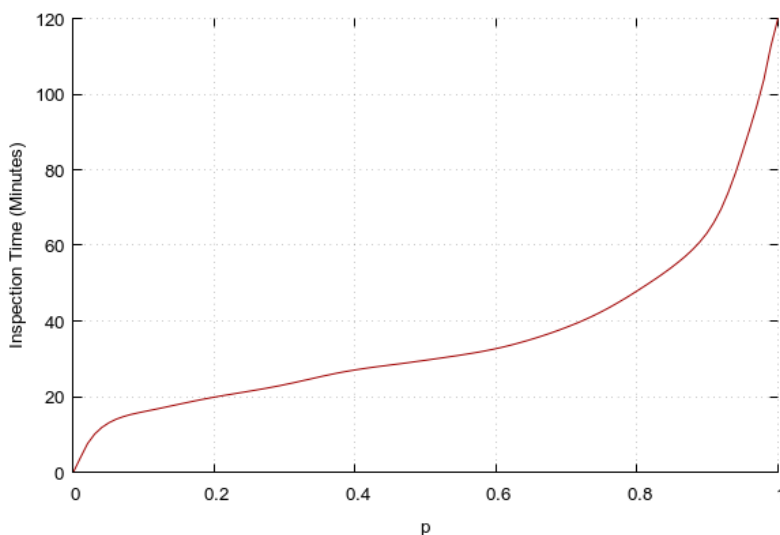


Figure 2.10: Inspection times for randomly sampled skimmer inspections. Some inspections are quick: for instance, the station may not have card readers. Typically inspections that find skimmers take more than an hour.

2.4.2 Inspection Time

Finally, skimmer inspections take a long time. I sampled 109SKI-origin reports and recorded the time taken in the inspection. Figure 2.10 presents a CDF of the inspection times. The average inspection time was 35.8 minutes ($\sigma = 21.9$ minutes). If there were a way to speed up inspection times, it could improve skimmer detection.

2.4.3 Skimmer Construction

Inspection documents can also provide hints on skimmer construction. As was seen in Section 2.2, some documents included potential MAC addresses and names. Unfortunately, most inspectors do not investigate whether a device is Bluetooth enabled. However, some are known to be [Cor17]. From this, we can isolate Bluetooth as a potential route of detection. Additionally, all the skimmers discovered within the sampled reports were internal. Thus, a robust method of internal skimmer detection may have high impact.

2.5 The Design of Bluetana

This chapter's research motivates development of a Bluetooth application for skimmer detection. In the next chapter, I will detail the design of such an application, named Bluetana. We know from the reports internal, Bluetooth-enabled skimmers are common. Because of this, Bluetooth scanning has the potential to discover these skimmers. Additionally, we know that inspections currently take a significant amount of time. Not only do inspections take time, they may not be comprehensive. Thus, a faster method of detection which is more comprehensive would be useful. Bluetana accomplishes both of these tasks. Finally, inspectors drive by and visit a large number of stations. Not every inspection is a pump inspection, as well (Section 2.2.1). A passive Bluetooth scanning application would allow for passive skimmer checking.

2.6 Related and Future Work

In this chapter, both keyword and phrase-based analysis were inaccurate in semantic classification. Random sampling was closer in determining the success rate of skimmer investigations. These same techniques were used to understand skimmer construction and investigator behavior. Since the documents are rich text by investigators, manual analysis was still needed to answer questions such as "Do inspectors check pumps with security seals?". A system for answering these queries would reduce the need for manual analysis. Similar tasks are already performed by the content-based filtering of recommendation systems. The classic content-based recommender system uses word frequency and a user profile (e.g. demographics) [Paz99]. A query could take the place of a user profile. However, this approach requires a semantic understanding of the document and the query. A 2011 survey by Lops notes several approaches for representing this semantic information. An example would be weighting documents by the frequency of rare terms [LDGS11]. However, at the time of writing, a system for performing the analyses presented by Lops in a free, quick, and automated manner does not seem to exist. In our case, Mechanical Turk

or a trained ANN may have been enough to answer most questions. However, answering each question would have required a separate round of testing or training.

