The Dark Net: De-Anonymization, Classification and Analysis

Rebecca Portnoff

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-5
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-5.html

March 7, 2018

Copyright © 2018, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

The Dark Net: De-Anonymization, Classification and Analysis

by

Rebecca Sorla Portnoff
B.S. Princeton University

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Wagner, Chair
Professor Vern Paxson
Professor David Bamman

Spring 2018

The dissertation of Rebecca Sorla Portnoff, titled The Dark Net: De-Anonymization, Classification and Analysis, is approved:

Chair Date

Date

Date

University of California, Berkeley

The Dark Net: De-Anonymization, Classification and Analysis

Copyright © 2018
by
Rebecca Sorla Portnoff

Abstract

The Dark Net: De-Anonymization, Classification and Analysis

by

Rebecca Sorla Portnoff
Doctor of Philosophy in Computer Science

University of California, Berkeley
Professor David Wagner, Chair

The Internet facilitates interactions among human beings all over the world, with greater scope and ease than we could have ever imagined. However, it does this for both well-intentioned and malicious actors alike. This dissertation focuses on these malicious persons and the spaces online that they inhabit and use for profit and pleasure. Specifically, we focus on three main domains of criminal activity on the clear web and the Dark Net: classified ads advertising trafficked humans for sexual services, cyber black-market forums, and Tor onion sites hosting forums dedicated to child sexual abuse material (CSAM).

In the first domain, we develop tools and techniques that can be used separately and in conjunction to group Backpage sex ads by their true author (and not the claimed author in the ad). Sites for online classified ads selling sex are widely used by human traffickers to support their pernicious business. The sheer quantity of ads makes manual exploration and analysis unscalable. In addition, discerning whether an ad is advertising a trafficked victim or an independent sex worker is a very difficult task. Very little concrete ground truth (i.e., ads definitively known to be posted by a trafficker) exists in this space. In the first chapter of this dissertation, we develop a machine learning classifier that uses stylometry to distinguish between ads posted by the same vs. different authors with 90% TPR and 1% FPR. We also design a linking technique that takes advantage of leakages from the Bitcoin mempool, blockchain and sex ad site, to link a subset of sex ads to Bitcoin public wallets and transactions. Finally, we demonstrate via a 4-week proof of concept using Backpage as the sex ad site, how an analyst can use these automated approaches to potentially find human traffickers.

In the second domain, we develop machine learning tools to classify and extract information from cyber black-market forums. Underground forums are widely used by criminals to buy and sell a host of stolen items, datasets, resources, and criminal services. These forums contain important resources for understanding cybercrime. However, the number of forums, their size, and the domain expertise required to understand the markets makes manual exploration of these forums unscalable. In the second

chapter of this dissertation, we propose an automated, top-down approach for analyzing underground forums. Our approach uses natural language processing and machine learning to automatically generate high-level information about underground forums, first identifying posts related to transactions, and then extracting products and prices. We also demonstrate, via a pair of case studies, how an analyst can use these automated approaches to investigate other categories of products and transactions. We use eight distinct forums to assess our tools: Antichat, Blackhat World, Carders, Darkode, Hack Forums, Hell, L33tCrew and Nulled. Our automated approach is fast and accurate, achieving over 80% accuracy in detecting post category, product, and prices.

In the third domain, we develop a set of features for a principal component analysis (PCA) based anomaly detection system to extract producers (those actively abusing children) from the full set of users on Tor CSAM forums. These forums are visited by tens of thousands of pedophiles daily. The sheer quantity of users and posts make manual exploration and analysis unscalable. In the final chapter of this dissertation, we demonstrate how to extract producers from unlabeled, public forum data. We use four distinct forums to assess our tools; these forums remain unnamed to protect law enforcement investigative efforts.

We have released our code written for the first two domains, as well as the proof of concept data from the first domain, and a sub-set of the labeled data from the second domain, allowing replication of our results.[1]

Professor David Wagner
Dissertation Committee Chair

[1]As of the filing of this dissertation, our code and data are available online at https://github.com/rsportnoff/anti-trafficking-tools and https://evidencebasedsecurity.org/forums/

To my daughter, Esther Eunmee, and all the other kids.

Contents

Contents ii

List of Figures iv

List of Tables v

1 Introduction 1
- 1.1 Classified Ads Selling Sex . . . 1
- 1.2 Cyber Black-Market Forums . . . 3
- 1.3 Child Sexual Abuse Material Forums . . . 4

2 Classified Ads Selling Sex 7
- 2.1 Introduction . . . 7
- 2.2 Related Work . . . 8
  - 2.2.1 Sex Trafficking Online . . . 8
  - 2.2.2 Bitcoin . . . 10
- 2.3 Datasets . . . 11
  - 2.3.1 Backpage . . . 11
  - 2.3.2 Bitcoin . . . 12
- 2.4 Author Classifier . . . 15
  - 2.4.1 Labeling Ground Truth . . . 15
  - 2.4.2 Models . . . 15
  - 2.4.3 Validation Results . . . 16
- 2.5 Linking Ads to Bitcoin Transactions . . . 18
- 2.6 Grouping Ads by Owner . . . 20
  - 2.6.1 Grouping by Shared Author . . . 21
  - 2.6.2 Grouping by Persistent Bitcoin Identities . . . 22
- 2.7 Case Study . . . 23
  - 2.7.1 Price Reconstruction . . . 23
  - 2.7.2 Linking Backpage Ads to Bitcoin Transactions . . . 24

2.7.3 Results . . . 26
2.8 Discussion . . . 33
2.9 Conclusion . . . 35

3 Cyber Black-Market Forums **36**
3.1 Introduction . . . 36
3.2 Related Work . . . 37
3.2.1 Ecosystem Analysis . . . 37
3.2.2 Classification for Black-Markets . . . 38
3.2.3 NLP Tools . . . 39
3.3 Forum Datasets . . . 40
3.4 Automated Processing . . . 41
3.4.1 Type-of-Post Classification . . . 42
3.4.2 Product Extraction . . . 45
3.4.3 Price Extraction . . . 50
3.4.4 Currency Exchange Extraction . . . 52
3.5 Analysis . . . 54
3.5.1 End-to-end error analysis . . . 54
3.5.2 Broadly Characterizing a Forum . . . 54
3.5.3 Performance . . . 54
3.6 Case Studies . . . 55
3.6.1 Identifying Account Activity . . . 56
3.6.2 Currency Exchange Patterns . . . 57
3.7 Conclusion . . . 57

4 CSAM Forums **60**
4.1 Introduction . . . 60
4.2 Related Work . . . 61
4.2.1 Anomaly Detection in Social Network Users . . . 61
4.2.2 CSAM . . . 62
4.3 Forum Datasets . . . 66
4.4 Extracting Producers . . . 67
4.4.1 Labeling Ground Truth . . . 68
4.4.2 Features . . . 68
4.4.3 Public Data . . . 68
4.4.4 Validation Results . . . 69
4.5 Conclusion . . . 75

5 Conclusion **76**

A **87**

List of Figures

2.1 Example of how a new transaction is added to Bitcoin's peer-to-peer network. 14
2.2 Example of a false positive case with possibly flawed ground truth. . . 18
2.3 Example of a false negative case where the ads appear in different sections. 19
2.4 Linking Ads to Bitcoin Transactions 19
2.5 Shared Authors 20
2.6 Persistent Bitcoin Identities 22

3.1 Example post and annotations from Darkode, with one sentence per line. We underline annotated product tokens. The second exhibits our annotations of both the core product (*mod DCIBot*) and the method for obtaining that product (*sombody*). 46
3.2 Number of transactions of each type observed in each forum. Numbers indicate how many posts had each type of transaction. If multiple currencies appear for either side of the transaction, then we apportion a fraction to each option. Colors indicate... 58

4.1 Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 1 70
4.2 Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 2 71
4.3 Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 3 72
4.4 Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 4 73
4.5 Number of Producers in top % of Ranked Users, with varying Principal Components: All Forums 74

List of Tables

2.1 Backpage 12
2.2 Classification accuracy for same vs. different author. 17
2.3 Distribution of Ads by Category during 4-week Case Study 22
2.4 Price calculation correctness. 24
2.5 Transaction to ad linking correctness (Exact Match - EM, Multiple Match - MM). 26
2.6 Multiple match transaction to No. ads matched. 27
2.7 Number of Pairs of Authors Linked within SA Wallet using location and Jaccard 27
2.8 SA Wallets with Zero FP by Location and Zero FP by Jaccard threshold 0.9 28
2.9 SA Wallets with between 0%-25% of pairs of authors linked by Jaccard threshold 0.9 29
2.10 Number of Transactions Linked to Exact Match within PBI single exact match Wallet using author, area code and location 31
2.11 Number of Pairs of Authors Linked within MEM Wallet using author, area code and location for Exact Matches 32
2.12 Final SA Cluster Statistics by Jaccard > 0.9 33
2.13 Final PBI Statistics 33

3.1 General properties of the forums considered. 40
3.2 Number of labeled posts by class for type-of-post classification. 44
3.3 Classification accuracy on the buy vs. sell task. MFL refers to a baseline that simply returns the most frequent label in the training set. Note that the test sets for within-forum evaluation... 44
3.4 Classification accuracy on the buy vs. sell vs. currency exchange vs. other task. Omitted entries indicate categories with too little ground-truth data of that type to robustly evaluate. 44

3.5 Results of the product extractor (trained on all training data) on the test sets of two forums. We report results for two baselines as well as for two variants of our system. Bolded results represent statistically significant improvements over... 48
3.6 Cross-forum evaluation of the post-level product extractor. We report product type F-measure on the test sets for three variants of the post-level system: one trained on Darkode, one trained on... 50
3.7 Evaluation of the **Regex**- and **SVM**-based price extractors. 51
3.8 Validation results for the currency exchange extractor. We report results for all four models, evaluating extraction of all fields (left), and for only the currencies being exchanged (right). We assess... 53
3.9 Top frequently-occurring stemmed nouns in Darkode and Hack Forums from two methods: simple frequency counts, and... 55
3.10 Accounts case study. We have a three-class classification task of posts: they deal in original accounts, in bulk/hacked accounts, or not in accounts at all. Compared to a grep baseline, a method based on our... . 57

4.1 General properties of the forums considered. 66

A.1 SA Wallet `1A3Bj` Statistics 88
A.2 SA Wallet `1Abgk` Statistics 88
A.3 SA Wallet `1ASPo` Statistics 88
A.4 SA Wallet `1BT6w` Statistics 88
A.5 SA Wallet `1D1di` Statistics 88
A.6 SA Wallet `1E4RK` Statistics 89
A.7 SA Wallet `1Gsuis` Statistics 89
A.8 SA Wallet `1Hre7` Statistics 89
A.9 SA Wallet `1Kh3x` Statistics 89
A.10 SA Wallet `1KpyX` Statistics 89
A.11 SA Wallet `1KtCW` Statistics 89
A.12 SA Wallet `1Kyoc` Statistics 89
A.13 SA Wallet `1LetZ` Statistics 90
A.14 SA Wallet `1LYEQ` Statistics 90
A.15 SA Wallet `1MGDy` Statistics 90
A.16 SA Wallet `1MheR` Statistics 90
A.17 SA Wallet `1Mufv` Statistics 90
A.18 SA Wallet `1N7V4` Statistics 90
A.19 SA Wallet `1P7n4` Statistics 90
A.20 SA Wallet `1PesE` Statistics 91
A.21 SA Wallet `1yVFE` Statistics 91

A.22 SA Wallet 12Xis7 Statistics 91
A.23 SA Wallet 14FCt Statistics 91
A.24 SA Wallet 14XUU Statistics 91
A.25 SA Wallet 16iD4 Statistics 91
A.26 SA Wallet 17idT Statistics 91
A.27 SA Wallet 18tTg Statistics 92
A.28 SA Wallet 194iD Statistics 92
A.29 SA Wallet 198xk Statistics 92
A.30 PBI Wallet 1 EM 1M58i Statistics 92
A.31 PBI Wallet 1 EM 1EyHa Statistics 92
A.32 PBI Wallet 1 EM 1AFN6 Statistics 92
A.33 PBI Wallet 1 EM 1MZG3 Statistics 92
A.34 PBI Wallet 1 EM 14eoU Statistics 92
A.35 PBI Wallet 1 EM 16THW Statistics 92
A.36 PBI Wallet 1 EM 14nrA Statistics 92
A.37 PBI Wallet 1 EM 1Nzho Statistics 93
A.38 PBI Wallet 1 EM 1Ks4n Statistics 93
A.39 PBI Wallet 1 EM 1EKEg Statistics 93
A.40 PBI Wallet 1 EM 1JntX Statistics 93
A.41 PBI Wallet 1 EM 1EChh Statistics 93
A.42 PBI Wallet 1 EM 1Aah7m Statistics 93
A.43 PBI Wallet 1 EM 18SzY Statistics 93
A.44 PBI Wallet 1 EM 18SDy Statistics 93
A.45 PBI Wallet 1 EM 1DLkr Statistics 94
A.46 PBI Wallet 1 EM 1FhgN Statistics 94
A.47 PBI Wallet 1 EM 1AnJA Statistics 94
A.48 PBI Wallet 1 EM 1LGmB Statistics 94
A.49 PBI Wallet 1 EM 14LMur Statistics 94
A.50 PBI Wallet 1 EM 1NBrV Statistics 94
A.51 PBI Wallet 1 EM 1P6eT Statistics 94
A.52 PBI Wallet 1 EM 1BZ91 Statistics 94
A.53 PBI Wallet 1 EM 1DFvc Statistics 94
A.54 PBI Wallet 1 EM 1MAoG Statistics 94
A.55 PBI Wallet 1 EM 12r6t Statistics 95
A.56 PBI Wallet 1 EM 158DE Statistics 95
A.57 PBI Wallet 1 EM 18j9y Statistics 95
A.58 PBI Wallet 1 EM 1K7rX Statistics 95
A.59 PBI Wallet 1 EM 13Ges Statistics 95
A.60 PBI Wallet 1 EM 1KCJ5 Statistics 95
A.61 PBI Wallet 1 EM 1GSWt Statistics 95

A.62 PBI Wallet 1 EM 1DaRLu Statistics 95
A.63 PBI Wallet 1 EM 16ZvA Statistics 95
A.64 PBI Wallet 1 EM 1Nuf8 Statistics 96
A.65 PBI Wallet 1 EM 16Yyx Statistics 96
A.66 PBI Wallet 1 EM 1NbRn Statistics 96
A.67 PBI Wallet 1 EM 1Lune Statistics 96
A.68 PBI Wallet 1 EM 1J1Cc Statistics 96
A.69 PBI Wallet 1 EM 1N7Af Statistics 96
A.70 PBI Wallet 1 EM 13gqd Statistics 96
A.71 PBI Wallet 1 EM 19cYw Statistics 96
A.72 PBI Wallet 1 EM 16qR6 Statistics 97
A.73 PBI Wallet 1 EM 189Bu Statistics 97
A.74 PBI Wallet 1 EM 1EKp8 Statistics 97
A.75 PBI Wallet 1 EM 13qSM Statistics 97
A.76 PBI Wallet 1 EM 1NVVV Statistics 97
A.77 PBI Wallet 1 EM 1L8rv Statistics 97
A.78 PBI Wallet 1 EM 14cax Statistics 97
A.79 PBI Wallet 1 EM 15ZAE Statistics 97
A.80 PBI Wallet 1 EM 1Fvev Statistics 97
A.81 PBI Wallet 1 EM 1EUAF Statistics 97
A.82 PBI Wallet 1 EM 1LDTv Statistics 98
A.83 PBI Wallet 1 EM 1Hu8a Statistics 98
A.84 PBI Wallet 1 EM 1EF3p Statistics 98
A.85 PBI Wallet 1 EM 1HEDb Statistics 98
A.86 PBI Wallet 1 EM 1JWm4 Statistics 98
A.87 PBI Wallet 1 EM 18FBc Statistics 98
A.88 PBI Wallet 1 EM 15kcN Statistics 98
A.89 PBI Wallet 1 EM 1L1Pd Statistics 98
A.90 PBI Wallet 1 EM 1Jjm5 Statistics 98
A.91 PBI Wallet 1 EM 1M4aa Statistics 99
A.92 PBI Wallet 1 EM 19mbJ Statistics 99
A.93 PBI Wallet 1 EM 1573u Statistics 99
A.94 PBI Wallet 1 EM 16CmR Statistics 99
A.95 PBI Wallet 1 EM 1BPCN Statistics 99
A.96 PBI Wallet 1 EM 1N4FE Statistics 99
A.97 PBI Wallet 1 EM 15Ztx Statistics 99
A.98 PBI Wallet 1 EM 1DqxW Statistics 99
A.99 PBI Wallet 1 EM 1Fv7D Statistics 100
A.100 PBI Wallet 1 EM 1FpkQ Statistics 100
A.101 PBI Wallet 1 EM 166vQ Statistics 100

A.102 PBI Wallet 1 EM 1JY63 Statistics 100
A.103 PBI Wallet 1 EM 1Amh5 Statistics 100
A.104 PBI Wallet 1 EM 168GD Statistics 100
A.105 PBI Wallet 1 EM 14BJR Statistics 101
A.106 PBI Wallet 1 EM 1P6DB Statistics 101
A.107 PBI Wallet 1 EM 1KFPo Statistics 101
A.108 PBI Wallet 1 EM 1EerL Statistics 101
A.109 PBI Wallet 1 EM 1JCre Statistics 101
A.110 PBI Wallet 1 EM 133b7 Statistics 101
A.111 PBI Wallet 1 EM 1CDaj Statistics 101
A.112 PBI Wallet 1 EM 1Jsgm Statistics 101
A.113 PBI Wallet 1 EM 16Syp Statistics 101
A.114 PBI Wallet 1 EM 13dKY Statistics 102
A.115 PBI Wallet 1 EM 1LSju Statistics 102
A.116 PBI Wallet 1 EM 17Xoc Statistics 102
A.117 PBI Wallet 1 EM 1Je2z Statistics 102
A.118 PBI Wallet 1 EM 1JgQK Statistics 102
A.119 PBI Wallet 1 EM 19TW2 Statistics 102
A.120 PBI Wallet 1 EM 1LxPF Statistics 102
A.121 PBI Wallet 1 EM 1AZ6T Statistics 102
A.122 PBI Wallet 1 EM 17dKq Statistics 102
A.123 PBI Wallet 1 EM 1NJA5 Statistics 103
A.124 PBI Wallet 1 EM 189T Statistics 103
A.125 PBI Wallet 1 EM 1Hjq Statistics 103
A.126 PBI Wallet 1 EM 19S7 Statistics 103
A.127 PBI Wallet 1 EM 19HaT Statistics 103
A.128 PBI Wallet 1 EM 1MXgv Statistics 103
A.129 PBI Wallet 1 EM 1789c Statistics 103
A.130 PBI Wallet 1 EM 1BTZ Statistics 103
A.131 PBI Wallet 1 EM 1Gb3 Statistics 103
A.132 PBI Wallet 1 EM 1DPE Statistics 103
A.133 PBI Wallet 1 EM 14y1o Statistics 104
A.134 PBI Wallet 1 EM 135S Statistics 104
A.135 PBI Wallet 1 EM 14ywq Statistics 104
A.136 PBI Wallet 1 EM 1ExUN Statistics 104
A.137 PBI Wallet 1 EM 1GEDX Statistics 104
A.138 PBI Wallet 1 EM 1DvN Statistics 104
A.139 PBI Wallet 1 EM 1HLqs Statistics 104
A.140 PBI Wallet 1 EM 1J3t Statistics 104
A.141 PBI Wallet 1 EM 1NmYX Statistics 104

A.142 PBI Wallet 1 EM 1NT5 Statistics 104
A.143 PBI Wallet 1 EM 1H5 Statistics 104
A.144 PBI Wallet MEM 14psc Statistics 105
A.145 PBI Wallet MEM 1LzN Statistics 105
A.146 PBI Wallet MEM 1H4h Statistics 105
A.147 PBI Wallet MEM 1GWt Statistics 105
A.148 PBI Wallet MEM 1DCQ Statistics 105
A.149 PBI Wallet MEM 13DE Statistics 105
A.150 PBI Wallet MEM 1CJy Statistics 105
A.151 PBI Wallet MEM 1BBey Statistics 105
A.152 PBI Wallet MEM 15Gd Statistics 106
A.153 PBI Wallet MEM 16pp Statistics 106
A.154 PBI Wallet MEM 18zU Statistics 106
A.155 PBI Wallet MEM 1Fox Statistics 106
A.156 PBI Wallet MEM 1BQ Statistics 106
A.157 PBI Wallet MEM 1KXo Statistics 106
A.158 PBI Wallet MEM 1Hyt Statistics 106
A.159 PBI Wallet MEM 17uj Statistics 106
A.160 PBI Wallet MEM 1FUk Statistics 106
A.161 PBI Wallet MEM 1Nrs Statistics 106
A.162 PBI Wallet MEM 16L5 Statistics 107
A.163 PBI Wallet MEM 1NYR Statistics 107
A.164 PBI Wallet MEM 1AP1 Statistics 107
A.165 PBI Wallet MEM 1K2J Statistics 107
A.166 PBI Wallet MEM 1Dja Statistics 107
A.167 PBI Wallet MEM 1B6G Statistics 107
A.168 PBI Wallet MEM 1BdD Statistics 107
A.169 PBI Wallet MEM 18Vf Statistics 107
A.170 PBI Wallet MEM 1Ng Statistics 107
A.171 PBI Wallet MEM 1Gvn Statistics 108
A.172 PBI Wallet MEM 1LYx Statistics 108
A.173 PBI Wallet MEM 13jV Statistics 108
A.174 PBI Wallet MEM 1FUv Statistics 108
A.175 PBI Wallet MEM 1Fs4 Statistics 108
A.176 PBI Wallet MEM 1K2H Statistics 108
A.177 PBI Wallet MEM 15db Statistics 108
A.178 PBI Wallet MEM 1BrD2 Statistics 108
A.179 PBI Wallet MEM 197Z Statistics 108
A.180 PBI Wallet MEM 1AA Statistics 108
A.181 PBI Wallet MEM 1A6M Statistics 109

A.182 PBI Wallet MEM 1P5Z Statistics 109
A.183 PBI Wallet MEM 1PeV Statistics 109
A.184 PBI Wallet MEM 1KUT Statistics 109
A.185 PBI Wallet MEM 1JNQ Statistics 109
A.186 PBI Wallet MEM 1F3w Statistics 109
A.187 PBI Wallet MEM 14Qo Statistics 109
A.188 PBI Wallet MEM 1MD Statistics 109
A.189 PBI Wallet MEM 1DpT Statistics 109
A.190 PBI Wallet MEM 1L6E Statistics 109
A.191 PBI Wallet MEM 1MCS Statistics 110
A.192 PBI Wallet MEM 13vH Statistics 110
A.193 PBI Wallet MEM 1Lew Statistics 110
A.194 PBI Wallet MEM 1G2S Statistics 110
A.195 PBI Wallet MEM 12T9 Statistics 110
A.196 PBI Wallet MEM 13Yo Statistics 110
A.197 PBI Wallet MEM 13af Statistics 110
A.198 PBI Wallet MEM 16qB Statistics 110
A.199 PBI Wallet MEM 1Ejb3 Statistics 110

Acknowledgments

This thesis would not have been possible without the generous guidance and support of my advisor, Professor David Wagner. I would also like to thank my collaborators Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, Professor Damon McCoy, Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Professor Kirill Levchenko and Professor Vern Paxson who also contributed to this work. I would particularly like to thank Professor McCoy and Professor Paxson for their help in connecting me to persons and projects throughout my time at UC Berkeley.

In addition, there are several institutions without whom this work would not have been possible. I would like to acknowledge the following funding sources: the National Science Foundation Graduate Research Fellowship Program, the Office of Naval Research, the Center for Long-Term Cybersecurity, Amazon "AWS Cloud Credits for Research", the U.S. Department of Education and Giant Oak. I would like to thank Google and Thorn for gifts that made this work possible, as well as Chainalysis for the use of their tools, and Scraping Hub and SRI for their assistance in collecting data for this work. The opinions in this thesis are those of its author and do not necessarily reflect the views of the sponsors.

Outside of research collaboration, I would also like to thank my family. To my sister-in-law Ruth, who walked my baby to sleep countless times during her visit so I could finish writing; to my mother-in-law and father-in-law Lynn and Marc, whose love and encouragement I greatly value. To my sister Jinju, without whose friendship, wisdom, summer visits and multiple hour long phone calls I would not have finished this PhD with my sanity intact. I owe a special thanks to my Oma, who sacrificed her time, energy and health more than any person I know, pouring out herself so that her daughters could achieve their goals. I also owe a special thanks to my Dad, who first introduced me to Computer Science, taught me how to code, co-authored my first paper with me, brainstormed research ideas with me, and made it possible for me to get to where I am today. To my dear husband Jacob, whose support, patience, computer science savvy and good humor carried me through to the end.

Finally, to my Father in Heaven, who led me to this work and will lead me through whatever comes next.