Despite current efforts and the design of Bluetana, skimming will remain a problem. The GasBuddy database reports that there are 135,968 in the United States. It would be difficult for investigators to cover all stations at all times. However, preventative mechanisms are being developed and implemented. Newer Veriphone pumps operate using a chip-only reader. Authorities in other countries have recovered skimmers which operate on chip-based POS systems [Kre17a]. These devices, called shimmers, may allow criminals to continue skimming. Pump alarms and advanced locks can also prevent skimming. However, time will be needed to upgrade all the pumps in the United States.

2.7 Summary

Thus concludes my survey of AZDWM fuel station inspection reports from 2016 to 2018. From these documents, we can both better understand skimming and investigator behavior. The difficulty of investigation and the construction of skimmers suggest many potential solutions. One such solution will be discussed in the next chapter: Bluetana. This application detects skimmers by leveraging criminals' preference for Bluetooth. Contained within this chapter's survey are also useful lessons in empirical analysis. Notably, that heuristic methods of human-generated text analysis have the potential for bias. Finally, we understand that skimming is a problem that will not go away any time soon. Criminals have many directions in which they can adapt and many places to hide. Prevention will require further work on the problem.

2.8 Acknowledgements

Special thanks to Nishant for motivating this survey. Had we listened to you from the start it would have saved us time. Thank you as well to the UCSD CSE Building cleaning staff for waking me up many mornings. Thank you to my advisors, Aaron and Kirill, for pushing on this analysis. Though it didn't make it in the paper, doing this was highly instructive. Additionally, thank you for letting me take time out of other projects for this thesis.

Chapter 3

A Better Mechanism: Bluetana

“So, in the interests of survival, they trained themselves to be agreeing machines instead of thinking machines. All their minds had to do was to discover what other people were thinking, and then they thought that, too.”

— Kurt Vonnegut, *Breakfast of Champions*

In the last chapter, AZDWM documents revealed many skimming devices are Bluetooth enabled. This motivated the development of an Android application, Bluetana, for skimmer detection. Bluetana was ultimately deployed to over twenty government investigators. The application was also responsible for the recovery of over 60 skimmers. The fraud prevented is in the range of hundreds of thousands to millions of dollars (Section 1.1). In this chapter, I detail the development of Bluetana. I also detail the set of systems allowing for the application’s crowd-sourced deployment. For educational purposes, I also note some of the mistakes made during implementation.

Many features made Bluetana usable in the field by a modest number of inspectors. This includes the implementation of a *kiosk* mode. Kiosk mode made Bluetana the only accessible application on the phone. This ensured inspectors didn’t misuse the phones or accidentally close the application. Another necessity were a fast, accessible-anywhere API endpoints. This allowed us to push remote updates and get live reports from the phones. To build this server, we cleverly leveraged existing consumer infrastructure. Finally, I built caching mechanisms into the

application for collected records. This guaranteed scan records would not be lost until they were successfully uploaded.

Beyond implementation, this chapter details the front-end features added to the application. Many of these systems were critical in spurring skimmer investigations during inspections. The chapter gives a full description of the mechanisms used highlighting suspicious devices. These mechanisms are undergoing constant change, and must adapt as skimmers adapt. Thus, I conclude with a discussion of correct classification. Classification remains difficult due to the number of Bluetooth devices seen by inspectors.

3.1 Related Work

There do exist systems and tools currently on the market for skimmer detection. One example, developed by Scaife et al., is SkimReaper, a device for detecting external skimmers [SPT18]. This credit-card shaped device can detect the second read head of external skimmers. However, this device cannot detect internal skimmers. Other skimmer detection apps leveraging Bluetooth also exist on the app store. Scaife et al. reviewed these applications in a survey of skimmer protection mechanisms [SBP⁺19]. They found these applications use a limited feature set in detection. The apps match potential skimmers by MAC, name, or attempting to connect. Apps attempting to connect to the device are most concerning. This can remove forensic evidence recording the last paired MAC address.