Chapter 1

Introduction

The Internet facilitates interactions among human beings all over the world, with greater scope and ease than we could have ever imagined. However, it does this for both the well-intentioned and malicious alike. As criminals increasingly make use of the Internet for a wide variety of ploys, their activity becomes more and more sophisticated, and requires an equally modern response. In this dissertation, we developed tools and techniques that can widely be used to analyze, classify and de-anonymize criminal forums and networks online. To that end, we focused on three main domains of criminal activity on the clear web and the Dark Net: classified ads advertising trafficked humans for sexual services, cyber black-market forums, and Tor onion sites hosting forums dedicated to child sexual abuse material (CSAM).

1.1 Classified Ads Selling Sex

Sex trafficking and slavery remain amongst the most grievous issues the world faces, supporting a multi-billion dollar industry that cuts across all nationalities and people groups [66]. With the advent of the Internet, many new avenues have opened up to support this pernicious business, including sites for online classified ads selling sex [51].

Although these ad sites provide a significant source of potentially incriminating data for law enforcement, monitoring these sites is unfortunately a labor-intensive task. The rate of new ads per day can reach into the thousands, depending on the website [70]. In addition, the nature of the advertising content can have a uniquely damaging psychological toll on its viewers. Picking out signs of trafficking requires domain expertise, creating an additional barrier for analytics. This problem space is made all the more difficult by the dearth of ground truth, e.g., ads known to be tied to trafficking activity vs. other consensual activity.

In conversation with our NGO and law enforcement collaborators, we have found

that there is a need for tools able to group ads by true owner. Such a tool would allow officers to confidently use timing and location information to distinguish between ads posted by women voluntarily in this industry vs. those by women and children forcibly trafficked. For example, groups of ads—posted by the same owner—that advertise multiple different women across multiple different states at a high ad output rate, is a strong indicator of trafficking. In this case, our goal is to distinguish which ads are owned by the same person or persons. This information can then be used to find traffickers, connections between pimps, or even trafficking networks.

All of the existing work in this problem space to date uses hard identifiers like phone numbers and email address links to define ownership [22]. Law enforcement considers this to be unreliable (as criminal organizations regularly change their phone numbers/use burner phones, and the cost of creating a new email address is low) but acknowledge it is the best link currently available. In fact, most of the work in this domain has focused on understanding the online environment that supports this industry through surveys and manual analysis ([13, 70, 51]). Almost no work has been done in building tools that can automatically process and classify these ads [22].

The aim of this chapter is to develop and demonstrate automatic techniques for clustering sex ads by owner. We designed two such techniques. The first is a machine learning stylometry classifier that determines whether any two ads are written by the same or different author. The second is a technique that links specific ads to publicly available transaction information on Bitcoin. Using the cost of placing the ad and the time at which the ad was placed, we link a subset of ads to the Bitcoin transactions that paid for them. We then analyze those transactions to find the set of ads that were paid for by the same Bitcoin wallet, i.e., those ads that are owned by the same person. As far as we are aware, this is the first work to explore this connection between paid ads and the Bitcoin blockchain, and attempt to link specific purchases to specific transactions on the Bitcoin blockchain.

In addition to reporting our results using our stylometry classifier on test sets of sex ads labeled by hard identifier, we apply both our tools to 4 weeks of scraped sex ads from Backpage, a well known advertising site that has faced multiple accusations of involvement with trafficking [48]. We assess the information gain between the set of owners found using just hard identifiers, vs. when adding our two methodologies. In summary, our contributions are as follows:

- We develop a stylometry classifier that distinguishes between sex ads posted by the same vs. different authors with 90% TPR and 1% FPR.
- We design a linking technique that takes advantage of leakages from the Bitcoin mempool, blockchain and sex ad site to link a subset of sex ads to Bitcoin public wallets and transactions.

- We propose two different methodologies that combine our classifier, our linking technique, and existing hard identifiers to group ads by owner.
- We evaluate our techniques on 4 weeks of scraped sex ads from Backpage, relying on the data automatically extracted using those two methodologies. We rebuild the price of each Backpage sex ad, and analyze the output of our two different methodologies.

1.2 Cyber Black-Market Forums

As technology evolves, abuse and cybercrime evolve with it. Much of this evolution takes place on underground forums that serve as both marketplaces for illicit goods and as forums for the exchange of ideas. Underground forums play a crucial role in increasing efficiency and promoting innovation in the cybercrime ecosystem. Cybercriminals rely on forums to establish trade relationships and to facilitate the exchange of illicit goods and services, such as the sale of stolen credit card numbers, compromised hosts, and online credentials.

Because of their central role in the cybercriminal ecosystem, analysis of these forums can provide valuable insight into cybercrime. Indeed, security practitioners routinely monitor forums to stay current of the latest developments in the underground ([46, 47]). Journalist Brian Krebs, for example, relied on forum data when he alerted Target to an ongoing massive data breach based on an influx of stolen credit card numbers being advertised for sale on an online forum [46]. Information gleaned from forums has also been used by researchers to study many elements of cybercrime ([30, 32, 35, 58, 79, 91]).

Unfortunately, monitoring these forums is a labor-intensive task. To unlock this trove of information, human analysts must spend considerable time each day to stay current of all threads and topics under discussion. Understanding forums also requires considerable domain expertise as well as knowledge of forum-specific jargon. Moreover, a forum may be in a foreign language, creating an additional barrier for the analyst. Often, what one wants from a forum is not a deep understanding of a particular topic, but an aggregate summary of forum activity. For example, one may want to monitor forums for an uptick in offers to sell stolen credit cards, a strong indicator of a major data breach. In this case, the goal is to extract certain structured information from a forum. Continuing the example, the task is first to identify offers to sell credit card numbers and then extract from the post information like quantity and price. We can then use this structured data to carry out analyses of market trends, like detecting a sudden increase in supply.

In this chapter, we aim to develop and demonstrate automatic techniques for extracting such structured data from forums. Although extracting structured data from

unstructured text is a well-studied problem, the nature of forum text precludes using existing techniques that were developed for the well-written English text of the Wall Street Journal. In contrast, forum posts are written in their own specialized and rapidly evolving vocabulary that varies from forum to forum and ranges from ungrammatical to utterly incomprehensible. As a result, off-the-shelf Named-Entity Recognition (NER) models from Stanford NER perform poorly in this dataset. Another approach is to use regular expressions to identify occurrences of the words related to the type of a post, well-known products, and prices. This simplistic approach also fails because different users use different words for the same products.

Rather than aiming for complete automatic comprehension of a forum, we developed a set of natural language processing building blocks aimed at a set of precise tasks related to trade that a human analyst might require when working with forum data. As we show in this chapter, this approach allows us to extract key elements of a post with high confidence, relying on a minimal set of human labeled examples. By focusing on extracting specific facts from a post, our tools make automatic analysis possible for text inaccessible using conventional natural language processing tools.

In summary, our contributions are as follows:

- We develop new natural language processing tools for a set of precise data extraction tasks from forum text of poor grammatical quality. In comparison with a simple regular expression-based approach, our approach achieves over 9 F-point improvement for product detection and over 40 F-point improvement for price extraction.

- We evaluate our tools on a set of eight underground forums spanning three languages.

- We present two case studies showing how to use our tools to carry out specific forum analysis tasks automatically and accurately, despite the poor quality of the data.

1.3 Child Sexual Abuse Material Forums

Behind every image and video of child pornography, there is a real child who is being sexually abused and victimized [6]. The advent of the Internet and the Dark Net has not only markedly increased the proliferation of CSAM, but also created a new space for pedophiles to find and target future victims for abuse [11]. Most of the work in this domain has focused on CSAM in peer-to-peer networks, with researchers performing measurement studies ([90] [37] [77]) or building classifiers to detect CSAM ([64] [20] [74]). To the best of our knowledge, no one has yet undertaken any analysis or

tool-building geared towards processing and classifying data on CSAM forums hosted on Tor onion sites.

This is a missed opportunity. Tens of thousands of pedophiles visit these forums every day, sharing CSAM, tips on how to groom children online and in person, and advice on how to remain secure and safe from law enforcement. Some of these people who visit are currently, actively sexually abusing children, and sharing this content with their peers.

In conversation with our NGO and law enforcement collaborators, we have found that there is a need for tools able to distinguish between producers (those who are actively, hands on abusing children and producing CSAM content) and those who are consuming said content. Such a tool would allow officers to focus their attention on finding both the abusers and the children who are currently being abused, saving valuable time and effort that otherwise would need to be spent manually reading through posts. Promptly finding active abusers could mean the difference between a child sexually abused for weeks, instead of years.

In this chapter, we aim to develop and demonstrate automatic techniques for extracting these producers from forums. Specifically, we formulate this as an anomaly detection problem. As there is very little ground truth for which users are producers, we use an unsupervised approach: principal component analysis (PCA). PCA is a well-known dimensionality-reduction technique that has been used in prior work to detect network traffic anomalies. It has also been used in one previous work to uncover anomalous user behavior in online social networks [88]. Although using PCA for anomaly detection is a well-studied problem, using PCA to uncover anomalous user behavior is mostly unexplored. Using it to uncover anomalous user behavior in forums is entirely unexplored, and requires taking into account both the unstructured nature of the text, and an organizational structure that varies from forum to forum.

In this work, using the PCA anomaly detection framework from [88], we experiment and find a feature set that is the most discriminating between producers and normal users. In summary, our contributions are as follows:

- We develop a set of features that is most discriminating between producers and normal users on publicly viewable forum data.
- We evaluate our feature set on four distinct Tor CSAM forums.
- We evaluate our feature set on four combined Tor CSAM forums.

The remainder of this dissertation examines tool and technique design and implementation, case studies, experimental results and prior work. Each chapter covers one of the three domains we introduced above. In the final chapter we present concluding remarks and a reiteration of key findings. Portions of the work reported in this

dissertation have previously been reported in preliminary form in earlier papers: [68], [67].

Chapter 2

Classified Ads Selling Sex

2.1 Introduction

Sex trafficking and slavery remain amongst the most grievous issues the world faces, supporting a multi-billion dollar industry that cuts across all nationalities and people groups [66]. With the advent of the Internet, many new avenues have opened up to support this pernicious business, including sites for online classified ads selling sex [51]. Monitoring these sites is a labor-intensive task; the rate of new ads per day can reach into the thousands, depending on the website [70]. An automated solution is necessary to parse through this data to provide useful insights to law enforcement. In this chapter, we develop and demonstrate automatic techniques for clustering sex ads by owner (e.g., the person or persons who paid for the placement of the ad). Our goal is to distinguish which ads are owned by the same person or persons. This information can then be used to find traffickers, connections between pimps, or even trafficking networks.

The rest of this chapter is organized as follows. Section 2 provides the necessary background for the rest of the chapter. Section 3 outlines Backpage and Bitcoin, which we analyzed and used to evaluate our tools. Section 4 describes the methodology for building our stylometry classifier, covering ground truth labeling, the model we built, and validation results. Section 5 describes our linking technique. Section 6 describes our two proposed methodologies that combine our classifier, linking technique and existing hard identifiers (phone numbers and email addresses) to group ads by owner. Section 7 reports our findings when exploring the 4-weeks of scraped sex ads from Backpage, and Section 8 discusses limitations and future work. We conclude with reiteration of key contributions and findings.

2.2 Related Work

2.2.1 Sex Trafficking Online

Ecosystem Analysis Much of the research in this area to date has focused on surveys, manual analysis and meta-studies to better understand the existing online environment which allows for and encourages the trafficking of humans for sexual service ([13, 14, 40, 51, 70]).

These surveys all found that the majority of US-based trafficking victims are advertised online. In Bouché's survey of 115 sex trafficking survivors [14], 63% of participants reported being advertised online. Of those, almost half reported that they were advertised on Backpage; Craigslist and Facebook were the other most popular websites for advertising. The survivors interviewed were contacted via 14 different anti-trafficking organizations. Survivors completed paper surveys remotely, and then either mailed or scanned and emailed their responses. Additionally, 77 of the participants consented to follow-up interviews, where the audio recordings of the interviews were kept for qualitative analysis. Latonero et al. [51] outlines several criminal cases and news stories of traffickers using online classified sites such as Backpage to sell their victims ([27, 86, 87]). In one chilling case from 2010, New York gang members reportedly advertised girls as young as 15 on Backpage, beating and starving them if they did not make at least $500 a day performing sexual services [73].

Sykiōtou [82] outlines how traffickers recruit victims on the Internet. They discuss a large set of existing research and resolved criminal cases in this space, and observe two main avenues of recruitment. First, offering fake jobs (ranging from mail-order bride agencies to traditional employment such as waitress/model/home help), and second, Internet chat sites, where traffickers will "befriend" and lure in young victims. They found that victims are normally, but not always, recruited in their own countries, and on reaching their destination, are then taken over by a local contact. Bouché [14] found that the more recently trafficked respondents were significantly more likely to have met their controller on the Internet (r=0.22), although the majority (81%) of participants still reported meeting their controller in person. A report released by Shared Hope International [4] notes that "pimps, madams, and escort agencies recruit new members through their own websites ... and Facebook accounts."

Some researchers ([40] [13]) have also focused their attention on "john boards", online forums where men trade information with one another on buying sex with women. Janson et al. [40] manually analyzed posts made on the USA Sex Guide from June 1, 2010 to August 31, 2010 by men who buy sex in Illinois. Forum members on the USA Sex Guide made 2,466 entries during this time period, of which 1,684 posts (68.2%) were analyzed. The authors found that these forums are used by men to normalize the buying of sex, act as a training ground to inculcate men throughout Illinois in the

etiquette and social organization of the commercial sex industry in their communities, and provide a space to share tips/advice on the "best" places to buy commercial sex.

Blevins and Holt [13] manually analyzed an un-named john board website that had active forums in multiple U.S. cities, focusing on analyzing the unique language used by forum members. The authors learned several insights, including (1) johns promote the notion that paid sexual encounters are normal, giving higher status to "pooners" (those with extensive experience purchasing sex), (2) johns treat sex workers and sexual acts as a commodity, referring to sex workers by object-specific acronyms and regularly describing the build/physique of sex workers in a way that emphasizes them as services or goods rather than human beings, and (3) johns stress the importance of sexual acts and the way that sex is experienced with prostitutes, with many of the posts dedicated to depicting the types of sex acts and services certain prostitutes provide.

Classifying Sex Ads Automatic analysis in the sex trafficking space is still fairly sparse, as this is an area of research just recently gaining interest in the larger computing community. What does exist has primarily focused on using machine learning to detect instances of human trafficking in escort advertisements, and using machine learning and social network analysis to detect human trafficking entities and networks in general ([22, 38]). Our study is the first to create automated techniques using a combination of text-based and financial data for conducting large-scale clustering of escort ads.

In [22], Dubrawski et al. present a bag-of-words machine learning model to identify escort ads from Backpage that likely involve human trafficking. Using phone numbers of known traffickers as ground truth, with a false positive rate of 1% they achieve a true positive rate of 55%. They also present an entity resolution logistic regression model to group ads authored by the same person, or advertising the same person/group of people. Using personal features (age, race, physical characteristics) and operational characteristics (locations, movement patterns) with hard identifiers as ground truth, the authors conducted a small empirical evaluation with a balanced test set of 500 pairs of ads, achieving a 79% true positive rate at a false positive rate of 1%. They were also able to demonstrate some stand-alone cases where their model successfully tracked one author's ad record over the course of a year, even with phone number and a few other characteristics changing between advertisements.

Ibanez and Suthers [38] analyze Backpage sex ads using semi-automated social network analysis to detect human trafficking networks going into, and operating within, the state of Hawaii. In order to focus their attention on ads indicating trafficking, the authors first analyzed these ads for signs of trafficking derived from a list of indicators produced by the United Nations Office on Drugs and Crime [7] and the Polaris Project [5]. These indicators include: different ages used, different aliases used, multiple locations, third person language, references to ethnicity, and incalls only. 82% of the ads contained one or more indicators and 26% contained three or more indicators. They

then built a graph representing the movement of these escorts by extracting the state of origin for phone numbers listed in the ad (using the area code), and the various locations where the ad was listed. 208 total phone numbers were analyzed, and of those, 165 indicated movement. From that set, the authors discovered a potential trafficking network going from Portland, Oregon to Hawaii, as well as smaller trafficking networks within Hawaii proper.

2.2.2 Bitcoin

Bitcoin is a decentralized peer-to-peer pseudonymous payment system where users can transfer bitcoins among one another in the form of *transactions* that exchange this digital currency[1]. Succinctly described, a bitcoin is owned by a public encryption key, typically called a *wallet* or *address*. Transactions in practice are performed using the ECDSA signature scheme, where the owner of a bitcoin signs a statement agreeing to transfer ownership of bitcoin to another wallet (i.e., public key). Bitcoin is pseudonymous in that all transactions from a single wallet are linked to the same owner, but the same person can use many different wallets and these transactions will not be directly linkable. There have been many prior studies that point out limitations in the pseudonymous property of Bitcoin, when used in practice, and present methods for linking chains of bitcoin transactions from different wallets to the same owner ([9, 56, 71]). In this work, we leverage some of these bitcoin linking methods.

As there is no central authority, senders first broadcast their transactions across the Bitcoin peer-to-peer network, which consists of individual volunteer nodes that each maintain the full state of the network. Upon receiving the transactions, each node stores them into a temporary storage area known as the "mempool". Transactions in the mempool *may* be selected for *mining*, a process that is meant to secure bitcoin transactions and ensure the integrity of the distributed ledger (i.e., *blockchain*). There is no guarantee whether and when a transaction will be included into the blockchain, but if this happens, the transaction will be removed from the temporary mempool and included in the permanent blockchain.

Many merchants will wait until a valid bitcoin transaction is included in the blockchain before considering it completed, which will take on average 5 minutes to occur. However, the delay time between when a client broadcasts a bitcoin transaction over the Bitcoin peer-to-peer network and when it is included in the blockchain is variable and can take hours when the network is overloaded. Due to this delay some merchants, such as Backpage, choose to not wait and accept the payment as completed once a valid bitcoin transaction appears in their mempool. By not waiting the merchant is accepting the risk of the customer performing a double spend attack [43] that causes

[1]The standard is to use Bitcoin to refer to the system and bitcoin or BTC to refer to the digital currency.

the transaction to the merchant to be invalidated before it completes. To the best of our knowledge, there is no prior method for linking an online ad posting to the bitcoin transaction that paid for the online ad.

2.3 Datasets

2.3.1 Backpage

In this work, we focus our analysis and case study on data from Backpage, one of the most popular sites for online classified ads selling sex [51]. Backpage is widely known to be a popular domain used by traffickers to advertise their victims ([14] [51] [70]). Ernie Allen, president and CEO of the National Center for Missing and Exploited Children, makes the danger clear: "[O]nline classified ads make it possible to pimp these kids to prospective customers with little risk [for the pimp]" [73]. Obviously, in order to protect the pimp/trafficker, none of these ads explicitly state any coercion; the ads are either written as if from the perspective of the victim herself, or describing the victim being sold in the third person, with no mention of a pimp or trafficker in either case.

Backpage has been running since 2004, with listings all over the world. Any person who has an email address can register for an account on Backpage and post ads. Although the site offers a wide variety of different types of classified listings (e.g., automative, rentals, furniture), in this work we focus our attention on the "adult entertainment" listings, which contain about 80% percent of the U.S. market for online sex ads in America [48]. There are several different sub-categories in this section, namely: escorts, body rubs, strippers/strip clubs, dom/fetish, trans, male escorts, phone/websites, and adult jobs. On July 1, 2015, Visa and MasterCard stopped processing transactions for adult listings on Backpage, which caused Backpage to switch to Bitcoin payments for all paid adult ads. GoCoin, a third-party Bitcoin payment processor company, currently manages all Bitcoin payments for Backpage adult ads. As of January 9, 2017, the adult listings section of the website has been blocked, in response to ongoing legal action against Backpage for their role in the marketing of minors. All of our data was collected before that point.

We have two different forms of access to this data. First, we have a scrape containing 1,164,663 unique ads from January 2008 to September 2014. We define "author" to be an entity tied to a set of hard identifiers that co-occur in any given ad: i.e., if hard identifiers A and B occur in one ad, and hard identifiers B and C occur in another ad, the author of those two ads consists of the hard identifiers A, B and C. By processing all the ads and linking together phone numbers and email addresses, we discerned that we have 336,315 authors in this dataset. This data was used to build and assess our authorship classifier. Second, we conducted a scrape from December 11, 2016 to January 9, 2017, collecting all adult ads placed in the United States every hour. This data was

used in our case study. Using the same definition of authorship as above, this scrape contains a total of 741,275 unique ads and 141,056 authors.

Dates	No. Unique Ads	No. Authors	Locations
1/2008-9/2014	1,164,663	336,315	Global
12/2016-1/2017	741,275	141,056	United States

Table 2.1. Backpage

2.3.2 Bitcoin

A registered user can post Backpage ads for free, but premium features, such as posting a single ad across multiple locations or bumping an ad to the top of a listings page, will require payment. For adult entertainment ads, bitcoin or a hand mailed check are the only acceptable forms of payment. GoCoin processes all bitcoin payments on Backpage.

Each purchase of premium features, however many, is represented as a single invoice. Users also have the option to deposit an arbitrary amount of bitcoins as credits; each purchase paid for via credit would withdraw funds from those pre-deposited credits. For each invoice, Backpage dynamically generates a fresh wallet address that belongs to GoCoin, along with the bitcoin amount. A user can either transfer bitcoins from his own personal wallet address into the fresh address, or he can use a third-party wallet service such as Paxful. Bitcoins received by the fresh address are subsequently aggregated into some central wallet address of GoCoin, along with bitcoins received by fresh addresses for other users.

When a user transfers bitcoins into the fresh wallet address, the corresponding transaction typically appears on the Bitcoin peer-to-peer network within seconds. Once Backpage sees the transaction on their mempool, the premium features take effect and the ad appears on the listings page, without the user having to wait for the transaction to be confirmed into the blockchain. For example, if a user purchases a premium feature that posts an ad across multiple locations, the timestamp at which the ad appears across multiple locations is approximately the timestamp at which the transaction is propagated on the Bitcoin network. We discovered this by placing ads ourselves and comparing the timestamp at which the ads appeared on Backpage, with the timestamp at which the transaction first appeared on the peer-to-peer network.

In fact, this timing proximity allows us to link Bitcoin transactions, as they first appear on the peer-to-peer network, with Backpage ads. Before we can establish such links, however, we need to know exactly when a transaction first appears on Bitcoin's peer-to-peer network. To this end, we build a tool that snapshots the state of the network at a fine granularity.

In particular, a collaborator (Danny Yuxing Huang) runs the default Bitcoin client at our research institution. The client maintains the up-to-date blockchain, and it also allows us to query the mempool state via the `getrawmempool` API call. The mempool state is dynamic; new transactions are broadcast over the Bitcoin peer-to-peer network, while some existing transactions are removed from the mempool as they are confirmed into blocks. To this end, we set up an automated script that saves a snapshot of the mempool state every minute. Using these per-minute snapshots along with the timestamps of the snapshots, we can find the earliest timestamp at which our Bitcoin client received a transaction. These timestamps, which we call *mempool timestamps*, estimate the first time a transaction appears on the Bitcoin network. Since our Bitcoin client runs on a low-latency gigabit research network, we assume the mempool timestamps are a reasonable approximation for the true timestamps at which the transactions were sent.

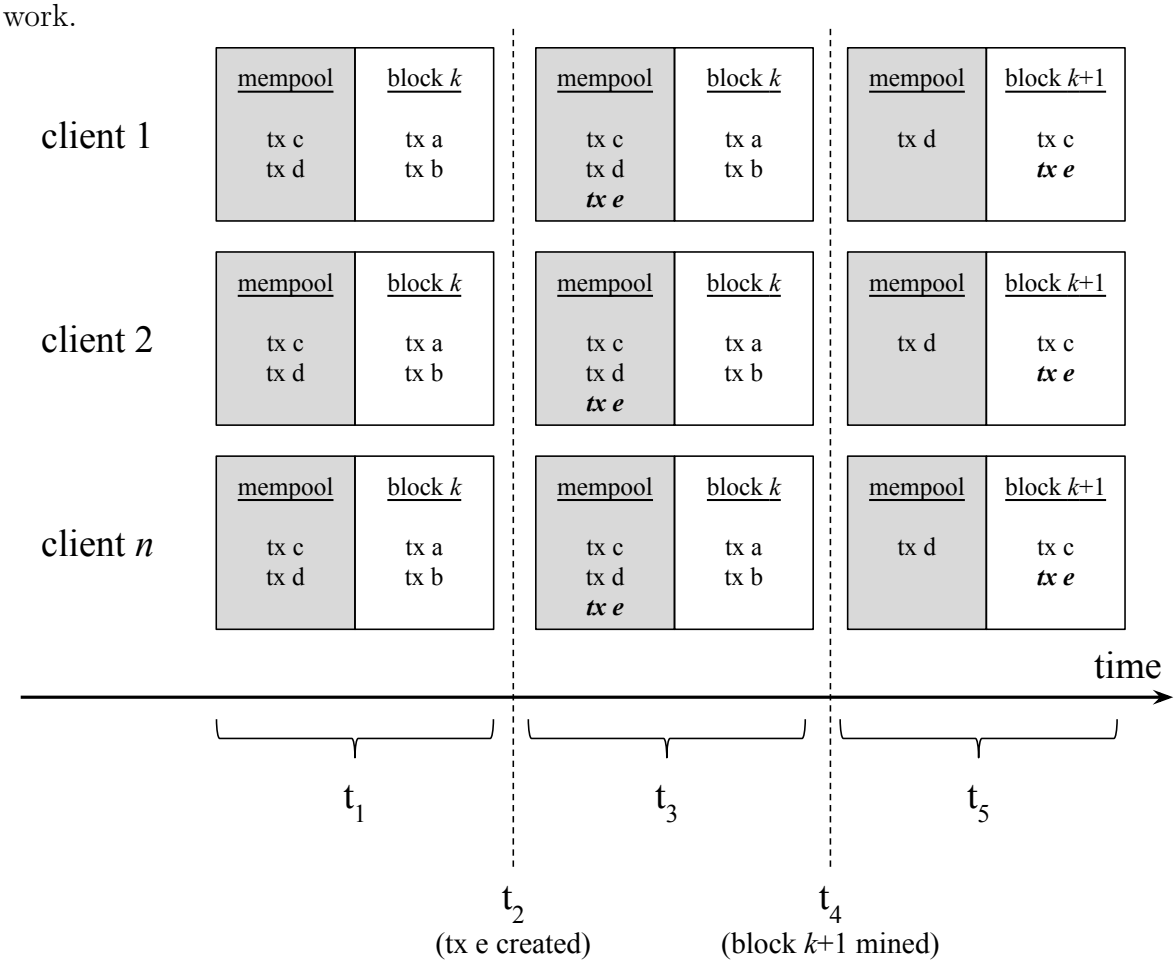
To illustrate how we use the methodology above to link the timestamps of Bitcoin transactions and Backpage ads, we consider a hypothetical example as shown in Figure 2.1, which depicts a peer-to-peer network of n Bitcoin clients. Each of the clients maintain two pieces of state: the mempool and the blockchain. Time t_1 shows a snapshot of the n nodes. All of them currently have block k confirmed, which includes transactions ("tx") a and b. At the same time, all the n clients have both transactions c and d held in the mempool, waiting to be confirmed into the next block $k+1$.

At time t_2, let us assume that someone purchased an escort ad using transaction e. This new transaction is spread across Bitcoin's peer-to-peer network. At time t_3, which we assume is a few seconds after t_2, transaction e appears in the mempools of all the clients in the network. Since we maintain a Bitcoin client ourselves and we snapshot its mempool every minute, we are likely to detect the presence of transaction e also in t_3. Here, t_3 is the mempool timestamp for transaction e. Backpage is also likely to detect the presence of transaction e, and typically will post the corresponding ad within a minute.

At this point, however, transaction e remains unconfirmed, as it is in the mempool rather than the blockchain. Let us say that at time t_4, block $k+1$ is mined, and the miner of the block decides to include transaction e. Subsequently, at time t_5, transaction e is removed from the mempool and added to the global blockchain (as is transaction c, for the same reason). Because the transaction is confirmed, we can now use Chainalysis to identify if this transaction e sent bitcoins to GoCoin (explained later in this section).

We continuously snapshot the mempool state from Oct 24, 2016 to Jan 20, 2017 and obtained 16,767,921 transactions that were later confirmed into the blockchain. Not all the transactions are relevant to our analysis. We focus on transactions that are likely to have sent bitcoins to GoCoin. We use two methods to identify transactions to GoCoin.

Figure 2.1. Example of how a new transaction is added to Bitcoin's peer-to-peer network.



(1) Chainalysis Labels Chainalysis is a private company that clusters and labels identities on the blockchain. In particular, it repeatedly deposits bitcoins into and from GoCoin, so that Chainalysis can obtain a list of fresh Bitcoin wallet addresses generated for every deposit it makes. Even though these wallet addresses are specific to one user, eventually the bitcoins from them are transferred, along with other user deposits, to GoCoin's central wallets. Since other users' deposit wallet addresses appear in the same transaction inputs as Chainalysis' wallet addresses, Chainalysis can cluster all these addresses together and label them as GoCoin. In this way, Chainalysis is able to discover wallet addresses used for making payments to GoCoin. Through its subscriber-only API, we can check if a particular transaction made payments to GoCoin.

(2) GoCoin Heuristics While Chainalysis' technique provides us with the ground truth for transactions to GoCoin, it is unable to discover *all* GoCoin transactions. To account for false negatives, we develop heuristics to identify possible GoCoin transactions. By analyzing many GoCoin transactions that Chainalysis identified, we found that GoCoin transactions have the following features: (i) the fresh wallet address appeared in exactly two transactions—one for receiving bitcoins from the user, and the other for sending the bitcoins into some aggregation wallet address; (ii) the deposited bitcoin amount is always less than 1 BTC and has between 3 and 4 decimal places (e.g., "0.0075 BTC"); (iii) the bitcoins are aggregated along with other bitcoins that follow Feature (ii); and (iv) all these bitcoins are aggregated into a single multi-signature wallet address (i.e., a wallet requiring more than one key to authorize a Bitcoin transaction) that starts with the number "3". We label any wallet address that meets all four conditions as GoCoin-heuristic. We note that this technique may introduce false positives—i.e., transactions that resemble GoCoin transactions but in reality are not GoCoin.

During the period when we snapshot the mempool state, we labeled 753,929 distinct wallet addresses as either Chainalysis-GoCoin or Heuristic-GoCoin. Of these addresses, 1.5% are Chainalysis-GoCoin only, 69.6% are Heuristic-GoCoin only, and 29.0% have both labels.

2.4 Author Classifier

For any given two ads appearing on a site, we extract the authorship similarity, determining whether the ads are written by the same or different author. We take a supervised learning approach, labeling a randomly sampled subset of the data with ground truth and using those annotations to train our classifier to label the rest. We build a binary classifier that takes a pair of ads as inputs and outputs 'same' if the ads are written by the same author or 'different' otherwise.

2.4.1 Labeling Ground Truth

Our ground truth labeling uses hard identifiers (phone numbers and email addresses) to define ground truth authorship. Any set of co-occurring phone numbers/email addresses within the full set of ads is considered a unique author. We labeled all the data available to us in this way.

2.4.2 Models

We consider two models for authorship classification. For both models, we experimented with using multiple different machine learning algorithms for training. We

achieved best overall performance using logistic regression. We train the logistic model by coordinate descent on the primal form of the objective [26] with ℓ_2-regularization.

WritePrints Limited This model uses a limited section of the Writeprints [92] feature set,[2] consisting mainly of counts of characters, words and punctuation. This feature set has been widely used for authorship attribution. We consider the WritePrints feature set to be appropriate for our domain because the ads are similar to other short texts, such as tweets and product reviews, where this feature set has been used successfully. For each pair of ads, we extract the Writeprints limited feature set for each ad, resulting in two feature vectors. We obtain the final feature values by normalizing globally, then subtracting these two vectors and taking the absolute value of each coordinate. We use this model as a baseline for evaluating the performance of our Jaccard and Structure model.

Jaccard and Structure This model uses a variety of text-based features: word unigrams, word bigrams, character n-grams, parts of speech, and proper names, as well as a structural feature: the location and spacing of line breaks in the post. We extracted the parts of speech using the Stanford POS tagger [3]. We extracted proper names by matching all unigrams to a list of names we compiled from various online resources, and from reading through the ads. For each pair of ads, we extract the relevant set (e.g., all adjectives) appearing in each ad, and calculate the Jaccard (on the text-based values) and cosine (on the structural values) similarity of the two sets.

2.4.3 Validation Results

We assessed our classifier on strictly non-overlapping sets of authors between the training and testing datasets, in order to ensure that the classifier was learning the concept of 'same' vs. 'different', and not just learning the stylometry for the particular set of authors. Ultimately, we built three separate training/testing datasets.

Building Validation Sets In our initial pre-process, we removed all ads that were exact duplicates of each other (leaving only one copy of each duplicate in the final set) as well as all ads that had fewer than 50 words in the ad. We defined an exact duplicate as ads that were byte-for-byte identical. From this set, we randomly sampled 5,000 authors with at least two ads each. From each of these authors, we randomly sampled two ads. The resulting 10,000 ads were used to create 2,500 'same' instances (where

[2]We do not use the full set of Writeprints features, since it is too computationally expensive to run on larger sets of pairs.

Model	TPR Average	FPR Average
Jaccard & Structure LR	89.54%	1.13%
Writeprints Limited LR	83.06%	16.93%
Jaccard & Structure SVM	87.72%	0.81%
Writeprints Limited SVM	85.53%	14.50%
Jaccard & Structure Naive Bayes	85.90%	2.32%
Writeprints Limited Naive Bayes	75.45%	24.55%
Jaccard & Structure Random Forest	89.15%	1.25%
Writeprints Limited Random Forest	86.05%	13.95%
Jaccard & Structure Random Subspace	89.76%	1.30%
Writeprints Limited Random Subspace	85.07%	14.94%
Jaccard & Structure AdaBoost	90.31%	1.87%
Writeprints Limited AdaBoost	81.34%	18.66%

Table 2.2. Classification accuracy for same vs. different author.

each same instance represents two ads written by the same author). We then randomly sampled 5,000 pairs of authors, where the two authors in each pair are distinct from each other, from the original sample of 5,000 authors. We then randomly sampled one ad from each of the authors in a given pair. These two ads (one for each author in a pair) were used to create 5,000 'different' instances (where each different instance represents two ads written by different authors).

We repeated this process three times, with non-overlapping sets of 5,000 authors. In this way, we created three separate training/testing datasets, with each one consisting of 7,500 instances: 5,000 different instances and 2,500 same instances. We chose this class balance in order to reflect the underlying nature of the data (i.e., there are more ad pairs with different authors than the same author). We evaluate our tool with all six different training/testing combinations, training on one of the datasets and separately testing on the other two, for all pair-wise combinations of the three datasets.

Results In all cases, the same vs. different author classifier is effective, achieving 89.54% true positive rate and 1.13% false positive rate on average. This indicates that the classifier is not just learning to distinguish ads written by a specific set of authors, but is learning the concept of same vs. different in general. This is necessary for this domain of sex trafficking, where new victims are recruited daily, and there is no guarantee of a permanent set of traffickers persisting through time. In addition, the classifier significantly and consistently outperforms the baseline Writeprints Limited model; the accuracy for the same author class improves slightly, and the accuracy for the different author class improves dramatically in all cases.

We reviewed a random sample of the false positive and false negative cases from

Figure 2.2. Example of a false positive case with possibly flawed ground truth.

Title: Visiting and can't wait to meet you!
Ad: Allow me to pamper you with my limitless skills and talents.
I have a flawless body with curves in all the right places.
I am extremely down to earth , smart, sexy, and sophisticated and adventurous girl who loves to have fun!
Are you ready for a hot 1-on-1?

Real & Independent
Non-Rushed
Ultimate Relaxation
Utmost Discretion

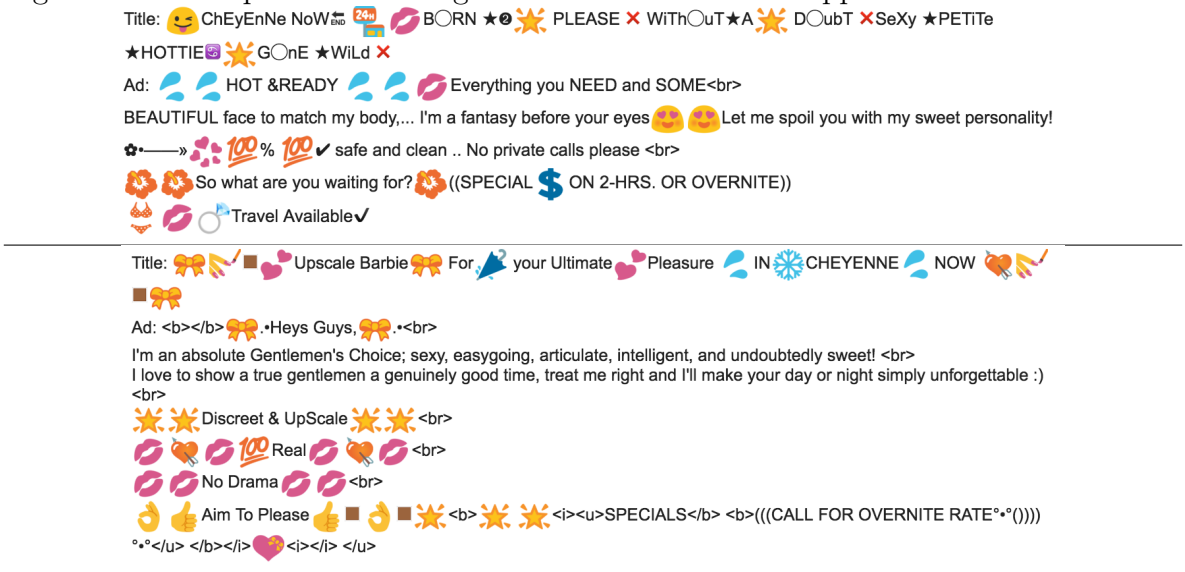
Title: 💋 Domincan Beauty Queen 💎 Treat Yourself to high class ($)
Ad:. 💋 .•Allow me to pamper you with my limitless skills and talents. High Class Companionship at it's finest 💋 I am an extremely down to earth, smart, sexy, sophisticated and open minded girl who just loves to have fun! 💋 I am 100% independent...100% accurate photos 💋. I am looking for that special someone who wants to spoil me....and get spoiled in returns !• 💋 THANK YOU; !!!! 💋 💋 !!!!!. Hourly Dates Overnights and weekend dates Kimmy 💋 💋 💋 💋 more beautiful in person....guarantee good time

our authorship classifier. Since we use phone numbers and email addresses as ground truth, it is possible (and in some cases, appears to be the case) that the false positives are actually true positives. Figure 4.5 shows one such case, where the ad posters used different phone numbers and different formatting to present the exact same textual content; our authorship classifier considered them to be written by the same author. It is possible one author randomly selected, copied and pasted the text from the other, and that there is no shared owner; given the lack of definitive ground truth it is not possible to know for sure. For false negative cases, we found that the classifier misidentifies ad pairs where the writing style is completely different (Figure 2.3).

2.5 Linking Ads to Bitcoin Transactions

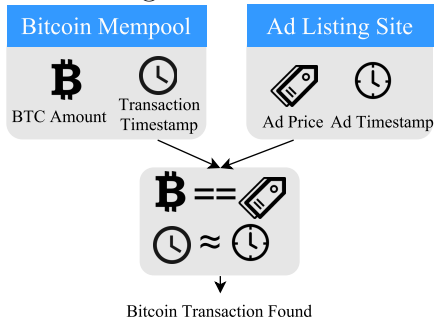
This section describes our method for linking a particular bitcoin transaction to its corresponding ad. Suppose A is the set of ads on the target ad site S where an individual ad $a \in A$. T is the set of bitcoin transactions whose output wallet belongs to the Bitcoin payment processor for site S. We construct an undirected bipartite graph, $G = (V, E)$, where the set of vertices is $V = A \cup T$, and the set of edges E contains an edge between two nodes, $a \in A$ and $t \in T$, if t is a possible transaction for a. We consider t to be a possible transaction for a if the cost of posting a equals the value of t, and the difference between the timestamp of a's appearance on the listings page and the timestamp when t is first observed on either the mempool or the blockchain is less than a threshold. The threshold depends on the particular ad site. For example, the threshold for Backpage is one minute; we discovered this threshold by placing dozens of ads and observing the timestamps. Backpage accepts the payment for an ad as completed, and posts said ad on S, as soon as t appears in their mempool. We then

Figure 2.3. Example of a false negative case where the ads appear in different sections.



observe that transaction in our mempool within one minute. In this case, the value of the threshold is simply the amount of time it takes for the transaction t to appear in our mempool: one minute.

Figure 2.4. Linking Ads to Bitcoin Transactions



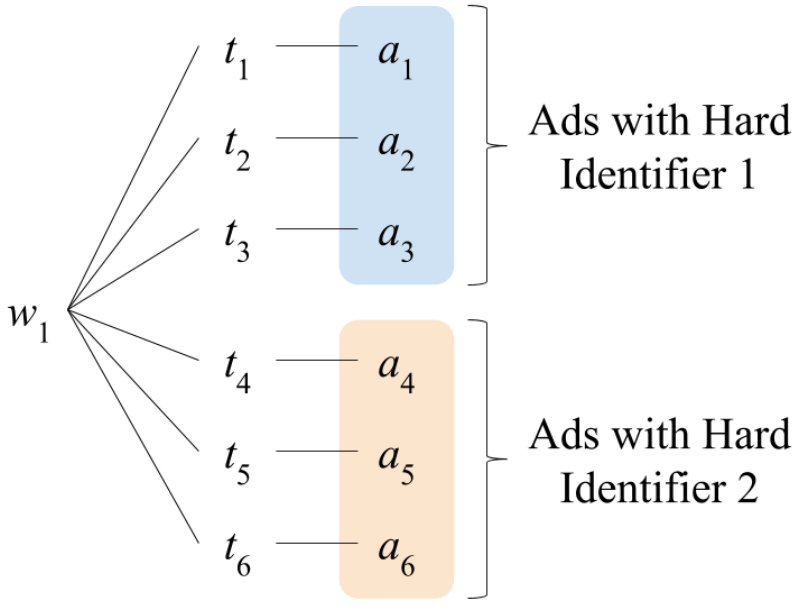
The edge between transaction t and ad a could be a false positive if the transaction, in reality, is not linked to the ad. For example, if a was not paid for in bitcoin, and there happened to be a transaction t around the same time with the same cost, that would result in a false positive. The edge could also be a false positive if the transaction's output wallet is mistakenly labeled as belonging to the Bitcoin payment processor for site S. False negatives, where there is a missing edge between t and a, are also possible. False negatives can occur if we wrongly reconstructed the price of a and thus failed to link a to the corresponding t.

2.6 Grouping Ads by Owner

We propose two methods for grouping ads by owner: grouping by shared author (e.g., a set of co-occurring hard identifiers) and grouping by persistent Bitcoin identities. The former was designed by a collaborator (Danny Yuxing Huang). Our methods assume two sources of data: the cost to post each ad a in A, and the timestamp of a's appearance on the target site. The mechanism for collecting this timestamp data and rebuilding the cost will vary from site to site. In section 2.7, we demonstrate how we did both for Backpage.

Both methods have T to be a set of bitcoin transactions to GoCoin, such that each transaction $t \in T$ has exactly one input wallet address w that is not multi-signature (i.e., does not require more than one key to authorize a Bitcoin transaction), and at least one of the output wallet addresses in t is either labeled as Chainalysis-GoCoin or Heuristic-GoCoin. The ultimate goal is to map an ad a to its true owner wallet address w, from among all W (the set of wallet addresses on the blockchain). Our first method tackles this problem by finding a wallet address w that links together multiple existing authors. Our second method focuses on persistent Bitcoin identities (defined below).

Figure 2.5. Shared Authors



2.6.1 Grouping by Shared Author

In practice, our inference typically results in a mapping between a transaction t and an ad a that is many-to-many. This many-to-many mapping between t and a makes it difficult to map a wallet address w to a. To this end, we construct a subgraph $G' \subset G$, where G is an undirected graph with A, T, and W as the vertices. In G, an edge between a w node and a t node exists if wallet w is the sole input wallet address of transaction t. We require subgraph G' to satisfy all of the following criteria, applied in order.

1. Each t node should be adjacent to exactly one w node, because we already require every transaction in G to have a single input wallet address. However, we allow each w to be adjacent to one or more t, as a wallet address may be used across multiple transactions.

2. With exactly two hops (from w to t to a), each w node should be able to reach at least three a nodes with the same author. This reachability suggests that w is likely to be the true owner for at least three of the a nodes; the presence of the shared author reduces the probability of having incorrect edges between t and a.

3. Each t should be adjacent to exactly one a. By transitivity, each a can reach exactly one w with two hops. In other words, one cannot find another wallet address, other than w, that can be mapped to ad a. This criteria attempts to further reduce the probability of incorrect edges between t and a.

   In an effort to link together multiple authors, we add one more criteria:

4. With exactly two hops, each w must be able to reach at least two sets of a nodes with different authors. This suggests that these authors are likely to be related, in that they might have all used w to pay for the ads.

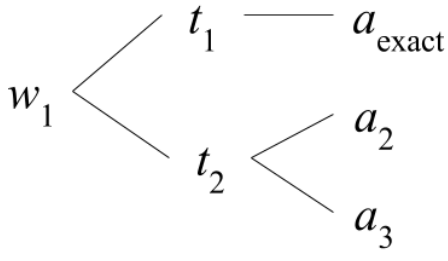
We define the resulting w in G' to be a `shared author` wallet (SA wallet). The last criteria allows us to find at least two sets of ads of different authors that are mapped to the same w.

Figure 2.5 shows a hypothetical example of a subgraph G' that satisfies all our criteria. In particular, the wallet address w_1 is associated with two groups of transactions and ads that are linked to two distinct authors. Within each group, there is a one-to-one mapping between each transaction and ad; for example, transaction t_i is linked to ad a_i for $i = 1, 2, \ldots, 6$. In this way, w_1 is the SA wallet for these six ads.

2.6.2 Grouping by Persistent Bitcoin Identities

Within the set of input wallet addresses for transactions $t \in T$, any w that repeatedly uses the same wallet address for multiple transactions t is a `persistent bitcoin identity` (PBI). We further reduce this set of PBI's by only keeping those where at least one t is adjacent to exactly one a, and that ad a is also adjacent to only that t – i.e., they are an 'exact match.' When an exact match is found, we consider the PBI to be the owner of the matching ad. Figure 2.6 shows a hypothetical example of wallet address w_1 as a PBI for ad a_{exact}, as there is a one-to-one mapping between t_1 and a_{exact}.

Figure 2.6. Persistent Bitcoin Identities



For any remaining transactions that are not exact matches, e.g., multiple a's are linked to the same t, we use the author classifier (Section 2.4) to find the most likely true link. In our hypothetical example in Figure 2.6, t_2 fits that category, linked to both a_2 and a_3. For each such a_i that is linked to the transaction t, we run our binary author model on each pair a_{exact}, a_i. The ad a_i that belongs to the a_{exact}, a_i pairing with the highest probability of being written by the same author is selected as the matching ad.

Ad Type	**Adult Jobs**	**Body Rubs**	**Datelines**	**Escorts**	**Fetish**	**Male Escorts**	**Strippers**	**Transsexual Escorts**
Unique Ads	14,143	83,158	5,443	555,394	14,227	27,638	6,245	35,027
Unique Postings	46,160	382,843	27,495	1,805,174	43,166	55,351	16,881	159,778
Avg Num Locations	1.10	1.04	1.74	1.02	1.10	1.04	1.35	1.05

Ad Price	**Adult Jobs**	**Body Rubs**	**Datelines**	**Escorts**	**Fetish**	**Male Escorts**	**Strippers**	**Transsexual Escorts**	**Total**
Free	12,553	60,704	3,732	447,319	11,729	24,583	4,095	24,976	589,961
$1-5	1,029	3,991	813	13,703	1,660	2,196	1,392	5,967	30,751
$5-20	393	7,825	437	71,622	549	557	492	2,682	84,557
$20-100	157	7,331	161	16,987	269	194	261	1,250	26,610
>$100	11	3,307	298	5,763	20	8	5	152	9,564

Table 2.3. Distribution of Ads by Category during 4-week Case Study

2.7 Case Study

To validate our methods, we placed 33 Backpage ads – 32 paid, and one free – and used these as ground truth. In accordance with IRB requirements, all of the ads were placeholder ads, rather than realistic mockups. Eleven of these were paid for using our personal `1Ejb3` persistent wallet address. The remainder were paid for using the payment processor Paxful. The ads were placed from Dec 12, 2016 until Dec 24, 2016. The price of the ads ranged from $2-$20. For all eleven `1Ejb3` ads, the main contributor of the cost was bumps; for Paxful, the main contributor was location and sponsor (see Section 2.7.1 for more details). 87% (29) of the ads were in Escorts category. The ads were posted in 27 distinct US states and regions.

2.7.1 Price Reconstruction

To reconstruct the price of an ad we reverse engineer the exact algorithm used by Backpage. The price algorithm is highly dynamic, and required systematic experimentation to rebuild. To that end, we placed several ads ourselves, studied the HTML of the price payment page, and reproduced the algorithm (see Algorithm 1). To calculate the price of an ad, we need to know how often an ad was 'bumped' (e.g., appeared on the main listing page exactly an hour later), how often an ad was 'reposted' (e.g., appeared on the main listings page in the same hour every day for X number of days, where the number of days must be in 4-day increments), how many weeks an ad was 'sponsored' (e.g., appeared in a thumbnail highlighted on the side of the main listings page as well as in its normal position on the body of the listings page) and how many locations to which the ad was posted.