3.2 Construction of an Application for Skimmer Detection

Most of the “skimmer detection” work is actually back-end data analysis. Thus, the front-end design of a skimmer detection application is trivial. The application approximates a run-of-the-mill Bluetooth scanner. Bluetana consists of two pages, a scan page and settings page.

The application consists of 72 code files and 4,678 lines in total. The primary page allows the user to choose to turn scanning on or off and show a list of nearby Bluetooth devices.

The settings page allows for the user to adjust the behavior of scanning and the application. Inspectors can set the app to scan only while the phone is connected to power. This makes starting a scan as easy as plugging the phone in and preserves battery. Another option, to start the app at boot-up, makes data collection more robust. By requiring less user interaction, we increase the probability that users scan. The settings page also provides versioning metadata and several override switches. These are provided in case there are bugs in the application.

3.2.1 Key Limitations

Using Android for skimmer detection has one major drawback. Bluetooth Service Discovery Protocol (SDP) look-ups require a connection to the other device. SDP features offered by a device are an effective method of fingerprinting [Won05]. As mentioned above, connecting to a skimmer can destroy useful forensic evidence. It may be possible to circumvent this restriction by sacrificing application stability. Due to stability concerns, I did not implement these workarounds.

A key limitation to Bluetana specifically was the lack of constant geolocation recording. Records were not recorded from gas stations with no Bluetooth devices. Location recording could be used to better assess Bluetooth's prevalence. For now, only stations with at least one measurable Bluetooth device appear in our analysis.

3.2.2 The Benefits and Challenges of Crowdsourcing

Detection of skimmers requires a significant amount of travel. One must visit a large number of stations to find even a single skimmer. Thus, crowd-sourcing Bluetana was necessary for the completion of this study. This required the applications adoption by the users and making the design easy to use. It also required sophisticated mechanisms for remote updates from and to

the phone.

Bluetan implements a seamless, infallible remote update mechanism. On application boot, the phone makes an API query to our server for a new application version. If one is found, the app attempts to download the new application if possible using Android's `DownloadManager` API. This allows the app to resume downloading even with a shoddy connection. The Play Store was avoided due to the sensitivity of the application. Serving remote updates in this manner causes no issue beyond the initial install. This initial install requires the user to "enable untrusted apps". The application also checks for updates before closing if a fatal exception occurs. In this case, the application can recover in case a broken update is pushed to the phones.

Another challenge is providing the API endpoints for file upload and download. Coding these routes securely and efficiently takes time (and isn't much fun). So facsimile endpoints were created by hosting files on Google Drive. This requires some care in distribution and control of API keys. However, these endpoints provided timestamps on data, dynamic content, and speedy hosting. For research projects, this is a counter-intuitive but robust infrastructural route. A separate scraper can be written to parse uploaded files into a SQL database. The Drive did start lagging on direct access from the *browser* after some time. Otherwise, no significant (non-ethical) issues were encountered using Drive in this manner.

Of Bluetana's front-end features, the most challenging to implement was a kiosk mode. This mode makes Bluetana the only application accessible on the phone. Kiosk mode prevents accidental exit from Bluetana and misuse of the phone. Making a separate APK for this mode would complicate version control. Thus, I implemented it as a dynamic feature of the application. Bluetana looks for the existence of an inaccessible-without-root file to enter this mode. It then swaps the APK's application manifest description to allow extended permissions. Complications arose in life-cycle management at startup. Running an application as the home screen causes partial garbage-collection at inopportune times. The application's initial implementation tied Bluetooth scanning to *fragments* rather than *services*. Fragments are UI layers intended to be

Table 3.1: Functions and constants provided by the Android API on device discovery. All the functions return immediately. However, `getName` can return `null` until the paging step of discovery is complete.