In order to collect this information, we wrote a scraper that scraped, every hour, all of the listings pages for every region (totaling 67) in the United States, for every adult entertainment sub-category. Backpage includes a minute-granularity timestamp indicating when an ad was posted; we extracted this timestamp during the scrape to determine the first appearance of an ad, and subsequent bump, repost and sponsor appearances. We ran our scraper for 4 weeks, shutting it down on January 9th when Backpage took down its adult entertainment section. In parallel, we also ran a separate scraper that collected the pricing information for each of the 67 regions, as the cost of a bump, repost, sponsor or location varied depending on the region to which it was posted. Backpage has one main anti-scraping policy that we had to work around: after a certain number of hits from a particular IP address, there is a forced timeout. In order to avoid this, we set up hundreds of HTTP proxy tunnels. For each page we scraped, we randomly selected and used one of those proxies.

Based on prior scrapes, we observed that for all of the adult entertainment sub-categories except for Escorts, the price stayed constant from week to week; so we ran

Ad Type	No. Correct Posting Pattern	No. Correct Price
Non-sponsor	24 of 25	24 of 25
Sponsor	8 of 8	0 of 8
All	32 of 33	24 of 33

Table 2.4. Price calculation correctness.

our pricing scraper on the non-Escorts sub-categories across all 67 regions four times, once for each week of the scrape. We observed more variability in the Escort pricing, and therefore scraped that sub-category's pricing once a day. We also noted that the sponsor pricing was significantly more variable than either bump, repost or location; the pricing followed what seemed to be a surge pattern where the prices varied every 15 minutes. We did not have the computing capacity to run our pricing scraper on all 67 regions every 15 minutes for four weeks, so instead we ran the pricing scraper at this rate for just one region, Los Angeles. Even scraping every 15 minutes did not allow us to reconstruct the price of our sponsored ads in Los Angeles with complete accuracy. In a previous experiment, we had scraped the pricing for Los Angeles as often as possible for one day (about every 8 minutes) and noted one instance where the sponsor price changed in as little as 10 minutes. Because this was rare, we elected during the 4-week scrape to reserve computing resources for other tasks.

Finally, we ran one last scraper that collected the first page of the main listings page for each region, for each adult entertainment sub-category, once a day, in order to collect the set of sponsored ads.

We found that for all non-sponsor ground truth ads, we correctly calculated the exact price for 24 of 25 total. For the wrong ad (paid for using Paxful), the hourly scraper missed one bump that happened to occur right on the hour, so the price was calculated incorrectly. For the eight sponsor ground truth ads, the posting pattern was correctly extracted, but the price was incorrect due to the high variability of sponsor pricing, with predicted prices varying within +/-5% of the true price. As a result, we decided to not include any sponsor ads in the rest of the study, leaving 95% (143,908 of 151,482) of the paid ads available for our analysis. None of the 11 ads placed using our personal `1Ejb3` wallet address were sponsor ads; eight of the 21 ads placed using Paxful were sponsor ads.

2.7.2 Linking Backpage Ads to Bitcoin Transactions

Before attempting to group the ads by true owner using our two methods, we first had to link the scraped, paid ads to the set of transactions T (as defined in section 2.5). To that end, we:

1. Constructed the w and t vertices and edges using the blockchain dataset

Algorithm 1 Price Recreation

```
procedure DETERMINE SERVICES
    ads ← sort(occurrences of this ad)
    bumps ← 0
    reposts ← 0
    sponsorWeeks ← number of weeks ad was sponsored
    lastAd ← ads.pop
    for ad in ads do:
        if ad.hour = lastAd.hour + 1 then
            bumps ← bumps + 1
        else if ad.day = lastAd.day + 1 and ad.hour = lastAd.hour then
            reposts ← reposts + 1
        else
            call Reconstruct Price
        end if
        lastAd ← ad
    end for
end procedure
procedure RECONSTRUCT PRICE
    priceInfo ← price info for each location at time of ad
    totalPrice ← 0
    for price in priceInfo do:
        if priceInfo.size > 1 then
            totalPrice ← totalPrice + price.base
        end if
        totalPrice ← totalPrice + (price.getBump * bumps)
        totalPrice ← totalPrice + (price.getRepost * reposts)
        totalPrice ← totalPrice + (price.getSponsor * sponsorWeeks)
    end for
    return totalPrice
    if not ads.empty then
        goto Determine Services
    end if
```

Payment Type	No. EM	No. EM is true ad	No. MM	No. MM contain true ad
`1Ejb3`	7 of 11	7 of 7	4 of 11	4 of 4
Paxful	1 of 8	1 of 1	7 of 8	6 of 7
All	8 of 19	8 of 8	11 of 19	10 of 11

Table 2.5. Transaction to ad linking correctness (Exact Match - EM, Multiple Match - MM).

2. Construct the a vertices using the Backpage scrape dataset

3. Constructed an edge between t and a if their timestamps were within one minute of each other (using the mempool timestamps dataset), and if the ad a's predicted cost was within 2% of one of t's GoCoin output values in US Dollars

Of the 11 GoCoin ground truth transactions processed with our personal `1Ejb3` wallet address, all satisfied the requirements to belong in set T. Seven were an exact match for the correct ground truth ad. All four of the remaining transactions matched to two ads, one of which was the correct ground truth ad. Of the 13 non-sponsor GoCoin ground truth transactions processed with Paxful, eight satisfied the requirements to belong in set T (the other five had multiple input wallet addresses). Of those eight, one was an exact match for the correct ground truth ad. Six of seven of the remaining transactions matched to multiple ads, one of which was the correct ground truth ad. One transaction matched to multiple ads, none of which was the correct ad. This occurred because a delay in Backpage caused the ad to appear with an initial timestamp 2 hours after the actual time in which the ad was placed.

Overall, of the 67,117 transactions in the set T found during the course of the 4-week study, 6,402 were exact matches to one ad, 58,657 matched multiple ads, and 2,058 were a match to one ad, where that ad matched to multiple transactions.

2.7.3 Results

2.7.3.1 Using Shared Authors

Using the methodology in Section 2.5.1, we constructed the graph G and subsequently the subgraph G'. By applying the four criteria in Section 2.5.1, we found 29 SA wallets. Each of the SA wallets mapped to multiple ads with different authors (see the Appendix for detailed tables on each wallet, and various features of author).

For each of the SA wallets, we used area codes on phone numbers, locations, and Jaccard similarity in order to verify whether any of those authors we had found were actually linked. Specifically, we calculated whether each pair of authors posted ads to a common location, or had at least one ad title or ad body with Jaccard similarity greater than some threshold.

Payment Type	Transaction	No. Ads matched
1Ejb3	c036d56	2
1Ejb3	5001910	2
1Ejb3	22dfe5b	2
1Ejb3	47612d0	2
Paxful	422ab70	6
Paxful	8cf7510	12
Paxful	428f494	13
Paxful	7e9fefc	17
Paxful	188552c	19
Paxful	d692422	21
Paxful	631f99a	29

Table 2.6. Multiple match transaction to No. ads matched.

Wallet	Total Pairs Authors	Location	Jac > 0.2	Jac > 0.3	Jac > 0.5	Jac > 0.8	Jac > 0.9
1A3Bj	28	5	15	4	4	3	3
1Abgk	1	0	1	1	0	0	0
1ASPo	1	1	1	1	1	1	1
1BT6w	36	8	32	21	15	10	3
1D1di	3	1	1	0	0	0	0
1E4RK	36	4	16	6	4	2	1
1Gsuis	10	4	9	2	0	0	0
1Hre7	1	0	1	1	1	0	0
1Kh3x	3	1	1	0	0	0	0
1KpyX	1	0	1	0	0	0	0
1KtCW	1	1	1	0	0	0	0
1Kyoc	55	10	32	10	6	3	1
1LetZ	45	11	38	25	20	10	2
1LYEQ	136	46	75	13	3	1	1
1MGDy	1	0	0	0	0	0	0
1MheR	15	1	15	10	5	0	0
1Mufv	1	1	1	1	1	1	1
1N7V4	28	7	19	7	5	3	1
1P7n4	3	1	3	2	2	1	1
1PesE	36	2	21	21	18	10	3
1yVFE	1	0	1	1	1	1	0
12Xis7	6	0	6	1	1	0	0
14FCt	78	15	65	40	16	5	4
14XUU	6	3	6	3	1	1	1
16iD4	21	7	15	2	1	1	1
17idT	3	0	1	0	0	0	0
18tTg	15	7	13	5	2	2	2
194iD	10	1	7	1	0	0	0
198xk	3	0	1	1	1	0	0

Table 2.7. Number of Pairs of Authors Linked within SA Wallet using location and Jaccard

SA Wallet	Linked Authors	Demographics	Locations
1ASPo	110831, 110828	japanese & taiwanese girls, 'outcall only'	SF Bay area
1KtCW	37140, 39230	latina, asian & western girls	Los Angeles
1Mufv	110830, 110828	japanese girls	SF Bay area

Table 2.8. SA Wallets with Zero FP by Location and Zero FP by Jaccard threshold 0.9

Table 2.7 shows these numbers for each SA wallet. Of the 29 wallets, three have all 100% of pairs of authors post ads to a common location. Nine have between 25%-50% of pairs of authors linked by location, nine have between 0%-25% of pairs of authors linked by location, and eight do not have any pairs of authors linked by location. Looking at the most stringent Jaccard metric with threshold of 0.9, two have all 100% of pairs of authors linked by Jaccard. One has between 25%-50% of pairs of authors linked by Jaccard, 12 have between 0%-25% of pairs of authors linked by Jaccard, and 13 do not have any pairs of authors linked by Jaccard.

We tested the effectiveness of the Jaccard metric by removing the author information on a subset of linked SA wallets, to see whether we found the true, known links between hard identifiers that were shared by the same author (e.g., if author 125 is made up of phone numbers X and Y, do we find that using the Jaccard metric we label ads from X and from Y as being linked by Jaccard). We found this to consistently be the case for Jaccard with threshold of 0.8. In order to be as strict as possible in our assessment, from this point on we focus our assessment on the Jaccard metric with threshold of 0.9.

Of the three SA wallets with zero false positives by location link, and the two SA wallets with zero false positives by Jaccard with threshold of 0.9 link, there are three total distinct wallets: `1ASPo`, `1KtCW` and `1Mufv` (see Table 2.8). Each of these SA wallets link together two authors, where the ads posted by the linked authors within each wallet advertise escorts with the same demographics, and to the same locations.

There is only one SA wallet, `1P7n4`, with between 25%-50% of pairs of authors linked by Jaccard. This wallet groups together three authors, two of which are linked by Jaccard. Both of these two authors (66852, 66789) advertise young asian girls; one author advertises in New Jersey, and the other in New Jersey and New York.

There are 12 authors that have between 0%-25% of pairs of authors linked by Jaccard (see Table 2.9). `1A3Bj` groups together eight authors, three of which are linked by Jaccard. For this wallet, we see that the escorts advertised share the same demographics, but are advertised across the country (in NY and CA). `1BT6w` (which groups together nine authors, four of which are linked by Jaccard), `1E4RK` (groups together nine authors, two of which are linked by Jaccard), `1LetZ` (groups together 10 authors, two of which are linked by Jaccard), `1LYEQ` (groups together 18 authors, two of which are linked by Jaccard), `1N7V4` (groups together eight authors, two of which are linked

SA Wallet	Linked Authors	Demographics	Locations
1A3Bj	36660, 36741, 37140	latin & asian girls & boys	NY, Los Angeles
1BT6w	85388, 110828, 110829, 110827	japanese, korean & taiwanese girls	SF Bay area & TX
1E4RK	24482, 10317	japanese & korean girls	CO, IL
1Kyoc	110827, 110828	asian girls	SF Bay area
1LetZ	24482, 10317	japanese & korean girls	CO, IL
1LYEQ	36660, 36741, 37140	latin & asian girls & boys	NY, Los Angeles
1N7V4	24482, 10317	japanese & korean girls	CO, IL
1PesE	118111, 24482, 14950, 10317	asian girls	TX, DC, IL, CO
14FCt	97790, 138909, 138908	asian girls	OR, WA
14FCt	118112, 10317	asian girls	TX, CO
14XUU	138909, 138908	asian girls	WA
16iD4	36660, 36741	latin & asian girls & boys	NY, Los Angeles
18tTg	41721, 21270	asian & latina girls	CA, GA, NY, FL, IL, CA, MA, LA
18tTg	72789, 74000	asian girls	NY

Table 2.9. SA Wallets with between 0%-25% of pairs of authors linked by Jaccard threshold 0.9

by Jaccard), `1PesE` (groups together nine authors, four of which are linked by Jaccard), `14FCt` (groups together 13 authors, with two sets of authors linked by Jaccard: one of size three and one of size two), `16iD4` (groups together seven authors, two of which are linked by Jaccard) and `18tTg` (groups together five authors, with two sets of authors linked by Jaccard: both of size two) all similarly advertise escorts with shared demographics, and in multiple locations across the country. `1Kyoc` groups together 11 authors, two of which are linked by Jaccard; `14XUU` groups together four authors, two of which are linked by Jaccard.

2.7.3.2 Using Persistent Bitcoin Identities

Using the criteria in Section 2.5.2, we found 114 PBI wallets with one exact match; we found 55 PBI wallets with multiple exact matches. As described in section 2.5.2, for any remaining transactions that are not exact matches, e.g., multiple ads are linked to the same t, we used the author classifier to find the most likely true link. See the Appendix for detailed tables on each wallet. For the single exact match wallets, the first row shows the author and other data of the exact match transaction, and the remaining show the same for the ad found to most likely be the true ad of the non-exact match transactions. For the multiple exact match wallets, we only include the author and other data of the exact match transactions.

For each of the single exact match PBI wallets, we used the author, area codes on phone numbers and locations in order to verify whether each of the non-exact match ad transactions appear to truly be from the same author as the exact match transaction. Specifically, we calculated how many of the non-exact match transaction ads had the same author, area code, and location as the exact match ad. We do not assess using Jaccard, as that is one of the features used in our classifier. Table 2.10 shows these

numbers for each single exact match PBI wallet. Of the 114 wallets, the majority - 86 - do not have any link between the non-exact match transaction ads and the exact match transaction ad.

When looking at the remaining 28 wallets, we observe that ten of them have at least two of the three metrics (author, area code and location) linking a subset of their non-exact match transaction ads to the exact match transaction ad: `16Zva`, `1N7Af`, `189Bu`, `1L8rv`, `1LSju`, `1NbRn`, `1LDTv`, `168GD`, `1P6DB` and `1H5`. `16Zva` groups together three authors. All three authors (37773, 113573, 36786) advertise girls in California. `1N7Af` has one author (45994), advertising young, asian girls in New York and Louisiana. `189Bu` also has one author (56512), advertising girls from Costa Rica, in Minnesota.

`1L8rv` also has one author (59987), advertising asian girls in Missouri. `1LSju` groups together three authors. All three authors (104638, 42837, 43145) advertise young black and latina girls in California. `1NbRn` groups together three authors (44282, 40532, 11185). These three authors advertise young, black girls in New Hampshire, California, Massachusetts and Colorado. `1LDTv` groups together two authors. Both authors (64220, 63787) advertise girls in Nevada. `168GD` groups together six authors (86108, 9910, 9222, 12145, 87608, 86528). These authors advertise girls in Texas, New York, and New Jersey. `1P6DB` groups together three authors (86703, 123265, 121049). These three authors advertise black and latina girls in Florida and Texas. `1H5` has one author (10), advertising a single woman in Colorado.

For the multiple exact match PBI wallets, we first focused our assessment on just the exact matches. We evaluated the multiple exact match PBI wallets in a similar way to the SA wallets, calculating whether each pair of exact match transactions had the same author, shared an area code, posted to a common location, or had at least one ad title or ad body with Jaccard similarity greater than 0.9. Then, looking at just those PBI multiple exact match wallets with linked exact match transactions, we assess those wallets in the same way we did the single exact match PBI wallets, using the author, area codes on phone numbers and locations to measure a link between the exact match transaction ads and the non-exact match transaction ads. Table 2.11 shows these numbers assessing the exact matches for each multiple exact match PBI wallet. Of the 55 wallets, the majority - 38 - do not have any link between the exact match transactions.

When looking at the remaining 17 wallets, we observe that four of them have at least two of the four metrics (author, area code, location and Jaccard) linking a subset of their exact match transactions: `1BdD`, `1A6M`, `1MD` and `16qB`. `1BdD` groups together two authors. Both authors (36677, 39708) advertise young girls in California; one author also advertises in Nevada. `1A6M` groups together two authors (62667, 63073), both advertising girls in Nevada. `16qB` groups together two authors, both (8, 9) advertising a single woman in Illinois. `1MD` has one author (11429), advertising a single woman across multiple states. Our own wallet, `1Ejb3`, satisfies only the Jaccard metric.

Wallet	Total Non EM	Author	Area Code	Location	Wallet	Total Non EM	Author	Area Code	Location
1M58i	1	0	0	1	18FBc	2	0	0	0
1EyHa	1	0	0	0	15kcN	2	0	0	0
1AFN6	1	0	0	1	1L1Pd	1	0	0	0
1MZG3	1	0	0	0	1Jjm5	2	0	0	0
14eoU	1	0	0	0	1M4aa	13	0	0	1
16THW	1	0	0	1	19mbJ	1	0	0	0
14nrA	1	0	0	1	1573u	1	0	0	1
1Nzho	17	0	0	0	16CmR	1	0	0	0
1Ks4n	1	0	0	0	1BPCN	1	0	0	0
1EKEg	4	0	0	0	1N4FE	1	0	0	0
1JntX	2	0	0	0	15Ztx	1	0	0	0
1EChh	2	0	0	2	1DqxW	3	0	0	1
1Aah7m	4	0	0	0	1Fv7D	1	0	0	0
18Szy	5	0	0	0	1FpkQ	1	0	0	0
18SDy	1	0	0	0	166vQ	3	0	0	0
1DLkr	4	0	0	0	1JY63	2	0	0	0
1FhgN	1	0	0	0	1Amh5	1	0	0	0
1AnJA	1	0	0	0	168GD	28	0	1	5
1LGmB	1	0	0	1	14BJR	1	0	0	0
14LMur	1	0	0	0	1P6DB	7	0	1	2
1NBrV	1	0	0	0	1KFPo	1	0	0	0
1P6eT	2	0	0	0	1EerL	7	0	0	0
1BZ91	2	0	0	1	1JCre	1	0	0	0
1DFvc	1	0	0	0	133b7	3	0	0	0
1MAoG	1	0	0	0	1CDaj	1	0	0	0
12r6t	1	0	0	0	1Jsgm	1	0	0	0
158DE	1	0	0	1	16Syp	1	0	0	0
18j9y	1	0	0	1	13dKY	3	0	0	0
1K7rX	1	0	0	0	1LSju	5	2	2	4
13Ges	19	0	0	0	17Xoc	2	0	0	0
1KCJ5	4	0	0	0	1Je2z	4	0	0	0
1GSWt	1	0	0	0	1JgQK	1	0	0	0
1DaRLu	1	0	0	0	19TW2	1	0	0	0
16Zva	6	2	5	6	1LxPF	5	0	0	1
1Nuf8	3	0	0	0	1AZ6T	4	0	0	0
16Yyx	1	0	0	0	17dKq	1	0	0	0
1NbRn	9	0	1	1	1NJA5	5	0	0	0
1Lune	1	0	0	1	189T	1	0	0	0
1J1Cc	1	0	0	0	1Hjq	3	0	0	0
1N7Af	1	1	1	1	19S7	1	0	0	0
13gqd	1	0	0	0	19HaT	1	0	0	0
19cYw	3	0	0	0	1MXgv	1	0	0	0
16qR6	1	0	0	1	1789c	2	0	1	0
189Bu	1	1	1	1	1BTZ	1	0	0	0
1EKp8	6	0	0	0	1Gb3	1	0	0	0
13qSM	1	0	0	0	1DPE	1	0	0	0
1NVVV	1	0	0	0	14y1o	2	0	0	0
1L8rv	1	1	1	1	135S	1	0	0	0
14cax	2	0	0	0	14ywq	2	0	0	0
15ZAE	1	0	0	0	1ExUN	2	0	0	0
1Fvev	1	0	0	1	1GEDX	3	0	0	0
1EUAF	1	0	0	0	1DvN	1	0	0	0
1LDTv	7	0	1	1	1HLqs	1	0	0	0
1Hu8a	1	0	0	0	1J3t	3	0	0	0
1EF3p	1	0	0	0	1NmYX	1	0	0	0
1HEDb	1	0	0	0	1NT5	1	0	0	0
1JWm4	4	0	0	1	1H5	1	1	0	1

Table 2.10. Number of Transactions Linked to Exact Match within PBI single exact match Wallet using author, area code and location

Of those four wallets, two have other non-exact match transactions: 1A6M and 1MD. In both cases, all of the non-exact match transaction ads (6 and 3, respectively) were linked to the exact match transaction ads, by either author, area code or location, or some combination of all three. When looking at our own wallet, 1Ejb3, we observe that all four of the non-exact match transaction ads are correct (e.g., the ad chosen using our classifier to be the true match for the non-exact match transaction is in fact the true ad).

Wallet	Total Pairs Authors	Author	Area Code	Location	Jac > 0.9	Wallet	Total Pairs Authors	Author	Area Code	Location	Jac > 0.9
14psc	6	0	0	1	0	13jV	1	0	0	1	0
1LzN	3	0	0	3	0	1FUv	1	0	0	0	0
1H4h	1	0	0	0	0	1Fs4	3	0	0	0	0
1GWt	10	0	0	2	0	1K2H	1	0	0	0	0
1DCQ	1	0	0	0	0	15db	3	0	0	0	0
13DE	1	0	0	0	0	1BrD2	1	0	0	1	0
1CJy	1	0	0	0	0	197Z	3	0	0	0	0
1BBey	3	0	0	0	0	1AA	3	0	0	0	0
15Gd	1	0	0	0	0	1A6M	1	0	1	1	0
16pp	1	0	0	0	0	1P5Z	1	0	0	0	0
18zU	3	0	0	0	0	1PeV	1	0	0	1	0
1Fox	1	0	0	0	0	1KUT	3	0	0	0	0
1BQ	1	0	0	0	1	1JNQ	1	0	0	0	0
1KXo	1	0	0	0	0	1F3w	1	0	0	0	0
1Hyt	1	0	0	0	0	14Qo	1	0	0	0	0
17uj	1	0	0	0	0	1MD	6	6	6	6	6
1FUk	1	0	0	0	0	1DpT	1	0	0	0	0
1Nrs	1	0	0	0	0	1L6E	3	0	0	1	0
16L5	1	0	0	0	0	1MCS	1	0	0	0	0
1NYR	1	0	0	1	0	13vH	1	0	1	0	0
1AP1	6	0	0	2	0	1Lew	3	0	0	0	0
1K2J	1	0	0	0	0	1G2S	3	0	0	0	0
1Dja	1	0	0	0	0	12T9	1	0	0	1	0
1B6G	1	0	0	0	0	13Yo	1	0	0	0	0
1BdD	1	0	1	1	0	13af	1	0	0	0	0
18Vf	3	0	0	0	0	16qB	1	0	0	1	1
1Ng	1	0	0	0	0	1LYx	1	0	0	0	0
1Gvn	3	0	0	1	0	1Ejb3	21	0	0	0	21

Table 2.11. Number of Pairs of Authors Linked within MEM Wallet using author, area code and location for Exact Matches

2.7.3.3 Bitcoin-based Owners

The following two tables summarize all the Bitcoin-based owners that grouped together multiple authors, found using both the SA and the PBI methodologies. We observed that many of the SA wallets appeared to be linked to each other, based on the fact that they paid for ads written by the same author. Table 2.12 shows the final seven 'owner' identities we extracted when grouping together SA wallets in this way. On the upper end, Cluster 2 spent $32,358.55 on ads in the four week period; on the lower end, Cluster 6 spent $5,208 on ads in that same four week period.

Cluster ID	Wallets	No. Authors	No. Post Locations	No. Ads Posted	$ Spent
Cluster 1	1PesE, 14FCt, 1E4RK, 1LetZ, 1N7V4, 1yVFE	5	5	433	$18,720.50
Cluster 2	1BT6w, 1Kyoc, 1ASPo, 1Mufv	6	2	437	$32,358.55
Cluster 3	1A3Bj, 1LYEQ, 16iD4, 1KtCW	4	2	3,154	$15,812
Cluster 4	14FCt, 14XUU	3	2	132	$7,900
Cluster 5	1P7n4	2	2	55	$5,670
Cluster 6	18tTg	2	2	1128	$5,208

Table 2.12. Final SA Cluster Statistics by Jaccard > 0.9

Table 2.13 shows the final nine 'owner' identities we extracted from the PBI methodology. On the upper end, wallet 1LDTv spent $2,890 on ads in the four week period; on the lower end, wallet 16qB spent $16 on ads in that same four week period.

Wallet	No. Authors	No. Post Locations	No. Ads Posted	$ Spent
16Zva	3	1	5	$1,180
1LSju	3	1	27	$232
1NbRn	3	4	14	$154
1LDTv	2	1	6	$2,890
168GD	6	4	63	$1,697.15
1P6DB	3	2	18	$1,790
1BdD	2	2	5	$535
1A6M	2	1	9	$785
16qB	2	1	2	$16

Table 2.13. Final PBI Statistics

The owner identities found using the SA methodology spent more money across the four week period than those found using the PBI methodology. However, on average the SA owners spent less per ad than the PBI owners (about $52 per ad vs. $118 per ad). Both methodologies grouped on average three authors within each owner, and about two locations within each owner. On a whole, the new Bitcoin-based owner identities consistently expanded the set of locations; what previously was a single author in a single location becomes a network of authors (and hard identifiers) across multiple locations. In addition, authors that previously looked fairly small and financially limited suddenly become part of a much larger financial entity when linked to a different author that has much more capital (as judged by the average price of an ad they purchase).

2.8 Discussion

Both grouping by shared author and grouping by persistent bitcoin identity have an obvious limitation: false positives and negatives on the link between transactions and ads. An exact match transaction is not necessarily correct simply by virtue of being an exact match, and just because multiple ads all map to the same wallet does not mean those mappings are correct.

In particular, there are numerous reasons why a transaction might be an exact match with an ad, but that pairing still be incorrect. The post listings scraper may have scraped an ad right on an hour boundary; the ad might have been paid for using credit; the transaction might not be a payment for an ad posting but for purchase of credit; the transaction might not be a true GoCoin transaction; the transaction might be GoCoin but not a payment for Backpage.

It is also entirely possible that the exact match transactions we found are in fact correct. This is another place where the lack of ground truth becomes quite problematic; if the results do not look correct, it is difficult to tell whether that is because it is showing us something new, or because something is wrong.

There are several avenues of approach future work could take. We can work to disambiguate Backpage credit payments on the Bitcoin blockchain from Backpage ad payments by analyzing ads and credit payments we make ourselves. We can show our data to law enforcement officers and work together to build a ground truth set that we can then use to validate or reject the correctness of our exact match transactions. We can use existing Bitcoin clustering techniques to link our Paxful transactions to each other, and then our stylometry model to tie those ads that match the Paxful transactions to the ads that match the transactions made using our persistent Bitcoin wallet.

In general, finding more owners is key. The results of our case study indicate that even with a small increase in linkage across hard identifiers using Bitcoin and stylometry, we can potentially find critical information (e.g., connections between previously unconnected ads that indicate movement across multiple states/geographic locations, with multiple parties involved, both of which are strong indicators of trafficking) that could help our NGO and law enforcement partners in their mission to find and rescue trafficked humans. It is important to remember the immense size of this data - the fact that we can narrow down from hundreds of thousands of ads to find these connections is potentially enormously helpful to those law enforcement officers who have to read through so many ads during an investigation. This both saves these officers a substantial amount of time and also protects them from some of the psychological repercussions of analyzing this data.

There is also value in simply just linking ads to transactions, even in the case where a trafficker is not reusing the same wallet. Law enforcement could potentially subpoena information from Paxful or some other wallet service; going from a set of ads of interest to the set of matching Bitcoin transactions would make it possible to then get explicit personally identifying information from a wallet service. Some of our law enforcement collaborators have also stressed the value of having the Bitcoin transaction matched to a target ad when it comes to building a case for the court, after the alleged trafficker or pimp has been arrested. Our success in matching transactions to the correct ad for some of the PBIs, SAs, and our ground truth, is encouraging to this end.

It is worth noting that none of this work is stymied by the fact that Backpage has shut down their adult entertainment section. The vast majority of those ads shifted over to the dating section of their website, where ads are also paid for using Bitcoin. It is also worth noting that perpetrators have no choice but to continue to use Bitcoin even after our published work: the original move to Bitcoin was because of Backpage's response to Visa and MasterCard's decision to stop processing transactions for adult listings on Backpage. Perpetrators have no choice but to use the payment platform provided by the advertising company. Even if Backpage changes the virtual currency it accepts as payment, as long as that virtual currency is implemented with publicly accessible ledgers, our techniques will continue to work (barring the use and development of more sophisticated mixes).

2.9 Conclusion

In this chapter, we proposed an automated and scalable approach for identifying sex trafficking using multiple data sources. We developed a stylometry classifier and a Bitcoin transaction linking technique to group sex ads by owner. To the best of our knowledge, this is the first such work to attempt to link specific purchases to specific transactions on the Bitcoin blockchain. We evaluated our approach using real world ads scraped from Backpage, and demonstrated that our approach can group multiple ads by their real owners. We are currently collaborating with multiple NGOs and law enforcement officers to deploy our tools to help fight human trafficking.

Chapter 3

Cyber Black-Market Forums

3.1 Introduction

As technology evolves, abuse and cybercrime evolve with it. Much of this evolution takes place on underground forums that serve as both marketplaces for illicit goods and as forums for the exchange of ideas. Underground forums play a crucial role in increasing efficiency and promoting innovation in the cybercrime ecosystem. Cybercriminals rely on forums to establish trade relationships and to facilitate the exchange of illicit goods and services, such as the sale of stolen credit card numbers, compromised hosts, and online credentials. Analysis of these forums to extract this structured data can provide valuable insight into cybercrime but is a labor-intensive task, requiring an automated solution. In this chapter, we aim to develop and demonstrate automatic techniques for extracting such structured data from forums.

The rest of this chapter is organized as follows. Section 2 provides the necessary background for the rest of the chapter. Section 3 outlines the eight underground forums which we analyzed and used to evaluate our tools. Section 4 describes the methodology for building our type-of-post classifier, product extractor, and price extractor, covering ground truth labeling, the models we built, and validation results. In sections 5 and 6, we analyze the results of using our tools on our underground forum data, in particular using two case studies to show how to use these building blocks to carry out specific forum analysis tasks. We conclude with reiteration of key contributions and findings.

3.2 Related Work

3.2.1 Ecosystem Analysis

There has been a substantial amount of work done in analyzing the infrastructure, size and economies of cyber underground markets ([8] [83] [34] [57]), as well as in the categorization and quantification of the goods and services being bought, sold and traded on these markets ([18] [35]). The vast majority of this work to date has been manual, with the purpose of either exploring a particular black-market forum/set of forums, or using the findings from these manual studies in order to build a holistic picture of how the cyber black-market economy operates.

Multiple Forums Herley and Florencio [34] and Moore et al. [57] explored the results of a sample of manual findings in an effort to provide a general economic understanding of cyber black-market forums. Based on data collected by [83] and [30], the authors of [34] posited that most of these black-market forums are classic examples of lemon markets, where the presence of "rippers" who cheat other participants in the market make it difficult for these underground markets to operate effectively in the long term. The authors of [57] focused on the class of criminal involved in the underground market, discussing the economic impact of criminals that operate trans-nationally, and can commit crime on a global and industrial scale. They tied into their analysis the unique way in which the victims of online crime often become participants in this criminal activity, in that their machines become compromised and can be used by botnets.

Afroz et al. [8] analyzed what aspects distinguish a sustainable forum from those that fail. They studied 5 different forums—AntiChat, BadHackerZ, BlackhatWorld, Carders and L33tCrew—ranging in time from the year 2002 to 2010. From the scrapes of these forums, they extracted a variety of data, included the number and content of public posts and private messages, the total number of users, and the number of lurkers (i.e., users who post neither public nor private messages). By manually studying the forum structure and content of posts/private messages, the authors learned several insights about the infrastructure of these black-market forums. These insights include: the main topics discussed in each forum, (e.g., AntiChat covers a broad array of cybercrime topics such as password cracking, email spam and stolen online credentials, whereas BadhackerZ specializes in the exchange of copies of pirated movies), the number of active users, and membership criteria and ranking. Ultimately, they found that successful black-market forums have the following five qualities: 1) they have easy/cheap community monitoring, 2) they show a moderate increase in new members, 3) they do not witness reduced connectivity as the network size increases, 4) they limit privileged access, and 5) they enforce bans/fines on offending members.

Single Forum Christin [18] performed a full-scale measurement study focusing specifically on Silk Road, an anonymous cyber black-market that operates as a Tor hidden service. Looking at data spanning eight months between the end of 2011 and 2012, they built a detailed picture of what goods were sold on the Silk Road, and the amount of revenue made by both the sellers and the operators of the market. [18] collected this data by crawling all the webpages on the site labeled "item", "user" (i.e., seller user information) and "category". After processing this data, they found approximately 220 distinct categories of goods offered, ranging from digital goods to various different narcotics. For the most part, they discovered that the market is used for controlled substances and narcotics, with most items sold available for at most 3 weeks. Most sellers only maintained a persistent identity on the site for about three months, although 112 sellers were present throughout the full 8 months of data they collected and analyzed. Additionally, by manually extracting the pricing information of various products, averaging this price over 29 days, then multiplying this average by the number of feedback responses gathered about the particular product, [18] roughly estimated the average sales on Silk Road to be $1.22 million per month.

Franklin et al. [30] studied an unnamed black-market public channel, commonly found on IRC (Internet Relay Chat) networks. By connecting to a particular channel on different IRC networks and logging all subsequent public messages, they were able to collect over 13 million messages over a 7 month period in 2006. Most of their work consisted of manual extraction and analysis of information—e.g., manually labeling a 3,789 line subset of the data and determining what proportion of messages contain sensitive information (such as SSNs and bank account numbers).

3.2.2 Classification for Black-Markets

Little work has been done on the automatic analysis and classification of underground market forum threads. Typically, work that has included an automation component ([30] [36] [58]) has focused specifically on one or a few underground markets. Rather than building tools to analyze and classify threads from markets in general, most research has focused on gathering a holistic view of the goods and services from particular market/markets, with the goal that this analysis will help provide some illumination of the mechanisms of black-markets in general. Our study is the first to create automated extraction techniques for conducting large-scale analyses of the products and pricing of goods offered on underground forums. Previous work used structured information (e.g., social graph, timestamps, usernames) ([58, 76, 32, 91]), handcrafted regular expressions [30] and manual annotations of a small set of posts to understand products and pricing [35]. Our tools can analyze unstructured texts in large scale with little manual effort.

Classifying Ads Franklin et al. [30] developed a fully automated technique to classify ads as sale ads or want ads. Using hand labeled training data from an unnamed black-market public channel, they trained two binary text classifiers. They achieved a precision of 68.4% and a recall of 42.6% for the sales ad classifier, and a precision of 57.1% and recall of 38.1% for the want ad classifier.

Motoyama et al. [58] analyzed a broader range of black-market forums, six in total: BlackHatWorld, Carders, HackSector, HackE1ite, Freehack, and L33tCrew. They developed a pseudo-automated rule based technique to classify ads into 18 hand-defined categories, including merchandise, banking information, drugs, mailing and "dropping" services, in which a person is hired to deliver illicit goods. The authors wrote over 500 regular expressions that grouped the data according to their 18 classes. These categories were built based on domain knowledge of the illicit goods, and by randomly sampling thread titles. With this method, they were able to classify 87% of the 14,430 threads from the Carders forum, and 77% of the 31,923 threads from the L33tCrew forum. For the Carders forum, they found that the top 5 most common classes were: payments, game-related, credit cards, accounts and merchandise. L33tCrew had the same classes except for replacing merchandie with software/keys.

Classifying Stolen Data Franklin et al. [30] also developed a pseudo automated technique to label messages and message content by type and frequency. They wrote and used regular expressions to identify messages containing credit card information or SSNs, extract the frequency at which common commands were issued, and learn the rate of new and repeated credit card and SSN arrivals into the market. Holz et al. [36] investigated keylogger-based stealing of online credentials. By setting up honeypots and spamtraps, [36] collected various keyloggers and harvested their data just as an attacker would have. Their classification task consisted of categorizing the stolen credentials into 4 distinct groups: banking, credit card, emails and social networking credentials. To do so, they built a unique model for each individual service provider (e.g., email provider, or banking provider) that kept track of which input fields in the HTTP POST credentials request contained the desired sensitive information. Using these models, they could then automatically extract from each piece of data both the credentials and the classification of the credentials. In total, they found 10,775 unique bank account credentials, 5,682 valid credit card numbers, 149,458 email passwords and 78,359 social networking credentials.

3.2.3 NLP Tools

Forum analysis with NLP tools. NLP techniques have proven useful for answering a range of scientific questions in various disciplines including the humanities [10] and the social sciences [60]. However, there has been relatively little work in specifically

Forum	Source	Primary Language	Date covered	Threads (Commerce)	Users
Blackhat World	Complete Dump	English	Oct 2005–Mar 2008	7270 (2.29%)	8,718
Darkode	Partial Scrape	English	Mar 2008–Mar 2013	7418 (27.94%)	231
Hack Forums	Partial Scrape	English	May 2008–Apr 2015	52,649 (97.34%)	12,011
Hell	Partial Scrape	English	Feb 2015–Jul 2015	1,120 (22.59%)	475
Nulled	Complete Dump	English	Nov 2012–May 2016	121,499 (32.81%)	599,085
Antichat	Complete Dump	Russian	May 2002–Jun 2010	201,390 (25.82%)	41,036
Carders	Complete Dump	German	Feb 2009–Dec 2010	52,188 (38.72%)	8,425
L33tCrew	Complete Dump	German	May 2007–Nov 2009	120,560 (30.83%)	18,834

Table 3.1. General properties of the forums considered.

applying NLP techniques to web forums ([44, 42]). Because of the high degree of domain dependence of NLP techniques [19], most out-of-the-box tools (like part-of-speech taggers or parsers) have various deficiencies in this setting, and in any case do not directly provide the information about forum posts in which we have the most interest.

NLP methodology. The problems we consider in this work differ from those in past NLP efforts on forum analysis ([44, 53, 89]). Our tasks broadly fall into the category of slot-filling information extraction tasks ([31, 81]), where the goal is to populate a set of pre-specified fields based on the information in the text. However, much of the recent work on information extraction in the NLP literature has aimed to extract a very broad set of relations for open-domain text [25], as opposed to focusing on domain-specific or ontology-specific methods [63]. The various kinds of information we consider (transaction type, products, prices) each necessitate different techniques: some tasks are formulated as classification problems with various structures, and our product extraction task is similar to named entity recognition [84] or entity detection [59]. We use a variety of supervised machine learning methods in this work, drawing on well-established conventional wisdom about what features prove most effective for each of our tasks.

3.3 Forum Datasets

We consider eight underground forums (Table 3.1): Blackhat World, Darkode, Hack Forums, Hell, Nulled, Antichat, Carders and L33tCrew. We collected the forum data in two ways: partial scraping (Darkode, Hack Forums, Hell) and complete publicly leaked database dumps that contain all public posts and metadata prior to the leak (Blackhat World, Nulled, Antichat, Carders and L33tCrew).

Blackhat World Blackhat World focuses on blackhat search engine optimization (SEO) techniques. The forum started in October, 2005 and is still active, although it

has changed in character over the past decade.

Darkode Darkode focused on cybercriminal wares, including exploit kits, spam services, ransomware programs, and stealthy botnets. We focused our attention on the four subforums that contained substantial amounts of commerce, ignoring twenty-eight other subforums unrelated to commerce. This forum was taken down in July of 2015 by a joint multinational law enforcement effort [61].

Hack Forums Hack Forums covers a wide range of mostly cybersecurity-related blackhat (and non-cybercrime topics), such as crypters (software used to hide viruses), keyloggers, server "stress-testing" (denial-of-service flooding) and hacking tools. The forum started in 2007 and is still active. For our analysis, we focus on the subforums in Hack Forums related to buy, sell, and currency exchange.

Hell Hell was an underground forum hosted as a Tor Hidden Service. It focused on credit card fraud, hacking, and data breaches. Hell made headlines when a hacker on the forum dumped the personal details of 4 million users of Adult Friend Finder, a dating website. The forum was shut down in July 2015 but relaunched in January 2016.

Nulled Nulled advertises itself as a "cracking community" specializing in leaks and tools for data breach. The forum was hacked on May 2016 and the full database of the forum was released publicly.

Non-English Forums We analyzed three non-English forums: Antichat, Carders and L33tCrew. Carders and L33tCrew were German-language forums that specialized in stolen credit cards and other financial accounts [8]. Both of the forums were leaked and closed. Our data spans the entire lifetime of the forums. Antichat is a Russian-language forum. Unlike the other forums, Antichat does not specialize on a single topic but rather covers a broad array of underground cybercrime topics such as password cracking, stolen online credentials, email spam and SEO [8].

3.4 Automated Processing

For each post appearing in a forum, we extract three properties —the type of transaction, the product, and its price—not explicitly marked. We take a supervised learning approach, labeling a small proportion of the data with ground truth and using those annotations to train a tool to label the rest. We divide the task of extracting all of this information into three sub-tasks. In every case, the input to the tool is a single

post, while output structure varies by task. In this section we describe the development of our tools, and results of evaluations that assess their effectiveness.

3.4.1 Type-of-Post Classification

Different forums use different conventions to mark different types of posts. For example, Darkode and Hack Forums have dedicated subforums for buy, sell and trade posts; on Carders and L33tCrew, buy posts start with "[S]" ("suche" means "seeking", i.e., buying) and sell posts start with "[B]" ("biete" means offering). The rest of the forums do not have any explicit tagging to mark the type of a post. Identifying the commerce section of a forum will significantly reduce the workload of an analyst, because fewer than 40% of the posts are related to commerce on the majority of the forums.[1]

The type-of-post classifier detects whether a post concerns buying or selling a product, exchanging currency, or none of these (e.g., an admin post). Due to the lack of ground-truth data on the non-English forums, the classifier only detects buy and sell posts on those forums. We use a variety of token- and character-level features robust across languages and domains.

3.4.1.1 Labeling Ground Truth.

To build a ground-truth dataset for the type-of-post classifier, we strip out the information that explicitly indicates the posting type. For the non-English forums (Antichat, Carders and L33tCrew), we consulted one German and one Russian native speaker to confirm the accuracy of the labels. For Antichat, we look for the words related to trade, buy or sell. For example, "prodayu" is the first person singular present tense of *to sell*, meaning "I am selling," and is often used in posts to offer a product for sale. By identifying threads with these words we constructed a training set, with one of three confidence levels assigned to each thread based on the words present—a confidence level of 3 indicates 100% confidence in the labeling, a level of 2 indicates less than 75% confidence, and a level of 1 indicates less than 50% confidence. Table 3.2 shows the final dataset size for each forum.

3.4.1.2 Models.

We consider two models for type-of-post classification.

Most-Frequent Label (MFL) This model returns the most frequent label in the training set. This approach can appear to do much better than 50% in some cases

[1]In our dataset, one exception is Hack Forums where 97% of the posts are commerce related because we only scraped the commerce section of the forum.

because of natural imbalances in the number of "buy", "sell" , "currency exchange" and "other" posts in a forum (see Table 3.2).

Support Vector Machine (SVM) This model uses text-based features: word unigrams, word bigrams, and character n-grams of lengths 1 to 6. We train the SVM by coordinate descent on the primal form of the objective [26] with ℓ_2-regularization. We also considered a range of other features: part-of-speech labels, parse dependencies, text from replies to the initial post, the length of the post, and rank and reputation of the user who authored the initial post. None of these additions appreciably improved performance, and so we do not include them in the final classifier.

3.4.1.3 Validation Results.

We assessed our classifier both within a forum and across forums. The first gives a direct measure of performance on the task we trained the classifier to do. The second gives a measurement of how well the classifier would generalize to an unseen forum in the same language. For Antichat, in addition to doing the standard evaluation, we also considered performance when using only the threads with a high confidence annotation level (level 3). In within-forum evaluations, we split the data 80% / 20% to form training and test sets. In cross-forum evaluations, we used 100% of the data.

English For Darkode, the buy vs. sell classifier is effective, achieving 98.2% accuracy overall, and 90.0% on the less common class. Our classifier is similarly effective on Hack Forums sell vs. currency (98.29% overall / 96.95% on the least common class) and Nulled buy vs. sell vs. other (95.27% overall / 85.34% on the least common class). When combining randomly sampled data from Darkode, Hack Forums and Nulled to get a dataset balanced between all four classes, we see uniformly high performance on all classes and 95.69% accuracy overall.

While Blackhat World and Hell are too small for within-forum evaluation, we can use the entire dataset as a test set to perform cross-forum evaluation. When training on Darkode and testing on Blackhat World, we see a performance drop relative to the within-forum evaluation on Darkode, but we achieve accuracy still well above the MFL baseline. The same holds when training on Nulled and testing on Blackhat World, Hell or Darkode. These results all indicate that the classifier generalizes well and analysts could use it in the future on other, completely unlabeled forums.

Non-English On both German-language forums (Carders and L33tcrew) we see high performance both within-forum and across-forum. Performance when evaluating on Carders runs consistently lower, probably because of the more even distribution of buy and sell threads. For Antichat, we also see high performance within-forum, but are

Forum	# Buy	# Sell	# Curr	# Other
Blackhat World	22	115	–	1
Darkode	1,150	205	14	1
Hack Forums	165	14,393	33,067	–
Hell	44	42	–	14
Nulled	2,746	8,644	49	1,025
Carders	8,137	5,476	–	–
L33tcrew	8,486	4,717	–	–
Antichat (all)	13,529	25,368	–	–
Antichat (confidence=3)	10,129	18,965	–	–

Table 3.2. Number of labeled posts by class for type-of-post classification.

	Accuracy (%)			
Train/Test Forum	**Buy**	**Sell**	**Overall**	**MFL**
Darkode	99.6	90.0	98.2	85.4
Darkode/BHW	85.7	90.6	90.5	84.8
Carders	95.65	89.52	93.20	60.0
L33tcrew	97.36	92.32	95.57	57.0
Carders/L33tcrew	95.75	88.21	93.05	64.27
L33tcrew/Carders	93.81	83.82	89.79	59.77
Antichat (all)	95.15	98.23	97.16	65.3
Antichat (confidence=3)	98.91	99.97	99.60	65.2

Table 3.3. Classification accuracy on the buy vs. sell task. MFL refers to a baseline that simply returns the most frequent label in the training set. Note that the test sets for within-forum evaluation comprise 20% of the labeled data, while the test sets for across-forum evaluation are 100% of the labeled data in the target forum.

	Accuracy (%)					
Train/Test Forum	**Buy**	**Sell**	**Curr**	**Other**	**Overall**	**MFL**
Hack Forums	—	96.95	98.89	—	98.29	69.67
Nulled	89.42	98.28	—	85.34	95.27	59.53
Darkode + Hack Forums + Nulled	92.86	95.72	98.35	96.26	95.69	27.42
Nulled/BHW	77.27	93.04	—	—	90.51	84.8
Nulled/Darkode	86.50	96.14	—	—	87.96	85.4
Nulled/Hell	86.36	85.71	—	—	86.1	51.2

Table 3.4. Classification accuracy on the buy vs. sell vs. currency exchange vs. other task. Omitted entries indicate categories with too little ground-truth data of that type to robustly evaluate.

unable to evaluate across-forum performance because we do not have another Russian forum. Focusing on the high-confidence threads, we see even higher performance, as would be expected.

These results indicate the robustness of our feature-set to variation in language as well as forum, though to generalize to further languages would require additional labeled data.

3.4.1.4 Limitations.

We investigated why the additional features we considered did not improve the accuracy any further. We found two general issues. First, most core NLP systems like part-of-speech taggers and syntactic parsers target formal, well-formed, grammatical text. The data we consider strays far from that setting, and performance of those tools suffers accordingly, making them less informative. Second, as a thread continues, the topic will often drift and lose focus, becoming less related to the original transaction. This noise explains why using features of later posts in a thread did not improve performance.

3.4.2 Product Extraction

Here we look at extracting the actual product being bought or sold in a thread. Our system outputs a set of spans in the text, each of which marks an explicit mention of the product. From this, we can extract a set of string representations of the product(s) being bought or sold. This task proves both very useful for analyzing criminal marketplace activity but also quite difficult. One general challenge is that a single post can mention many product-like items distinct from the item actually for sale, such as an account to contact for purchasing. We address this kind of ambiguity by building a machine-learned product extractor, which uses features that consider syntax and surface word context.

3.4.2.1 Labeling Ground Truth.

To start, while the task output is multi-word spans that describe products, we find manually annotating such spans a difficult and time-consuming process. To understand the challenge, consider this example from Darkode:

a keylogger coded completely in ASM

The correct span in this case could be keylogger, a keylogger, or the complete phrase. Linguistically, the first of these is a noun, the second is a base noun phrase (NP), and the third is a complex NP. We thus avoid defining rules on where to place the

0-initiator4856

TITLE: [buy] Backconnect bot

BODY: Looking for a solid backconnect bot .
If you know of anyone who codes them please let me know

0-initiator6830

TITLE: Coder

BODY: Need sombody too mod DCIBot for me add the following :
Update Cmd
Autorun Obfuscator (Each autorun diffrent and fud)
Startup Mod (needs too work on W7/VISTA)
Pm .