Field	Data	Type
<code>getAddress()</code>	Returns <code>BluetoothDevice</code> hardware address	String
<code>getBluetoothClass()</code>	Get device Bluetooth class	<code>BluetoothClass</code>
<code>getName()</code>	Get device Bluetooth name	String
<code>getType()</code>	Get Bluetooth device type	int
<code>EXTRA_RSSI</code>	Extra field in intent for signal strength	String

discarded. Thus, background tasks should almost always be implemented as services. These are lifecycle-independent data producers which the UI can consume. This led to more development time making Kiosk mode stable.

3.3 Bluetooth Scanning Implementation

The Android API's `BluetoothAdapter` class handles Bluetooth scanning. Methods of this object allow semi-direct control of the phone's Bluetooth module. Unfortunately, there are minor variations of this class for different hardware. I worked around these inconsistencies to make the application portable.

The Bluetooth adapter provides another class with Android `Intent` objects. These intents provide a `BluetoothDevice` class object whenever a Bluetooth device is discovered. Table 3.1 contains the functions and constants used for fingerprinting and localization. The class functions and their descriptions are from Android API site [Goo19]. Bluetana records this data into a CSV with 17 fields. Metadata such as location and time discovered are also recorded into this CSV.

3.4 Collecting Information

Bluetana collects data until the size of the CSV file reaches 30 kilobytes. At this point, the app compresses the file to around 4 kilobytes for upload. The CSV is then added to an upload

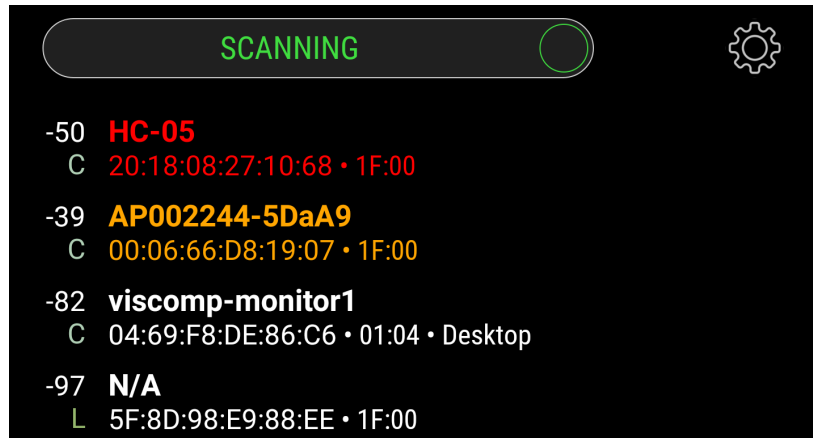


Figure 3.1: Screenshot of Bluetana’s suspicious device highlighting. Highly suspicious devices are highlighted red. Less suspicious devices are highlighted orange or yellow.

queue folder. Once a data connection is established, these files are uploaded to the drive. They are then moved to a “finished” cache on the phone. The drive is scraped every fifteen minutes for new files. If there are any, the records contained within them are input to a PSQl database.

Additionally, Bluetana contains a “hitlist” it routinely downloads from the drive. As new skimmer Bluetooth features are discovered, they are added to the hitlist. Device records which match features on the hitlist are highlighted within the app (Figure 3.1). If the device is highlighted red, inspectors are instructed to try an localize it. This increases the number of records collected for suspicious devices.

A larger number of records helped to prevent false positives. Using localization records, Bluetana was able to create heatmaps for each device (Figure 3.2). Skimmers will have a high signal strength close to a gas pump.

While other devices like car stereos may as well, this is less likely.

Live updates from each phone also allowed notifications to be created. This allowed our research team to quickly respond to new suspicious devices. It also allowed for the creation of a leaderboard and other metrics. These were integrated into a web interface used in retroactive analysis (Figure 3.3).