Figure 3.1. Example post and annotations from Darkode, with one sentence per line. We underline annotated product tokens. The second exhibits our annotations of both the core product (*mod DCIBot*) and the method for obtaining that product (*sombody*).

boundaries, and instead annotate the word common to all of the options—the head of the noun phrase, in this example, keylogger. Doing so provides a clearer definition of the annotations, enabling consistent labeling. Using automatic syntactic parsers we can subsequently recover the larger span (described further in Section 3.4.2.3), though we define our annotations over raw tokens to avoid tying ourselves to error-prone parser output.

Note that, when appropriate, we annotate both the outcome and the means of delivery. Figure 3.1 shows an example: *DCIBot* is closest to the core product, but *sombody* and *mod* are critical to the process and so we annotate them as well. However, we do not annotate features of products (*Update Cmd* in Figure 3.1), generic product references (*this*), product mentions inside "vouches" (reviews from other users), or product mentions outside of the first and last 10 non-whitespace lines of each post.[2] We make our full annotation guide available.[3]

We developed this approach through a series of rounds, first to investigate options for annotation methodologies, then to train annotators without security expertise. We annotated training, development, and test sets in several forums:

- Darkode training set (630 posts with 3 annotators per post, 30 with 7 annotators per post)
- Darkode development set and test set (both 100 posts with 8 annotators per post)
- Hack Forums training set (728 posts, 3 annotators per post, 30 posts with 8 annotators per post)

[2]This reduces the annotation burden on the small number of posts (roughly 4% on Darkode) that are unusually long—these posts are also often outliers with few product references.

[3]`cs.berkeley.edu/~jkk/www2017-product-annotation-guide.pdf`

- Hack Forums test set (140 posts, 4 annotators per post)

We used the Fleiss Kappa measurement of inter-annotator agreement [29] and found that our annotations had "substantial agreement" ($\kappa = 0.65$).

We derived the final annotations used by taking a majority vote among annotators; i.e., for each token in question, if at least half of annotators (rounded up) annotate it, then we treat it as ground truth. Roughly 95% of posts in Darkode and Hack Forums contained products according to this annotation scheme. We additionally pre-processed the data using the tokenizer and the sentence-splitter from the Stanford CoreNLP toolkit [54].

3.4.2.2 Models.

We consider two models for product extraction. In each case, our models deal with noun phrases as the fundamental units of products. We generalize ground-truth noun phrases from our headword annotation according to the output of an automatic parser [16].

Noun-phrase classifier We train an SVM to classify each noun phrase in the post as either a product or not. We structure our features around a set of key words that we consider, namely the first, last, and head words of the noun phrase, as well as the syntactic parent of the noun phrase's head, and up to three words of context on each side of these words. For each of these words, we fire features on its identity, character n-grams it contains, part of speech, and dependency relation to its parent. This gives us a rich set of contextual features examining both the surface and syntactic context of the noun phrase in question. For example, in Figure 3.1, when considering the noun phrase *a solid backconnect bot*, we fire features like *parent=for* and *parent-previous=looking*, the latter of which provides a strong indicator that our noun phrase corresponds to what the poster seeks. Finally we also use features targeting the noun phrase's position in the post (based on line and word indices), capturing the intuition that posters often mention products in a post's title or early in the post body.

We train the SVM by subgradient descent on the primal form of the objective ([69, 49]). We use AdaGrad [23] to speed convergence in the presence of a large weight vector with heterogeneous feature types. We trained all product extractors in this section for 5 iterations with ℓ_1-regularization.

Post-level extractor Rather than making decisions about every noun phrase in a post, we can support some kinds of analysis by a more conservative product identification scheme. If all we want to do is identify the general product composition of a forum, then we do not need to identify all references to that product in the body of the post,

but might instead just identify an easy one in, say, the post title. Therefore, we also consider a post-level model, which tries to select one noun phrase out of a post as the most likely product being bought or sold. Structuring the prediction problem in this way naturally lets the model be more conservative in its extractions, and simplifies the task, since we can ignore highly ambiguous cases if the post includes a clear product mention. Put another way, doing so supplies a useful form of prior knowledge, namely that the post contains a single product as its focus.

We formulate this version of the model as a latent SVM, where the choice of which product noun phrase to extract is a latent variable at training time. We use the same datasets, features, and training setup as before.

System	**Darkode**							**Hack Forums**							**Combined**
	Noun phrases			*Product types*			*Posts*	*Noun phrases*			*Product types*			*Posts*	*Posts*
	Prec	Rec	F_1	Prec	Rec	F_1		Prec	Rec	F_1	Prec	Rec	F_1		
Frequency	61.8	27.9	38.4	61.8	50.0	55.2	61.8	41.9	16.1	23.3	41.9	35.9	38.7	41.9	50.3
Dictionary	57.0	60.0	58.5	67.6	55.8	61.1	65.9	38.3	47.8	42.5	50.3	43.1	46.4	45.4	54.1
NP-level	75.0	**79.4**	**77.2**	74.4	**86.7**	80.1	90.6	62.1	**60.6**	**61.4**	53.9	**74.3**	62.5	73.5	80.7
Post-level	**93.8**	37.0	53.1	**93.8**	70.3	80.4	93.8	**81.6**	23.7	36.8	**81.6**	54.7	65.5	**81.6**	**86.6**

Table 3.5. Results of the product extractor (trained on all training data) on the test sets of two forums. We report results for two baselines as well as for two variants of our system. Bolded results represent statistically significant improvements over all other values on that metric (in the same column) according to a bootstrap resampling test with $p < 0.05$. Our post-level system achieves 86.6% accuracy on product identification overall, making it robust enough to support many kinds of analysis.

3.4.2.3 Validation Results.

We considered three different metrics to validate the effectiveness of our product extractor:

- Performance on recovering individual product *noun phrases*. We compute precision (number of true positives divided by the number of system-predicted positives), recall (true positives over ground truth positives), and F-measure (F_1, the harmonic mean of precision and recall).

- Performance on recovering *product types* from a post: we compare the set of product head words extracted by our automated system with those annotated in the ground truth (after lowercasing and stemming), and evaluate with precision, recall, and F_1.[4]

[4]Note that this is still an unnecessarily harsh metric in some cases: *mail list* and *emails* will not be considered the same product type, but getting both right may not be necessary for analysis.

- Evaluation of a single product chosen from the *post*, checking whether we accurately recovered a product from the post, or correctly decided that it contained no products.

The latter two of these better match the analysis we want to carry out in this work, so we will focus our discussion on them.

Table 3.5 shows these metrics for our noun phrase and post-level classifiers. Throughout this table, we train on a combined set of annotated training data from both Darkode and Hack Forums. We compare against two baselines. Our *Frequency* baseline takes the most frequent noun or verb in a post and classifies it as a product. This method favors precision: it will tend to annotate one token per post. Our *Dictionary* baseline extracts a gazetteer of products from the training data and tags any word that appears in this gazetteer. This method favors recall: it will severely over-extract words like *account* and *website*, and will only fail to recover products when they have never been seen in the training set.

Our learned systems outperform the baselines substantially on each of the forums we consider. Overall, we find consistently lower results on Hack Forums across all metrics. One possible reason is that Hack Forums posts tend to be longer and more complex (10.4 lines per post on average as opposed to 6.0 for Darkode), as well as exhibiting a wider variety of products: when we ran the extractor on 1,000 posts from each forum, the Darkode sample contained 313 distinct products and the Hack Forums sample contained 393.

Our post-level system performs well on both post-level evaluation as well as on product type evaluation, indicating that posts generally have only one product type. We use this system for the analysis going forward. Overall, this system achieves 86.6% accuracy on posts from these two forums: high enough to enable interesting analysis.

3.4.2.4 Limitations.

Performance on other forums Because we train our product extractor on data drawn from particular forums, we would expect it to perform better at prediction on those forums than on others. We can evaluate this limitation by training and evaluating the system on distinct forums among those we annotated. Table 3.6 shows variants of our system trained on just Darkode, just Hack Forums, or on both training sets (the condition from Table 3.5). In both cross-forum evaluation settings, performance of the extractor significantly degrades due to the reliance of the system on fine-grained features. Hack Forums contains many more posts related to online gaming, which are virtually absent in Darkode, so a Darkode-trained extractor does not perform as well on these due to having never seen the relevant terms before. In experiments, we found that our extractor was roughly twice as likely to make an error identifying a product

	Product Type		
Train/Test Forums	Prec	Rec	F_1
Darkode-Darkode	92.7	69.5	79.5
Darkode-Hack Forums	69.9	46.8	56.0
Hack Forums-Hack Forums	81.6	54.7	65.5
Hack Forums-Darkode	89.6	67.2	76.8
Both-Blackhat	82.2	64.5	72.3
Both-Darkode	93.8	70.3	80.4
Both-Hack Forums	81.6	54.7	65.5
Both-Hell	81.8	42.5	55.9
Both-Nulled	87.2	67.5	76.1

Table 3.6. Cross-forum evaluation of the post-level product extractor. We report product type F-measure on the test sets for three variants of the post-level system: one trained on Darkode, one trained on Hack Forums, and one trained on both (as in Table 3.5). When the system is missing data from a particular forum, its performance degrades; the combined system works well on a range of forums.

not seen in the training data. However, our extractor still works on several other forums with only small losses in performance.

Handling posts with multiple products One potential problem with the post-level approach is that we can only partially capture posts selling more than one product, since the system only returns a single noun phrase. We find that this does not commonly occur: we analyzed a sample of 100 posts from Darkode and Hack Forums, and found that only 3 of them actually reflected selling multiple products.

3.4.3 Price Extraction

For each thread, we want to extract the price of the product bought or sold and the payment method (e.g., USD, PayPal or Liberty Reserve). Price extraction proves challenging because we need to distinguish the actual price from any other price-like phrases. For example, consider the following sentence:

$150 worth of vouchers for $5

Here, $5 is the actual price of the product and $150 is not. We need to be able to extract "$5" from the post, while ignoring "$150".

	Regex			SVM		
Train/Test Forums	Prec	Rec	F_1	Prec	Rec	F_1
Darkode	-	-	-	97.7	97.7	97.7
Hack Forums	-	-	-	84.3	91.3	87.6
Darkode/Hack Forums	-	-	-	76.0	69.7	72.7
Hack Forums/Darkode	-	-	-	83.7	81.8	82.7
Both/Darkode	33.7	37.8	35.6	97.8	100.0	98.8
Both/Hack Forums	21.9	45.5	29.6	84.7	91.7	88.1
Both/Hell	24.3	47.2	32.1	83.8	64.6	72.9
Both/Nulled	23.8	42.4	30.5	87.4	66.1	75.2

Table 3.7. Evaluation of the **Regex**- and **SVM**-based price extractors.

3.4.3.1 Labeling Ground Truth.

On every post, we annotate the price of the product, the payment method, and the currency, unless the post states the price in US Dollars (which we skip annotating for convenience). We do not annotate the payment method in the absence of prices. We annotate prices on the same dataset used for product extraction.

3.4.3.2 Models.

We consider one baseline model and one machine-learning based model for price extraction:

Regex extractor Extracts all the numbers and known currencies from a post as the price(s) of the product mentioned in the post. Ignores any contextual information from the posts.

SVM based extractor Labels each token as a price or a payment method. The classifier uses token counts, position of a token in a post, parts-of-speech of the token, and membership in the Brown clusters as features. Brown clustering is a hierarchical clustering approach that creates clusters of similar words [15]. It can help disambiguate words used to refer to similar concepts.

3.4.3.3 Validation Results.

We evaluated both models on four forums: Darkode, Hack Forums, Hell and Nulled. We excluded Blackhat World for this analysis because of its low number of threads with prices. In our annotation dataset, 11.02% of the posts on Darkode mention pricing information. The rest of the posts usually ask the prospective buyer to

send a private message to the poster to negotiate price. We noticed the opposite on Hack Forums, where 49.45% of the posts mention price. Hell and Nulled also have a higher number of posts with prices than Darkode, 19% and 44% respectively.

The **Regex** extractor performs poorly compared to the **SVM** extractor (Table 3.7). In our dataset, 40% of the numbers and currencies mentioned in a post are related to prices. Without contextual information, the **Regex** extractor cannot recognize various ways of mentioning a price, and cannot distinguish regular numbers from prices.

For the **SVM** extractor, we achive both higher precision and higher recall when we train and test the model on the same forum. The accuracy on Hack Forums exceeds that for the other forums, perhaps due to Hack Forums much larger size than Darkode, thus providing more data for training the classifier. The majority of the errors occur for words used for both pricing and non-pricing information. For example, in the following sentence "pm" means private message: "Contact me via xmpp or pm"; in other contexts, "pm" can also mean Perfect Money, as in "Only accept PM or BTC."

3.4.3.4 Limitations.

The accuracy of the price extractor decreases when we train and test on separate forums. The discrepancy in accuracy may reflect different forums using different payment methods and discussing pricing information differently. For example, Bitcoin is one of the most used currencies on Hack Forums, but we never find it mentioned with a price on Darkode.

For some product categories, a price is not meaningful without a unit, which our current classifiers do not extract. For example, the price of 1,000 accounts is likely to be higher than the price of one account. While knowing the unit is important (especially if we want to compare the price of one category of product across multiple forums), in our dataset unit pricing is relatively rare. Only 4.09% of the posts on Darkode and 12% of the posts on Hack Forums mention a unit.

3.4.4 Currency Exchange Extraction

Some of the forums we considered contain large sections focused on exchanging money between electronic currencies and payment systems such as Liberty Reserve, Bitcoin, and PayPal. We treated these posts entirely separately, since the "product" is not a single noun phrase, and the price is not a single number. Instead, we consider the task of extracting several pieces of information: currencies offered, currencies desired, amounts offered, amounts desired, and exchange rates. For each of these, we wish to extract either a value, or a decision that the post does not contain it (for example, no amount appears).

	All Fields				Payment Methods Only				
Models	Prec	Rec	F_1	Ex.	Prec	Rec	F_1	Ex.	Rev
Fixed	69	58	63	29	80	64	71	61	14
Pattern	90	56	69	39	93	67	78	64	3
Classifier	81	79	80	46	81	83	82	61	6
Global	87	73	80	50	86	80	83	70	5

Table 3.8. Validation results for the currency exchange extractor. We report results for all four models, evaluating extraction of all fields (left), and for only the currencies being exchanged (right). We assess metrics of Precision, Recall, F-measure, percentage of fully matched posts (Ex.), and for currencies, the percentage of posts in which we find the transaction direction reversed.

Labeling Ground Truth Our annotators find this task much more clear-cut than product extraction or price extraction. Three annotators labeled 200 posts, 100 of which we used as a development set and 100 for validation. Two of the annotators also each labeled an additional 200 posts, producing a 400 post training set. In each case, we annotated tokens as either relating to what the post offers, what it requests, or the rate.

Models We considered two baselines and two learned models:

Fixed Order. Uses regular expressions to identify tokens corresponding to numerical amounts or known currencies. We consider the first amount and currency mentioned as offered, and the second amount and currency mentioned as requested.

Pattern-Based. Extracts patterns of token sequences from the training data, with infrequent words discarded, and numerical values and currencies collapsed into special markers. When a pattern matches text in a post, we marked the tokens as currency or amount according to the pattern.

Token Classifier. A learned classifier using local context to label each token as one of the pieces of information to be extracted.

Global Extractor. An extension of the token classifier that makes decisions about all tokens in the post simultaneously. This means decisions can interact, with the label for one token depending on labels chosen for other tokens.

Validation Results Table 3.8 shows validation results on Hack Forums. In the evaluation, we de-duplicate trades mentioned multiple times in a single post (i.e., when a single post describes the exchange more than once, the system only gets credit once for getting it right).

As expected, of the two baselines, the pattern-based approach has higher precision, but cannot raise recall. The learned models balance these two more effectively, leading

to further overall improvements in F-measure, and reaching 50% exact match on posts. In the remaining cases, the errors relate to a mixture of the three types of data of interest.

3.5 Analysis

3.5.1 End-to-end error analysis

To compute the end-to-end error of the type-of-post, product, and price classifiers, we manually evaluate 50 posts from Hack Forums and Nulled. For this evaluation, we consider the product classifier output as correct if it extracts the correct product noun phrase. Overall, 14% of the posts on Nulled and 16% of the posts on Hack Forums had at least one misclassification. For both of the forums, the three classifiers never made an error in the same post—understandable given the differing nature of the three classification tasks.

3.5.2 Broadly Characterizing a Forum

To get a shallow picture of the activity going on in a forum, we can simply assess the most frequently bought and sold products. The first two columns of Table 3.9 show the 10 most frequently occurring products in Darkode and Hack Forums extracted according to two methods: take the most frequent nouns, or take the most frequent product headwords. We much more consistently extract actual products, as opposed to other features of e-commerce like currencies. Moreover, they highlight interesting differences between the forums that the word frequency method misses: Darkode has a higher amount of activity surrounding malware installs and exploits, whereas Hack Forums has a larger amount of activity related to online gaming (*cod*, *boost*). This rough picture could provide an analyst with a starting point for more in-depth investigation.

Our product extractor also supports finer-grained analysis, with its prediction of complete noun phrases. If we collect the most frequent noun phrases (last column of Table 3.9), as opposed to headwords, we have a new frequency distribution that surfaces terms like *steam account* for Hack Forums, a gaming-related concept. The category *account* disappears and others rearrange because they are fragmented into subtypes. Accurately characterizing activity surrounding accounts poses a challenging task that we address in more detail in the next section.

3.5.3 Performance

We focused our evaluation of our automated tools on accuracy rather than runtime, because our tools execute quickly enough to enable in-depth, real-time analysis. For

Word freq.		Products		Product NPs
Darkode	Hack Forums	Darkode	Hack Forums	Hack Forums
pm	pm	install	account	crypter
price	vouch	account	service	space
site	service	traffic	crypter	service
traffic	account	email	space	setup
bot	am	bot	setup	cod
email	view	root	cod	crypt
u	paypal	exploit	crypt	boost
server	price	service	bot	steam account
anyone	method	rdp	boost	server
lr	time	site	server	method

Table 3.9. Top frequently-occurring stemmed nouns in Darkode and Hack Forums from two methods: simple frequency counts, and looking only at nouns tagged as products by our product extractor. The product extractor filters out numerous frequent but uninteresting concepts related to commerce (*price*, *lr*) and allows analysts to more quickly see differences of greater interest.

the type-of-post classifier, training the classifier from scratch and running it on the complete forum took less than 5 minutes on the English language forums (using four threads on a quad-core Macbook Pro). For the German and Russian language forums, it took 10 minutes. Our product extractor can also process the forums in 5-to-15 minutes (15-to-30 posts per second on a single core of a Macbook Pro). The price extraction and currency exchange pipelines had similarly fast runtimes, analyzing a forum in a few minutes.

3.6 Case Studies

The methods developed in Section 3.4 provide tools that an analyst can use to answer specific questions of interest. To demonstrate this, in this section we present two case studies. In Section 3.5.2 we showed that our product extractor can provide useful high-level characterization of the activity in a forum. Section 3.6.1 then shows how to take this starting point and extend it to a more fine-grained analysis of particular products. This analysis requires only a few simple rules that an analyst might write down in an hour or two of study, and shows what our methodology can provide "out of the box."

Then, in Section 3.6.2, we delve deeper into a subset of posts not handled well by our existing tools, namely those involving currency exchange. Tackling this part of the

underground economy requires developing additional extraction machinery: we show that we can use a process similar to that for annotating our product extraction dataset to build a currency exchange detection system here as well.

3.6.1 Identifying Account Activity

Noun phrases produced by our product extractor may not immediately expose the types of cybercriminal activity of interest to an analyst. Table 3.9 shows that the head *accounts* is very common across two forums, but these posts might correspond to users selling hacked accounts or merely selling access to one-off accounts that they legally own, a distinction of potential interest to an analyst. Knowing the type of account (*steam account* versus *instagram account*) does not necessarily help us narrow this down either.

To analyze account activity in more depth, we can use our product extractor as a starting point. We gather all posts related to accounts according to the product extractor: these include posts with product headwords *email*, *account*, or names of popular services from a small whitelisted set (eg., *hotmail*, *snapchat*).[5] After gathering these posts, we observed a simple rule: we find that plural headwords (*accounts*, *emails*) almost always reflect users trafficking in illegally acquired accounts, whereas singular headwords typically reflect users selling their own accounts.

We can evaluate the efficiency of this simple set of rules on top of our product extractor. To do so, one of the authors undertook a fine-grained labeling of a set of 294 forum posts distinct from those used to train the product extractor. This labeling distinguished original ("OG") accounts (58 posts out of our 294) from bulk/hacked accounts (28 posts out of 294). We can then evaluate the accuracy of our product extractor and rules in surfacing account posts, and correctly distinguishing between the two account classes.

Table 3.10 shows the results from our method on this dataset. We compare against a simple heuristic: we grep for occurrences of *accounts*, declare those to be bulk/hacked accounts, and then grep for occurrences of *account* in what remains.[6] Our method outperforms this metric by roughly 9 F_1, with gains in both precision and recall. Note that the F-measure here captures both how often we can surface account posts as well as how often we correctly distinguish between the two classes. Our method saves the analyst time (by having higher precision) and finds a higher number of relevant posts (recall) compared to our baseline.

[5]We further exclude a few common noun phrases that correspond to spamming services instead, namely *bulk email*, *mass email*, or *[number] email*.

[6]Expanding what we grep for improves recall but harms precision; for example, including *email* as well decreases F-score to 54.3.

	Prec	Rec	F_1
Grep Baseline	58.1	64.3	61.0
Product Extractor	69.0	71.4	70.2

Table 3.10. Accounts case study. We have a three-class classification task of posts: they deal in original accounts, in bulk/hacked accounts, or not in accounts at all. Compared to a grep baseline, a method based on our product extractor performs better at identifying relevant posts. Precision measures how frequently we obtain a correct extraction when we identify a post related to either type of account, and recall measures how many of the gold-standard account-type posts we identify and classify correctly.

3.6.2 Currency Exchange Patterns

In Figure 3.2, we show the result of extracting transactions from the three forums using our tool. We label rows with the payment mechanism offered and columns with the one sought. Each cell of the table shows the number of posts of the designated type. The most popular three payment mechanisms are Liberty Reserve (now defunct), Bitcoin, and Paypal.

By far the most popular exchange offered is Bitcoin for PayPal, both on Hack Forums and Nulled. We suspect the reason for the demand is that exchangers can profit by charging on average a 15% fee to exchange Bitcoin and other difficult to obtain currencies for PayPal (calculated using extracted amounts and rates).

One unusual value is the square for Hack Forums showing Bitcoin–Bitcoin transactions. These indicate mistakes in our analysis of the extractor's output, where we treated "coins" as referring to Bitcoin, when in some cases they mean other types of crypto-currencies. Fortunately, this issue rarely occurs.

We also found surprising to observe demand for moving money from Paypal, Bitcoin, and other payment mechanisms to credit cards. Further investigation of thirty posts showed that half of these reflect requests for someone to make purcahses using with a credit card, using some other means to repay them. The other half arose from a combination of errors in our extraction, mostly related to statements regarding "CC verified" paypal accounts. These errors contrast sharply with the high accuracy observed when spot-checking one hundred of the Bitcoin to Paypal transactions (97% correct), indicating that our accuracy depends significantly on currency.