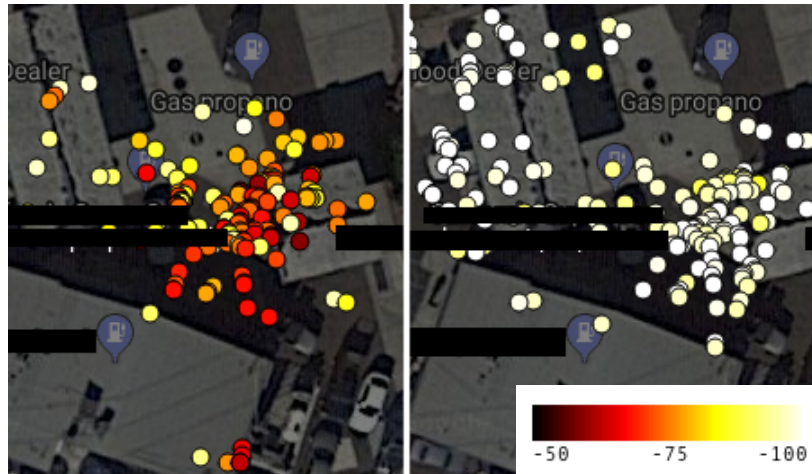


Figure 3.2: Screenshot comparing Bluetana’s localization data between a skimmer and non-skimmer. Skimmers, unlike other devices, have much higher signal strength near the gas pump. Sensitive data has been redacted.

SkimScan: An Empirical Study of Skimmers in the Wild

A project headed by Nishant Bhaskar, Maxwell Bland, Kirill Levchenko, Aaron Schulman

[Leaderboard](#) | List of who has seen the most skimmers, devices, and gas stations.

[Maps](#) | Map of all the seen devices, unclassified devices, and devices near gas stations.

[Metrics](#) | Metrics on what types of devices have been seen, database digger.

Figure 3.3: Screenshot of Bluetana’s web interface. The site includes a leaderboard as well as query engine for suspicious devices. Notifications of suspicious devices were pushed to team members via email.

3.4.1 Scan Time Reduction

The application also has settings to reduce the Bluetooth scan time. This helps to speed up device localization and records collected. Prior research demonstrates a majority of devices are detected within 10 seconds of a Bluetooth scan [PBK06]. However, this has two trade-offs. The first being that the application may not scan both Bluetooth frequency trains. These trains are used by devices during discovery to minimize signal conflicts. The second is that it prevents the device from performing some of the paging discovery stage. This stage is used for transferring the device name to the scanning device [MB01]. Due to these complications, inspectors phones were set to the android default scan time at startup.

3.4.2 Suspicious Device Flagging

We designed Bluetana's method of highlighting suspicious devices based upon discovered skimmers. Initially the hit-list contained information from the internet and existing applications. It then became clear inspectors were not finding Bluetooth Low Energy skimmers. All the discovered skimmers also had an unset Bluetooth device class ¹ and uncustomized MAC addresses. These features alone were effective in finding skimmers (Figure 3.4). However, we also performed edit-distance clustering on the device names. This revealed known products and common default names. Figure 3.5 presents a flowchart of the current flagging workflow.

3.5 Future Work and Open Problems

Bluetana managed to detect 64 skimmers, but the problem is not solved. Detection mechanisms could be extended to catch SMS-enabled skimmers. Figure 2.3 demonstrates that these skimmers exist. Additionally, persistent devices could be deployed for prevention. For example, it may be possible RF bands for payment data transfers. It may also be possible to

¹A set of bytes used to indicate what a device is (phone, stereo, etc.).

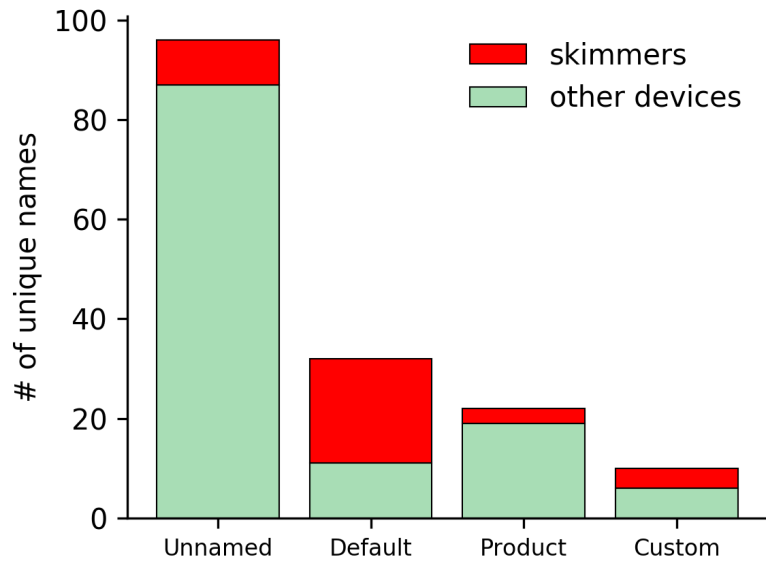


Figure 3.4: Plot of the distribution of skimmers in the last stage of flagging. After MAC, Device class, and Bluetooth classic filtering few devices remain.

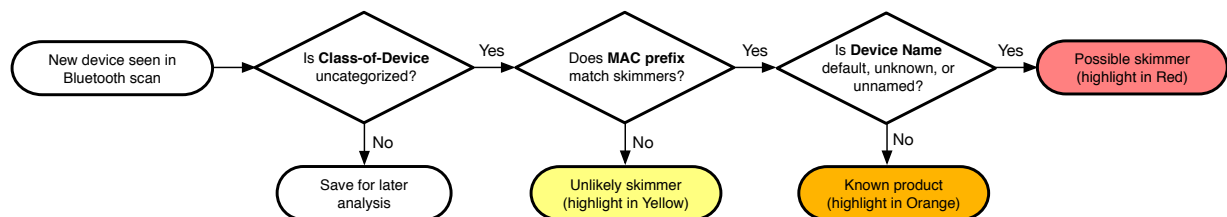


Figure 3.5: Decision tree for determining how Bluatana highlights a device. Orange and red devices suggest the performance of localization or manual inspection.

detect the power draw from internal skimmers. Criminals may also find ways of adapting their Bluetooth signature. In this case, more sophisticated localization mechanisms may be needed.

The analysis routines of Bluetana itself also have room for improvement. We relied on inspector's discovery of ground-truth data for our analysis. However, there may be skimmers they missed, exhibiting uncommon features. Notably, we did not analyze Bluetooth Low Energy modules in depth. Additionally, SDP might be used to detect more well-hidden skimmers. This would require developing workarounds for Android or low-cost embedded skimmer detection devices.

3.6 Summary

This chapter described the implementation of an Android application for skimmer detection. This application, Bluetana, although simple, was able to discover 64 skimming devices. Beyond Bluetana, this chapter also gives some insight into crowd-sourced application design. It notes the challenges in deploying an application to many users without the Play store. Finally, it was important that the application was adaptable and robust. This allowed continuous improvement after field deployment. In many ways, this was what allowed for Bluetana's success in skimmer detection.

3.7 Acknowledgements

Thank you to Aaron and Kirill for their motivation and guidance during development. Thank you to Nishant for being a friend and the other half in deployment and data analysis. Thank you to the thousands of unnamed engineers and scientists that made this project possible. And finally, thank you to the lovely staff of the AZDWM for your use of our application.

Chapter 3, in part, is a reprint of the material as it appears in USENIX Security 2019.

Bhaskar, Nishant; Bland, Maxwell; Levchenko, Kirill; Schulman, Aaron, In proc. USENIX Security, 2019. The dissertation/thesis author was a primary investigator and author of this paper.