3.7 Conclusion

In this chapter, we built several tools to enable the automatic classification and extraction of information from underground forums. We can apply our tools across a variety of forums, accommodating differences in language and forum specialization. We

Hack Forums

Have \ Want	Paypal	Bitcoin	(missing)	Liberty Reserve	Payza	Open Metaverse	Litecoin	Western Union	Moneypak	Moneybookers	Skrill	Web Money	Pay Safe Card	Bank Deposit	Dogecoins	Credit Card	Ukash	Perfect Money	Omnicoin	Alterpay	Moneygram	Other (24)
Other (24)	124	71	98	26	7	2	5	2	2	3	1	7	1	2	3	1			1	~		4
Moneygram	13	25	16	16	~	~	1	2	1	~	~	4				~		1				~
Alterpay	18	2	12	11	2					~				1		~						~
Omnicoin	20	8	25			10	1	~						1	2			~				1
Perfect Money	59	21	11	1	3	~	~	~		9	3	4	1	2	1		~				1	3
Ukash	17	9	21	10	1		1	~		2	1	2	4	~		~		1				~
Credit Card	41	14	26	13	2			~	2	~	2	~		~	~							1
Dogecoins	76	34	97			2	7	~	~	0	0	1		~				~	1			2
Bank Deposit	73	100	74	50	5	~	5	4	2	2	2	7	2		1	1	1	1	~		~	4
Pay Safe Card	58	19	45	21	1	~		~		3	~	1		~	~	~	4					2
Web Money	144	84	77	40	8	2	3	6	6	10	6		4	2	1	1	4	1		1	~	4
Skrill	147	116	63	7	10	6	4	1		2		5	4	2	1	1	0	~	~		~	2
Moneybookers	159	29	67	117	27		1	1	1		3	4	2	3		6	~	1		4		3
Moneypak	59	48	62	56	5	0	1	3		1	~	~	~	~		1		1		2	~	1
Western Union	204	226	102	148	6	5	10		16	3	8	12	1	4	1	3	1	3	~	1	1	3
Litecoin	352	94	289	19	5	25		9	8	~	10	5	2	2	14	1	1	~	4		1	15
Open Metaverse	202	88	201	~	1		6	1	1	~	3			2	2				6			1
Payza	286	39	120	109		1	2	1	2	7	3	2	1	3	~	2	0	2	~	3	1	6
Liberty Reserve	3,61	231	997		234		28	103	175	77	6	48	45	26		14	30			22	14	42
(missing)	2,41	1,51		715	203	152	88	47	166	82	47	45	61	27	46	9	21	9	18	10	1	56
Bitcoin	3,61	3	2,41	120	67	344	96	116	234	38	153	56	69	42	59	14	27	14	32	2	6	65
Paypal		3,21	1,22	2,11	554	227	155	27	123	113	97	48	221	30	60	25	18	24	20	39	3	106

Nulled

Have \ Want	Paypal	Bitcoin	(missing)	Pay Safe Card	Skrill	Steam	Other (11)
Other (11)	1	5	6	~			
Steam			1				
Skrill	4		~	0			
Pay Safe Card	2	0	2				
(missing)	46	17		5	0		2
Bitcoin	103		42	5	2	2	2
Paypal		32	20	10	2	2	1

Darkode

Have \ Want	Liberty Reserve	Web Money	(missing)	Paypal	Western Union	Moneybookers	Other (7)
Other (7)	~		3	1			
Moneybookers	5	2		1			~
Western Union	8	3	4	0			0
Paypal	11	10	9			3	
(missing)	11	13		4	0		0
Web Money	16		33	19	1	1	4
Liberty Reserve		27	34	27	6	5	6

Figure 3.2. Number of transactions of each type observed in each forum. Numbers indicate how many posts had each type of transaction. If multiple currencies appear for either side of the transaction, then we apportion a fraction to each option. Colors indicate less than 100 (yellow), 100–1,000 (orange), and 1,000+ (red). Values are rounded, with values between 0 and 0.5 indicated by a ˜ . Values for *(missing)* show when that side of the transaction was either not mentioned or not extracted (antepenultimate row in the tables). Values for "Other (#)" are the sum over other payment methods (combined to save space).

tested our tools on 8 different underground forums, achieving high performance both within-forum and across-forum. We also performed two case studies to show how analysts can use these tools to investigate underground forums to discern insights such as the popularity of original vs. bulk/hacked accounts, or what kind of currencies have high demand. Our tools allow for future researchers to continue this type of large-scale automated exploration to extract a holistic view of a single or several underground forums, as well as potentially provide support to law enforcement investigating cybercrime.

Chapter 4

CSAM Forums

4.1 Introduction

Behind every image and video of child pornography, there is a real child who is being sexually abused and victimized [6]. The advent of the Internet and the Dark Net has not only markedly increased the proliferation of CSAM, but also created a new space for pedophiles to find and target future victims for abuse [11]. Most of the work in this domain has focused on CSAM in peer-to-peer networks, with researchers performing measurement studies ([90] [37] [77]) or building classifiers to detect CSAM ([64] [20] [74]). To the best of our knowledge, no one has yet undertaken any analysis or tool-building geared towards processing and classifying data on CSAM forums hosted on Tor onion sites. The visitors to these sites number in the thousands, some of whom are currently, actively sexually abusing children, and sharing this content with their peers. These persons are known as producers. In this chapter, we aim to develop and demonstrate automatic techniques for extracting these producers from CSAM forums.

The rest of this chapter is organized as follows. Section 2 provides the necessary background for the rest of the chapter. Section 3 outlines the four CSAM forums which we analyzed and used to evaluate our tools. Section 4 describes the methodology for building our producer extractor, covering ground truth labeling, the models we built, and validation results. We conclude with reiteration of key contributions and findings.

4.2 Related Work

4.2.1 Anomaly Detection in Social Network Users

Non-PCA Most of the work in this space has focused on non-PCA, primarily supervised learning techniques to detect anomalous and/or misbehaving users in social networks ([52] [80] [12]). These works generally assume a clear distinction between misbehaving and non-misbehaving entities (e.g., spammers vs. legitimate users). This is not the case in our domain, where the difference is between a pedophile who is hands on abusing, and a pedophile who is not hands on abusing (in both cases, the users are participating in illegal activity, whether that be direct abuse or the ownership of CSAM).

Stringhini et al. [80] designed a set of features for a Random Forest classifier intended to distinguish between spammers and legitimate user accounts. Their data came from Twitter, Facebook and MySpace. In order to collect the necessary ground truth data, they created a set of honey-pot profiles across these three sites, using the content and messages received to then label the senders as either spammers or legitimate. For Facebook, with 173 spam bots and 827 legitimate accounts, they achieved a false positive rate of 2%, and a false negative rate of 1%. For Twitter, with 500 spam bots and 500 legitimate accounts, they achieved a false positive rate of 2.5%, and a false negative rate of 3%.

Benevenuto et al. [12] also focused on distinguishing between spammers and legitimate user accounts, solely on Twitter. They collected data from 54 million users, and hand labeled these users as either spammers or not spammers. Using a supervised machine learning approach with features they designed from this data, they were able to correctly classify 70% of the spammers, and 96% of the non-spammers.

Egele et al. [24] designed a set of features for a supervised machine learning platform intended to extract compromised user accounts from social networks. They focused their attention on Twitter and Facebook, training their model with a manually labeled dataset of misbehaving and non-misbehaving users.

PCA While PCA has been extensively explored for anomaly detection in network traffic, not much work has explored PCA to extract anomalous and/or misbehaving users. Specifically, Viswanath et al. [88] present a PCA based technique that models the behavior of normal users, and then flags those who deviate as anomalous. Their data comes from Facebook, with three sources for ground truth for misbehaving users: fake accounts, compromised accounts, and collusion networks. Their source for normal users came from vetted, technically savvy Facebook groups (in an attempt to avoid users who might be infected by malware), and a random sample of users from Facebook's people directory. In total, they collected data for 6.8K users. Their detector successfully

flagged 66% of misbehaving users, with a false positive rate of 0.3%.

4.2.2 CSAM

De-Anonymization All of the de-anonymization work in this space has focused on identifying sexual predators who groom children online ([39] [65] [45] [55]). Sexual predators online seek out underage victims who they can acquire information about, sexually desensitize, engage with in sexually explicit text or video chat, and (for some) eventually convince to meet in person. [33]

Pendar's [65] high level goal was to build an automatic recognition system of on-line predators, using machine learning combined with NLP text features. They conceptualize this problem as building a classifier that can distinguish, when given a chat where a pedophile is grooming a victim, between the victim and the pedophile. Obtaining ground truth in this space can be very difficult. As a result, Pendar made use of open provided chat data from the Web site www.perverted-justice.com. This site recruits volunteer contributors who pose as underage children in chatrooms; when a pedophile has been found, the Web site posts archives of all text chats with them online. In total, their dataset consisted of 701 chat logs (i.e., conversations between sexual predators and presumed underage victims), with the logs ranging in length from 269 and 42,220 words. Using standard word-level n-gram and TF-IDF features, [65] built a k-Nearest Neighbor (k-NN) classifier that achieved 94.3% classification accuracy with k = 30.

Kontostathis et al. [45] tackled the same problem, but with hand-coded features. With 288 chat logs from www.perverted-justice.com, ranging in length from 349 to 1,500 lines of text, they assigned specific terms and phrases in the logs to 8 large categories of predator communication, including approach (e.g., "are you safe to meet"), communicative desensitization (e.g., "i just want to gobble you up"), and reframing (e.g., "there is nothing wrong with doing that"). Using the J48 classifier, they achieved 60% classification accuracy in distinguishing text by predator from text by victim. They had more success using this same feature set and classifier in distinguishing the 288 predator-victim chats from a separate set of innocuous chats, achieving 93% accuracy. Finally, [45] also did some preliminary exploration into clustering predator by communication style, creating 8-dimensional vectors for each predator in their dataset. Each dimension consisted of a count of the number of phrases in each of eight luring categories: relationship, reframing, personal information, isolation, compliment, communicative desensitization, approach, and activities. Using k-means with k = 4 produced the minimum intra-cluster correlation.

Inches and Crestani [39] outline the results of the Sexual Predator Identification competition at PAN-2012. The competition had two main technical goals: identifying the predators among all the users in different conversations, and identifying the lines

of the predator conversations which are most distinctive in their predatory behavior. The dataset compiled for the competition consisted of chat logs from www.perverted-justice.com, adult sexual conversations from Omegle, and publicly available IRC logs: 357,622 conversations total. They proportioned their data to realistically reflect, in their opinion, the low proportion of predator conversations with respect to regular ones on the Internet. For the first task, the training and testing set each consisted of 65,000 unique conversations, with fewer than 4% labeled as predator conversations. Overall, less than 1% of the total author set consisted of predator authors. The contestant with the best performing classifier for the first set used both standard NLP features like n-grams, TF-IDF and bag-of-words feature representation, as well as predator specific vocabulary collected in other research like [45]. They achieved a true positive rate of 0.9804 and false positive rate of 0.7874, using a Support Vector Machine (SVM) model. No training data was provided for the second task. For the second task, the classifier with the highest recall, 0.8938, had a precision of 0.0915, and the classifier with the highest precision, 0.4510, had a recall of 0.1869.

Measurement Studies To date, most of the manual analysis the technical research community has conducted in this space consists of measurement studies of peer-to-peer (P2P) networks where CSAM is exchanged ([90] [37] [77] [50]). In addition to providing a better understanding of these networks, these studies also give suggestions on how to reduce the number of CSAM files available on said networks ([37] [78] [90]).

Hurley et al. [37] performed a comprehensive measurement study focusing on Gnutella and eMule, P2P networks used by some to share CSAM. The datasets spanned a one-year period from October 1, 2010 to September 30, 2011. They observed over 1.8 million distinct peers, from over 100 countries sharing CSAM on eMule, and over 700,000 peers on Gnutella. They found that most CSAM files were shared by a small set of aggressive users that were geographically diverse, and that most of these files were only available for a short amount of time. They also found a high degree of overlap among the CSAM files available on the two networks: 26,136 of the CSAM files on the eMule network (nearly 89%) were seen on the Gnutella network, and 97% of Gnutella peers were observed with at least one file that was also on the eMule network. Overall, on a daily basis, an average of 9,712 distinct files were available on both networks, with a peak of 32,020 files in one day.

Latapy et al. [50] and Steel [77] focus their attention on queries issued by users of the eMule and Gnutella networks. Specifically, they analyze what proportion of those queries are related to CSAM. [50] analyze two datasets of keyword-based search queries issued by users of the eMule system; the first spanning 10 weeks in 2007 and the second 28 weeks in 2009. In these datasets, they found that about 0.25% of queries relate to child pornography and about 0.2% of peers on the network are involved. [77] analyzes the supply of and demand for CSAM on the Gnutella network, using a dataset of both

queries and query hits collected over several weeks. They found that 1% of all queries and 1.45% of all query hits were CSAM-related, with the median age searched for being 13. They also found that while most of the available CSAM files are images, 99% of searches are for videos.

Wolak et al. [90] also performed a measurement study, focusing specifically on the Gnutella network. With access to a year of online CSAM sharing activity by U.S. computers and the IP addresses of said computers, the authors performed a lower-bound estimate measuring the amount of CSAM sharing in that particular network. This data was provided to them by law enforcement agencies. Of the 244,920 computers sharing 120,418 unique known CSAM files on Gnutella during the study year, they found that more than 80% shared fewer than 10 files during the course of the study. Most of the files were originating from less than 1% of the total computers involved.

In [37], Hurley at al. had the specific goal of finding the peers that, when removed, would minimize the number of files that are available for at least one day. As this problem is NP-hard, they instead offered and assessed several greedy heuristics for reducing the availability of CSAM by removing peers. Removing peers by either contribution or corpus size proved to be the most successful heuristics: if they were to remove the top 0.01% of 775,941 peers in the Gnutella network with the biggest corpus, only 59% of the known CSAM files would remain available in the network. The same proportions held true for the eMule network. Given this information, the authors recommended a triage strategy for law enforcement officers that would have them focusing on removing the most aggressive offenders, i.e., those that are online for the longest duration and share the largest amount of CSAM content. [90] came to a similar conclusion, stating that if law enforcement arrested the operators of high-contribution computers - i.e., the top 1%, 915 of 244,920 total - and took their files offline, the number of distinct known CSAM files available in the Gnutella P2P network could be reduced by as much as 30%.

In [78], Steel analyzed mobile device use for CSAM consumption on the Internet, as well as the global impact of deterrence efforts by various search providers operating on the open web (as opposed to the Dark Net). They found that mobile devices are a substantial platform for web-based consumption of CSAM, with tablets and smartphones representing 32% of all queries associated with CSAM conducted on Bing. Separately, they also found that blocking efforts by Google and Microsoft resulted in a 67% drop from 2013 to 2014 in web-based searches for CSAM.

Hurley et al. [37] also ranked, by aggressiveness, six different subgroups they found in the Gnutella and eMule P2P networks. Of those six, they found that the subgroup containing the top 10% of peers sharing the largest corpora was the most aggressive. The authors also found that most peers using a known Tor exit node did not do so consistently: on both networks, only 25% of peers who at some point used Tor to connect, did so every time they connected. Under 40% consistently used Tor to connect after their first use. As a result, the authors concluded that the use of Tor, as observed

in practice, poses only a small hurdle to investigators. It is important to note that the authors do not include in their assessment CSAM traffickers and child abusers who specifically gather on CSAM forums and chatrooms hosted on Tor, and who obviously benefit from Tor's anonymity.

Detecting CSAM Most of the research for automatic analysis and classification tools in the CSAM space has centered around identifying instances of CSAM content online ([64] ([20] [74] [17] [85]). Traditional solutions to organize CSAM use file hashes like MD5 sums to match seized material with databases of known CSAM maintained by law enforcement. The vast majority of the current research in this space, however, focuses on the automatic detection of CSAM material, both novel and already known. The detection of novel material is especially critical, as it indicates a new child suffering abuse. Below we outline some of the existing techniques to automatically classify CSAM.

Network Based Detection Shupo et al. [74] detect CSAM in network traffic at the packet level. They used a statistical feature extraction process on captured images, then compared these feature to feature vectors of known CSAM using various distance metrics (e.g., edit distance) for classification. They achieved false positive and false negative rates ranging between 0.2 at the lowest and 0.45 at the highest. Chopra et al. [17] summarize a variety of existing and potential work in this space, including CSAM detection systems that resemble existing network intrusion detection systems.

Content Based Image Retrieval Content-based image retrieval (CBIR) [75] detects CSAM using a query-by-example strategy on a database: given a questioned sample image, visually similar (as measured by color contrast and shapes) CSAM in the database is retrieved. Several forensic tools in this space use CBIR ([1] [2]); this technique is particularly useful for identifying images that are themselves unknown but come from well-known shoots, series, or locations.

Image Recognition Image recognition remains a fruitful research field for detecting pornography in general ([72] [41] [28] [21]), with some authors focusing their attention specifically on CSAM. Ulges and Stahl [85] used a bag-of-visual-words feature representation (i.e., discretizing images as collections of local, visually coherent patches referred to as visual words) with an SVM classifier. They achieved CSAM classification error rates of between 11 to 24%, where the non-CSAM data consisted of other (non-CSAM) pornographic images, and non-pornographic images from Flickr, Corel and the general web. Peersman et al. [64] designed a tool that classifies both image and video as CSAM or not CSAM, using color, skin-presence, visual word features, and

Forum	Source	Primary Language	Date covered	Users	Full Members
Forum 1	Complete Dump	English	Jul 2015–Oct 2016	72,391	42
Forum 2	Complete Dump	English	April 2016–Oct 2016	325,188	325,188
Forum 3	Complete Dump	English	April 2016–Oct 2016	708	708
Forum 4	Complete Dump	English	April 2016–Oct 2016	2,338	11

Table 4.1. General properties of the forums considered.

audio features from the video. The non-CSAM data consisted of other (non-CSAM) pornographic images and video collected from sites such as Redtube and Pornhub, as well as non-pornographic images and video collected from Flickr and Youtube. Using the SVM algorithm, they achieved 92% classification accuracy for image, and 95% for video. Polastro and Eleuterio [20] similarly designed a CSAM image detection tool, although their image component is restricted to detecting nudity.

Text-Based Detection Peersman et al. [64] also use text analysis in their larger system, building a SVM classifier that distinguishes filenames as CSAM files or adult pornographic media. They trained their classifier on 268 filenames of known CSAM content, and 10,000 non-CSAM filenames from adult pornography, using both pedophile keywords and standard n-grams as features. They achieved 92.9% precision, 52.5% recall and an F-score of 67.1%. Panchenko et al. [62] exclusively focused their attention on classifying filenames as either CSAM or non-pornographic at all, achieving a best performance of 97% accuracy.

4.3 Forum Datasets

We consider four forums hosted on Tor onion sites (Table 4.1). All four forums are from complete database dumps that contain all public posts, private messages and metadata prior to their retrieval.

Forum 1 The first forum is an English-speaking forum. The forum started in 2015, and is no longer active. The data that we have spans from July, 2015 to October, 2016. This forum was an invite-only forum, requiring the potential member to provide proof that they either had, or currently were, sexually abusing a child. Guest users (those who did not apply, or were denied access) were still allowed to browse most boards on the forum. Those who were admitted as full members were given access to boards that only their fellow producers could view. 72,391 unique users appear in our data; 42 of those were full members. Members used this forum to share CSAM content, discuss their own history of abusing children, share ideas for abuse with each other, and discuss their sexual preferences.

Forum 2 The second forum is a mainly English-speaking forum, with some boards present that provide a space for Russian, French, Spanish, Portuguese, German, Polish and Dutch speakers. The forum focused on CSAM of children between the ages of three and 17. Members primarily used this forum to share CSAM content, and discuss the abused children in said content. They also used the forum as a space to share news items about former victims, discuss references to pedophilia in the larger culture, share tips on how to best remain anonymous and safe from law enforcement, and spread advice on how to most effectively sexually abuse children. Membership was open to anyone with an email address and the tech-savvy to use a Tor browser. The forum started in 2016, and is no longer active. The data that we have spans from April, 2016 to October, 2016. 325,188 users appear in our data.

Forum 3 The third forum is an English-speaking forum. This forum focused on feet, toe, and sole fetish CSAM of girls under the age of 12. Members used this forum to share CSAM content, discuss the abused children in said content, and discuss their sexual preferences. Membership required users to submit CSAM of girls under the age of 12, with feet content. The forum started in 2016, and is no longer active. The data that we have spans from April, 2016 to October, 2016. 708 users appear in our data.

Forum 4 The fourth forum is an English-speaking forum. The forum started in 2016, and is no longer active. The data that we have spans from April, 2016 to October, 2016. This forum was an invite-only forum, requiring the potential member to provide proof that they either had, or currently were, sexually abusing a child. Guest users (those who did not apply, or were denied access) were still allowed to browse most boards on the forum. Those who were admitted as full members were given access to boards that only their fellow producers could view. 2,338 unique users appear in our data; 11 of those are full members. Members used this forum to share CSAM content they produced with each other.

4.4 Extracting Producers

For each forum, we build a ranked list of users. We take an unsupervised learning approach, building a model for normal user behavior and flagging those highly ranked users as anomalous - i.e., producers. We use PCA to extract the normal subspace, i.e., the top principal components of the m x n matrix X, where m is the number of users and n is the dimensions of the feature space. For a particular user, we calculate the L2 norm of the residual portion of their vector. A large L2 norm indicates an anomalous user; we rank all users by that L2 norm. We also combine all forum data to build a ranked list of users across all four forums.

4.4.1 Labeling Ground Truth

To build a ground-truth dataset, we use both a list of known producers provided to us by law enforcement, and the list of full members from Forums 1 and 4, all of whom were vetted by the admins of the forums and provided evidence that they currently were abusing, or had abused, children. In total, we have the usernames of 56 producers. 47 of those producers are full members or guest users in Forum 1. 20 of those producers are users in Forum 2, two are users in Forum 3, and 10 are full members or guest users of Forum 4.

4.4.2 Features

Our model uses a variety of time series, category, and text-based features. These features can be divided into 7 main groupings: posts, images, videos, thanks made, thanks received, unigram tokens, and usergroup (e.g., 'admin', 'mod', 'guest'). For each of the first 5 groupings, we extract three different types of features:

- Distribution of occurrences per day (e.g., how many occurrences per day)
- Distribution of occurrences per forum topic (e.g., how many occurrences per forum topic)
- Distribution of occurrences across forum topic per day (e.g., how many occurrences per forum topic per day)

In other words, for the first grouping we extract the number of total posts made by the user across the history of the forum, the distribution of these posts by day, the distribution of those posts by forum topic, and the distribution of those posts by forum topic per day. Links and videos refer to link content (mostly CSAM) posted by users in the form of a link to either an image or a video. Thanks made and thanks received refer to a functionality that exists across all forums that allow users to thank other users for particular posts. For unigram tokens, we extract the distribution of words per day (e.g., how many words are written per day), the distribution of counts per token in the vocabulary (where the vocabulary consists of all unique tokens posted by all users in a forum), and the distribution of counts per token in the vocabulary per day.

4.4.3 Public Data

Different forums have different levels of public visibility. Given that ultimately this ranking will be most useful when run on forums where investigators do not have inside knowledge of who is a producer, we extract our features only from those portions of the forums that would be publicly visible to an investigator (with an email address,

the ability to use a Tor browser, and access to certain CSAM). In Forums 1 and 4, that consisted of all non-producer boards and non-private messages; in Forums 2 and 3, that consisted of all the content in the forum, excepting private messages. While all of the forums indicate some public form of usergroup of a user (e.g., 'admin', 'mod', 'guest') none of these usergroups explicitly reveal whether a user is a producer, or belongs to the producer boards; that information is confidentially kept by the admin(s) only.

Identifying the producers amongst the full set of users will significantly reduce the workload of an investigator: in both Forums 1 and 4, fewer than 1% of the users are explicitly named producers. Of course, we know that amongst the other users there are producers who choose, for whatever reason, not to join the producer section. False negatives are a real issue in this space, which is why we choose to return a full ranked list of the users, rather than cutting off at a particular L2 threshold, protecting against the circumstance where a false negative is never investigated by law enforcement because the user is left off the list.

4.4.4 Validation Results

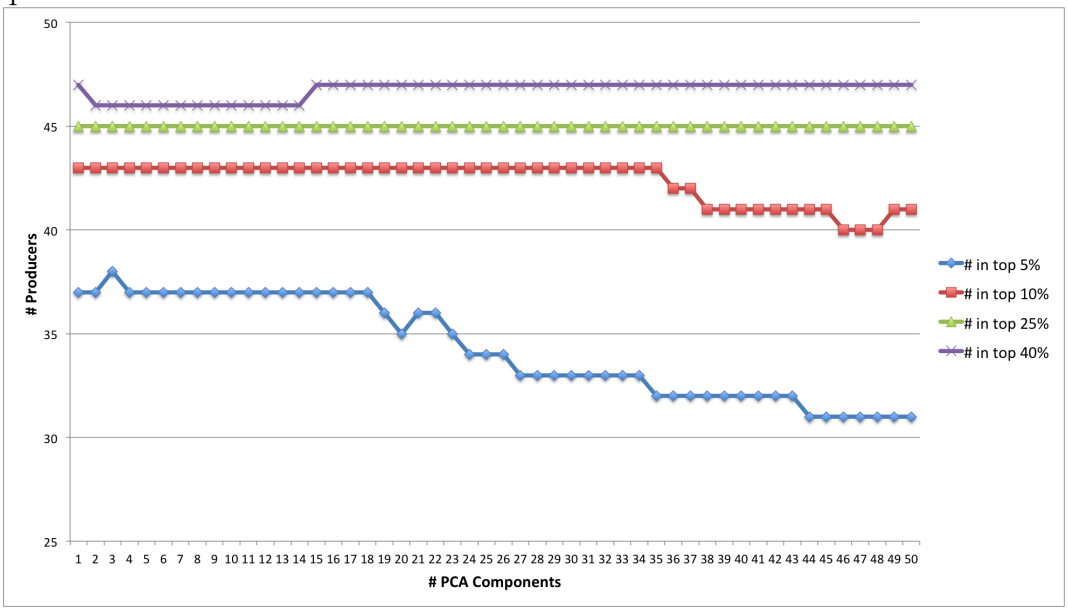
Each set of features returns a ranking of users, from largest to smallest L2 norm. We assess the performance of each feature set by looking at the L2 norms of the known producers. The best performing set of features is the one where the position of the lowest ranked producer (i.e., the one with the smallest L2 norm) is the highest, across all the feature sets. In other words, if feature set A returns a ranking where all the producers appear in the top 5,000 users, and feature set B returns a ranking where all the producers appear in the top 1,000 users, feature set B is considered the better performing set. When considering just the public data, that consisted of all the features except for thanks made, and unigram tokens. The results reported here are from the best performing feature set. We compare the performance of our best performing feature set against a simple baseline ranking by number of thanks received, where the user with the most thanks received is highest ranked.