Chapter 4

Conclusion

“And as the ends and ultimates of all things accord in some mean and measure with their inceptions and originals, that same multiplicit concordance which leads forth growth from birth accomplishing by a retrogressive metamorphosis that minishing and ablation towards the final which is agreeable unto nature so is it with our subsolar being.”

— James Joyce, *Ulysses*

This thesis presented two case studies. In the first, we uncovered the nature of skimming through the lens of AZDWM inspectors. We also discussed potential pitfalls for the fledgling empiricist. The second case study was the design of a crowd-sourced skimmer detection application. In this chapter, we addressed details of the Android API as well as system design. Many of the features which made Bluetana practical in the field were not noted in the original paper. The next two sections address the broader epistemological ground of this work.

4.1 On Crowdsourced Application Design

An individual observer cannot know everything: reality has a high degree of complexity. Thus, collaboration, the implementation of crowd-sourced applications, is necessary for knowledge. Crowd-sourcing can solve problems that need parallelism across space or time. It is also a rich ground for that which is non-computational. These systems must cooperate with

nontrivial-to-classify human input (chapter 2). This occurs through either the data they analyze or the system itself. Thus, the design of a crowd sourced systems can lead to an understanding of human behavior.

4.2 On Effective Data Analysis from Heterogeneous Sources

Information contained within this reality is non-homogeneous and avoids rigorous classification on account of the complexities generated via the combinatoric pairing of atomic deterministic systems that may be symbolically paralleled in logic. A large part of contemporary research is an attempt at a general solution to this “problem”. The closest we have come is the approximation of the human brain. However, approximation is not the only option. It may be possible to build automated systems for doing the analysis presented in this paper. These systems’s implementation may also be a crucial epistemological step in humanity’s development. They would certainly save a lot of time.

Bibliography

- [817] WFLA News Channel 8. Crooks caught on camera installing credit card skimmer in longboat key. <https://www.youtube.com/watch?v=eH4cZLVn0mA>, December 2017.
- [AHB⁺18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.
- [App15] Appeal from the US District Court for the Eastern District of Oklahoma, USA v. Konstantinov et al, 6:13cr62. United States Court of Appeals for the Tenth Circuit. <https://www.ca10.uscourts.gov/opinions/14/14-7050.pdf>, June 2015.
- [Bla18] Maxwell Bland. data-extract-pdf. <https://github.com/maxwell-bland/data-extract-pdf>, May 2018.
- [Bro18] Michael Brooks. Inspection #290206. <https://ctutools.azda.gov/PdfOriginals/7690-290206.PDF>, March 2018.
- [Cor17] Ephram Cordova. Inspection #284675. <https://ctutools.azda.gov/PdfOriginals/20340-269297.PDF>, August 2017.
- [Cri16] Criminal Complaint, USA v Cristea et al, 4:16cr182. US District Court for the Southern District of Texas. <https://www.courtlistener.com/recap/gov.uscourts.txsd.1357299.1.0.pdf>, April 2016.
- [Goo19] Inc. Google. Bluetoothdevice. <https://developer.android.com/reference/android/bluetooth/BluetoothDevice>, 2019.
- [Inc03] Adobe Systems Incorporated. Tounicode mapping file tutorial. <https://www.adobe.com/content/dam/acom/en/devnet/acrobat/pdfs/5411.ToUnicode.pdf>, May 2003.
- [Kre14] Brian Krebs. Gang rigged pumps with bluetooth skimmers. <https://krebsonsecurity.com/2014/01/gang-rigged-pumps-with-bluetooth-skimmers/>, January 2014.
- [Kre16] Brian Krebs. All about skimmers. <https://krebsonsecurity.com/all-about-skimmers/>, 2016.

- [Kre17a] Brian Krebs. Atm shimmers target chip based cards. <https://krebsonsecurity.com/2017/01/atm-shimmers-target-chip-based-cards/>, January 2017.
- [Kre17b] Brian Krebs. Gas pump skimmer sends card data via text. <https://krebsonsecurity.com/2017/07/gas-pump-skimmer-sends-card-data-via-text/>, July 2017.
- [Kre18] Brian Krebs. How to avoid card skimmers at the pump. <https://krebsonsecurity.com/2018/06/how-to-avoid-card-skimmers-at-the-pump/>, June 2018.
- [LDGS11] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.
- [LPS⁺03] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, and Robert Young. Icdar 2003 robust reading competitions. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 682–687. Citeseer, 2003.
- [MB01] Brent A Miller and Chatschik Bisdikian. *Bluetooth revealed: the insider’s guide to an open specification for global wireless communication*. Prentice Hall PTR, 2001.
- [Mem11] Memorandum and Order, USA v. Hristov et al, 1:10cr10056. US District Court for the District of Massachussetts. <https://www.courtlistener.com/recap/gov.uscourts.mad.127405/gov.uscourts.mad.127405.62.0.pdf>, April 2011.
- [oWM19a] Arizona Department of Weights and Measures. Credit card skimmers. <https://agriculture.az.gov/weights-measures/fueling/credit-card-skimmers>, 2019.
- [oWM19b] Arizona Department of Weights and Measures. Licensing. <https://agriculture.az.gov/weights-measures/licensing>, 2019.
- [Paz99] Michael J Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13(5-6):393–408, 1999.
- [PBK06] Brian S Peterson, Rusty O Baldwin, and Jeffrey P Kharoufeh. Bluetooth inquiry time characterization and selection. *IEEE Transactions on mobile computing*, 5(9):1173–1187, 2006.
- [Rip18] Rippleshot. State of card fraud: 2018. <https://www.aba.com/Products/Endorsed/Documents/Rippleshot-State-of-Card-Fraud.pdf>, 2018.
- [S.18] Ronnie S. Skim plus (bluetooth skimmer detection). <https://play.google.com/store/apps/details?id=com.rs.skimplus>, December 2018.
- [SBP⁺19] N. Scaife, J. Bowers, C. Peeters, G. Hernandez, I. N. Sherman, P. Traynor, and L. Anthony. Kiss from a rogue: Evaluating detectability of pay-at-the-pump card skimmers. In *2019 2019 IEEE Symposium on Security and Privacy (SP)*, pages 1208–1222, Los Alamitos, CA, USA, may 2019. IEEE Computer Society.

- [Sen15] Sentencing Memorandum of the United States, USA v. Aqel, 2:14cr270. US District Court for the Southern District of Ohio. <https://www.courtlistener.com/recap/gov.uscourts.ohsd.178108/gov.uscourts.ohsd.178108.47.0.pdf>, November 2015.
- [Smi07] Ray Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.
- [SPT18] Nolen Scaife, Christian Peeters, and Patrick Traynor. Fear the reaper: characterization and fast detection of card skimmers. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1–14, 2018.
- [Tec19] FlintLoc Technologies. Flintloc technologies: Foolproof skimming protection that pays for itself in months. <http://flintloc.com/>, 2019.
- [The17] The Newnan Times-Herald. Armenian skimmer leader pleads guilty. <http://times-herald.com/news/2015/06/armenian-skimmer-leader-pleads-guilty>, July 2017.
- [Tyd19] TydenBrooks. We care gas pump tamper-evident security labels. <https://tydenbrooks.com/gas-pump-tamper-evident-labels>, 2019.
- [Wea16] A.M. Weaver. Inspection #269297. <https://ctutools.azda.gov/PdfOriginals/20340-269297.PDF>, January 2016.
- [Wet16] Linda Wetzel. Inspection #268466. <https://ctutools.azda.gov/PdfOriginals/13529-268466.PDF>, June 2016.
- [Won05] Lih Wern Wong. Potential bluetooth vulnerabilities in smartphones. In *AIMS*, pages 123–132. Citeseer, 2005.