Forum 1 For each user in the forum, we first discard the set of users who never post any content in the forum; 4,480 users remained.

Figure 4.1 shows the number of producers that appeared in the top five, 10, 25, and 40% of ranked users, with principal components from one to 50. The number of principal components has the most effect on the number of producers that appear in the top five percent of ranked users, with the number of producers declining as the number of principal components increases. For all but 13 of the principal component values, all 47 producers appear within the top 40% of ranked users. For those 13, all 47 producers appear within the top 42% of ranked users.

The best performing number of principal components seem to be 15 to 18, in which 37, 43, 45 and 47 producers appear in the top five, 10, 25 and 40% of ranked users, respectively. When ranking by number of thanks received, 15 producers appear in the top five percent, 31 appear in the top 10%, 37 appear in the top 25%, and 38 appear in the top 40%.

Figure 4.1. Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 1



These results are promising: not only does our feature set outperform the baseline, but it also includes the majority (78%) of known producers in the top five percent of ranked users. The baseline only includes 32% of known producers in the top five percent of ranked users. This represents a significant time savings for law enforcement - given this ranking, law enforcement would only need to look through 224 users (rather than potentially the full set of 4,480) to find the significant majority of known producers on the forum.

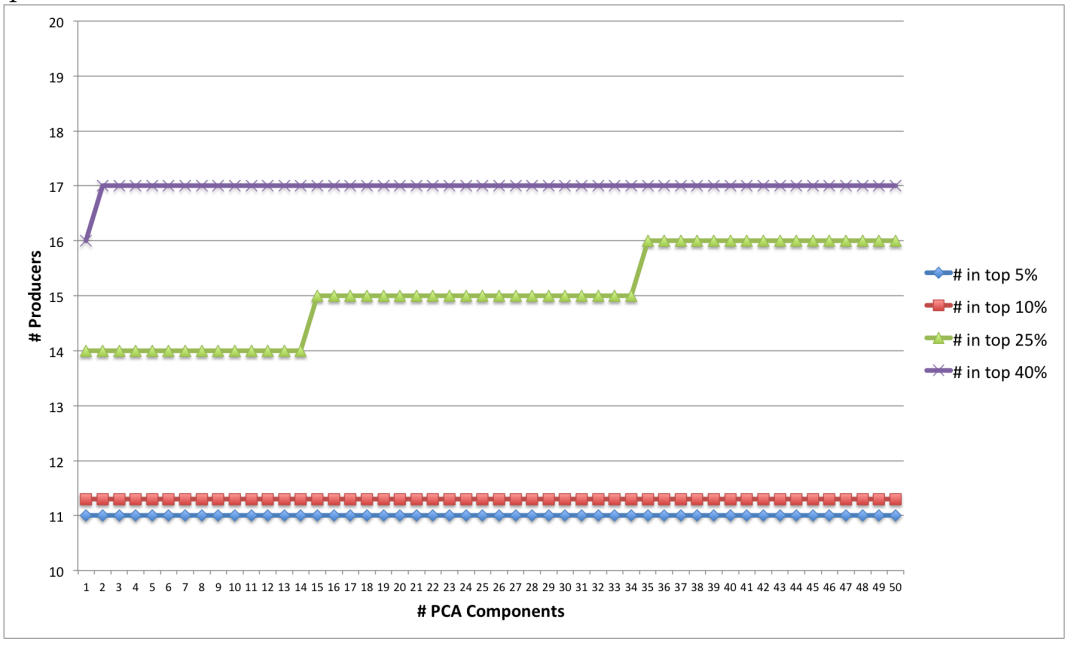
Forum 2 When discarding the set of users who never post any content in the forum, 9,836 users remained.

Figure 4.2 shows the number of producers that appeared in the top five, 10, 25, and 40% of ranked users, with principal components from one to 50. The number of principal components has the most effect on the number of producers that appear in the top 25% of ranked users, with the number of producers increasing as the number

of principal components increases. For all but one of the principal component values, 17 of the 20 producers appear within the top 40% of ranked users.

The best performing number of principal components seem to be 35 to 50, in which 11, 11, 16 and 17 producers appear in the top five, 10, 25 and 40% of ranked users, respectively. When ranking by number of thanks received, 10 producers appear in the top five percent, 11 appear in the top 10%, 12 appear in the top 25%, and 14 appear in the top 40%.

Figure 4.2. Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 2



While our feature set does outperform the baseline, the improvement is not significant. In either case (when using our feature set or the baseline) the majority (55%) of known producers appear in the top five percent of ranked users. Given this ranking, law enforcement would need to look through 492 users (rather than potentially the full set of 9,836) to find the majority of known producers on the forum.

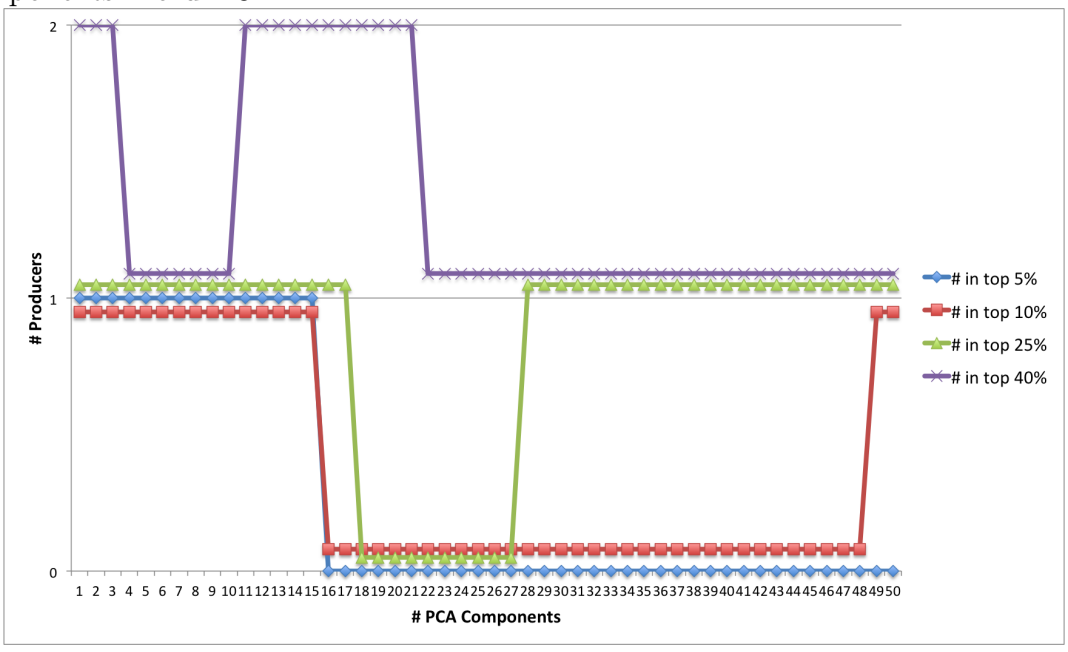
Forum 3 When discarding the set of users who never post any content in the forum, 196 users remained.

Figure 4.3 shows the number of producers that appeared in the top five, 10, 25, and 40% of ranked users, with principal components from one to 50. With only two producers in the set of users, the number of principal components did not have a significant effect on the number of producers that appear in the top percent. For

fourteen of the principal component values, both producers appear within the top 40% of ranked users.

The best performing number of principal components seems to be three, in which one, one, one and two producers appear in the top five, 10, 25 and 40% of ranked users, respectively. When ranking by number of thanks received, zero producers appear in the top five percent, zero appear in the top 10%, one appears in the top 25%, and two appear in the top 40%.

Figure 4.3. Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 3



These results are again promising: not only does our feature set outperform the baseline, fully half (50%) of known producers appear in the top five percent of ranked users. This represents a significant time savings for law enforcement - given this ranking, law enforcement would only need to look through 10 users (rather than potentially the full set of 196) to find the majority of known producers on the forum, and would only need to look through 67 users to find all known producers.

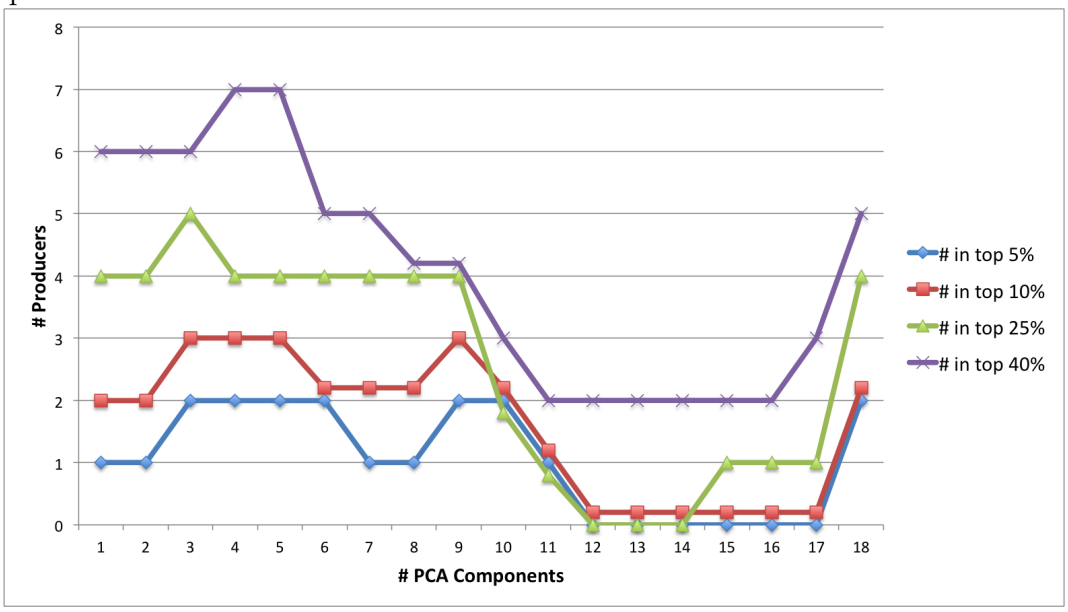
Forum 4 When discarding the set of users who never post any content in the forum, 18 users remained.

Figure 4.4 shows the number of producers that appeared in the top five, 10, 25, and 40% of ranked users, with principal components from one to 18. The number of principal components has the most effect on the number of producers that appear in

the top 40% of ranked users, with the number of producers generally decreasing as the number of principal components increases.

The best performing number of principal components seems to be four to five, in which two, three, four and seven producers appear in the top five, 10, 25 and 40% of ranked users, respectively. When ranking by number of thanks received, one producer appears in the top five percent, two appear in the top 10%, four appear in the top 25%, and six appear in the top 40%.

Figure 4.4. Number of Producers in top % of Ranked Users, with varying Principal Components: Forum 4



While our feature set does outperform the baseline, the improvement is not significant. In either case (when using our feature set or the baseline) the majority (60%) of known producers appear in the top 33% percent of ranked users. Given this ranking, law enforcement would need to look through 6 users to find the majority of known producers on the forum.

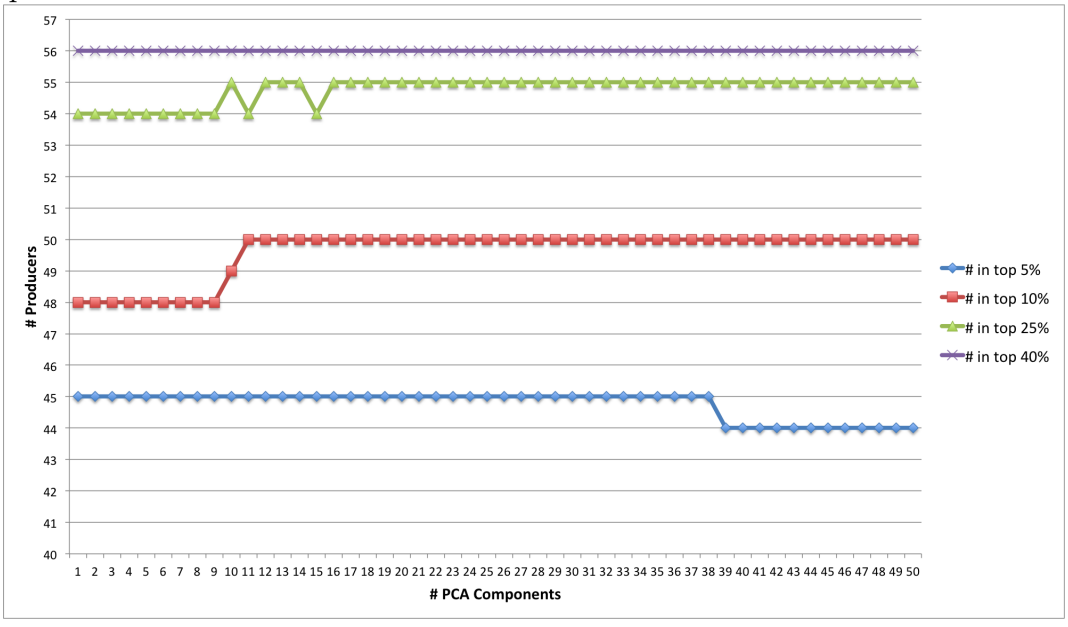
All Forums We first extract the full set of users across all four forums, defining a distinct user to be one with the exact same username. We discard the set of users who never post any content in any of the forums; 13,985 users remained.

Figure 4.5 shows the number of producers that appeared in the top five, 10, 25, and 40% of ranked users, with principal components from one to 50. The number of principal components has the most effect on the number of producers that appear in the top 10% of ranked users, with the number of producers increasing as the number

of principal components increases. For all of the principal component values, all 56 producers appear within the top 40% of ranked users.

The best performing number of principal components seem to be 16 to 45, in which 45, 50, 55 and 56 producers appear in the top five, 10, 25 and 40% of ranked users, respectively. When ranking by number of thanks received, 16 producers appear in the top five percent, 28 appear in the top 10%, 38 appear in the top 25%, and 42 appear in the top 40%.

Figure 4.5. Number of Producers in top % of Ranked Users, with varying Principal Components: All Forums



These results are promising: the vast majority (80%) of known producers appear in the top five percent of ranked users. The baseline only includes 29% of known producers in the top five percent of ranked users. This represents a significant time savings for law enforcement - given this ranking, law enforcement would only need to look through 700 users (rather than potentially the full set of 13,985) to find the significant majority of known producers on the forum.

After extracting this set of ranked users using all four forums, we provided the top 1,000 users to law enforcement for further labeling. Within that 1,000, they tagged 43 additional users (beyond the known producers) as users of known investigative interest. This feedback further demonstrates the promising nature of this work: not only do 47 of the known producers appear in the top 1,000 ranked users, but 5% of the remaining 953 users are also of known investigative interest, representing even more time savings.

4.5 Conclusion

In this chapter, we developed a feature set to enable the automatic ranking of users from Tor CSAM forums, where highly ranked users are anomalous. We tested a PCA based anomaly detection algorithm using our feature set on four different Tor CSAM forums, as well as a combined set of all four forums. We achieved high performance throughout, where performance is measured by percentage of known producers appearing in the top 5%, 10%, 25% and 40% of ranked users. We also verified our results from the combined forums assessment with law enforcement, finding even more users of investigative interest amongst those the algorithm ranked highly. Our tool potentially saves significant time for law enforcement officers in their investigative process, allowing them to focus their initial attention on a smaller subset of users that is likely to contain a disproportionate number of producers, as compared to the rest of the users on these sites. In this field, this type of saved time can mean the difference between a child abused for weeks, vs. years.

Chapter 5

Conclusion

This dissertation presented multiple tools and techniques that can widely be used to analyze, classify and de-anonymize criminal forums and networks online. Across the broad spectrum of criminal activity online, we focused on three main domains of criminal activity on the clear web and the Dark Net: classified ads advertising trafficked humans for sexual services, cyber black-market forums, and Tor onion sites hosting forums dedicated to child sexual abuse material.

In the first domain, we proposed an automated and scalable approach for identifying sex trafficking using multiple data sources. We developed a stylometry classifier and a Bitcoin transaction linking technique to group sex ads by owner. To the best of our knowledge, this is the first such work to attempt to link specific purchases to specific transactions on the Bitcoin blockchain. We evaluated our approach using real world ads scraped from Backpage, and demonstrated that our approach can group multiple ads by their real owners. We are currently collaborating with multiple NGOs and law enforcement officers to deploy our tools to help fight human trafficking.

There are several avenues of approach future work could take. We can work to disambiguate Backpage credit payments on the Bitcoin blockchain from Backpage ad payments by analyzing ads and credit payments we make ourselves. We can show our data to law enforcement officers and work together to build a ground truth set that we can then use to validate or reject the correctness of our exact match transactions. We can use existing Bitcoin clustering techniques to link our Paxful transactions to each other, and then our stylometry model to tie those ads that match the Paxful transactions to the ads that match the transactions made using our persistent Bitcoin wallet. In general, finding more connections between previously unconnected ads - i.e., finding more owners and grouping those ads by owner - is key. If those ads include movement across multiple states/geographic locations, with multiple parties involved, it is highly likely that a trafficker or trafficking ring is responsible.

In the second domain, we built several tools to enable the automatic classification

and extraction of information from underground forums. We can apply our tools across a variety of forums, accommodating differences in language and forum specialization. We tested our tools on 8 different underground forums, achieving high performance both within-forum and across-forum. We also performed two case studies to show how analysts can use these tools to investigate underground forums to discern insights such as the popularity of original vs. bulk/hacked accounts, or what kind of currencies have high demand.

Our tools allow for future researchers to continue this type of large-scale automated exploration to extract a holistic view of a single or several underground forums, as well as potentially provide support to law enforcement investigating cybercrime. Another potentially promising avenue for future work involves analyzing private messages. Typically, the final price for a product or service is settled in private messages between two users, and not in the public forum space. Given these private messages, we can perform an analysis using our price extractor to assess how well the publicly available pricing data reflects the true, final price. Additionally, we can build a tool that uses the public forum data to predict the true, final price indicated in the private messages.

In the final domain, we developed a feature set to enable the automatic ranking of users from Tor CSAM forums, where highly ranked users are anomalous. We tested a PCA based anomaly detection algorithm using our feature set on 4 different Tor CSAM forums, as well as a combined set of all four forums. We achieved high performance throughout, where performance is measured by percentage of known producers appearing in the top 5%, 10%, 25% and 40% of ranked users. We also verified our results from the combined forums assessment with law enforcement, finding even more users of investigative interest amongst those the algorithm ranked highly. Our tool potentially saves significant time for law enforcement officers in their investigative process, allowing them to focus their initial attention on a smaller subset of users that is likely to contain a disproportionate number of producers, as compared to the rest of the users on these sites. In this field, this type of saved time can mean the difference between a child abused for weeks, vs. years.

There is a wealth of possible steps future work could take. We can explore different forms of anomaly detection, beyond PCA. We can continue iterating with law enforcement to iteratively improve the existing model with updated ground truth. We can experiment with incorporating non-public elements of the forums (e.g., private messages and private boards) to see whether features from that data improves performance. We can also expand to a larger set of forums, adjusting the features as is necessary to reflect the different data available across these different forums.

Bibliography

[1] Ltu technologies. `www.ltutech.com`. 2014.

[2] Netclean technologies. `https://www.netclean.com`. 2016.

[3] Stanford POS Tagger. `https://nlp.stanford.edu/software/tagger.shtml`.

[4] The National Report on Domestic Minor Sex Trafficking: America's Prostituted Children. Technical report, Shared Hope International, 2009.

[5] Domestic Minor Sex Trafficking: The Criminal Operations of the American pimp. Technical report, Polaris Project, 2012.

[6] A Picture of Abuse. Technical report, Child Exploitation and Online Protection Centre, 2012.

[7] United Nations Office on Drugs and Crime, Human trafficking indicators. Technical report, United Nations, 2012.

[8] Sadia Afroz, Vaibhav Garg, Damon McCoy, and Rachel Greenstadt. Honor Among Thieves: A Common's Analysis of Cybercrime Economies. In *Proceedings of the eCrime Researchers Summit (eCRS)*, pages 1–11, 2013.

[9] Elli Androulaki, Ghassan Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. Evaluating User Privacy in Bitcoin. In *Financial Cryptography and Data Security: 17th International Conference*, pages 34–51, 2013.

[10] David Bamman, Brendan O'Connor, and Noah A. Smith. Learning Latent Personas of Film Characters. In *Proceedings of ACL*, 2013.

[11] Jamie Bartlett. *The Dark Net*. Random House, 2014.

[12] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting Spammers on Twitter. In *Proceedings of the Conference on Email and Anti-Spam*, 2010.

[13] Kristie R. Blevins and Thomas J. Holt. Examining the Virtual Subculture of Johns. *Journal of Contemporary Ethnography*, 38(5):619–648, 2009.

[14] Vanessa Bouche et al. A Report on the use of Technology to Recruit, Groom and Sell Domestic Minor Sex Trafficking Victims. Technical report, Thorn, 2015.

[15] Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-Based n-gram Models of Natural Language. In *Proceedings of ACL*, volume 18, pages 467–479, 1992.

[16] Danqi Chen and Christopher D Manning. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of EMNLP*, 2014.

[17] Munish Chopra, Miguel Vargas Martin, and Luis Rueda. Toward New Paradigms to Combating Internet Child Pornography. In *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, 2006.

[18] Nicolas Christin. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the International World Wide Web Conference*, 2013.

[19] Hal Daume III. Frustratingly Easy Domain Adaptation. In *Proceedings of ACL*, 2007.

[20] Mateus de Castro Polastro and Pedro Monteiro da Silva Eleuterio. Nudetective: A Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection. In *Proceedings of the IEEE Workshop on Database and Expert Systems Applications*, 2010.

[21] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. Bag-of-Visual-Words Models for Adult Image Classification and Filtering. In *Proceedings of the IEEE 19th International Conference on Pattern Recognition*, 2008.

[22] Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. Leveraging Publicly Available Data to Discern Patterns of Human-Trafficking Activity. *Journal of Human Trafficking*, 1(1):65–85, 2015.

[23] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[24] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. COMPA: Detecting Compromised Accounts on Social Networks. In *Proceedings of the Network and Distributed System Security Symposium*, 2013.

[25] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying Relations for Open Information Extraction. In *Proceedings of EMNLP*, 2011.

[26] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, pages 1871–874, 2008.

[27] Federal Bureau of Investigation, Jacksonville Division. Jury Finds New York Man Guilty of Sex Trafficking Women by Force, Threats of Force, and Fraud, 2010. `http://www.fbi.gov/jacksonville/press-releases/2011/ja021711.htm`.

[28] Margaret M. Fleck, David A. Forsyth, and Chris Bregler. Finding Naked People. In *Proceedings of the European Conference on Computer Vision*, pages 593–602, 1996.

[29] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[30] Jason Franklin, Vern Paxson, Adrian Perrig, and Stefan Savage. An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, pages 375–388, 2007.

[31] Dayne Freitag and Andrew McCallum. Information Extraction with HMM Structures Learned by Stochastic Optimization. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000.

[32] Vaibhav Garg, Sadia Afroz, Rebekah Overdorf, and Rachel Greenstadt. Computer-Supported Cooperative Crime. In *Proceedings of the 19th International Conference on Financial Cryptography and Data Security*, pages 32–43, 2015.

[33] Chad M. Harms. Grooming: An operational definition and coding scheme. *Sex Offender Law Report*, 8(1):1–6, 2007.

[34] Cormac Herley and Dinei Florencio. Nobody Sells Gold for the Price of Silver: Dishonesty, Uncertainty and the Underground Economy. In *Proceedings of the Workshop on Economics of Information Security and Privacy*, pages 33–53, 2009.

[35] Thomas J. Holt and Eric Lampke. Exploring Stolen Data Markets Online: Products and Market Forces. *Criminal Justice Studies*, 23(1):33–50, 2010.

[36] Thorsten Holz, Markus Engelberth, and Felix Freiling. Learning More about the Underground Economy: A Case-Study of Keyloggers and Dropzones. In *Proceedings of the European Symposium on Research in Computer Security*, 2009.

[37] Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls, Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine, Marc Liberatore, Brian Lynn, and Janis Wolak. Measurement and Analysis of Child Pornography Trafficking on P2P Networks. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 631–642, 2013.

[38] Michelle Ibanez and Daniel D Suthers. Detection of Domestic Human Trafficking Indicators and Movement Trends Using Content Available on Open Internet Sources. In *Proceedings of the 47th Hawaii International Conference on System Sciences*, pages 1556–1565, 2014.

[39] Giacomo Inches and Fabio Crestani. Overview of the International Sexual Predator Identification Competition at PAN-2012. *CLEF (Online Working Notes/Labs/Workshop)*, 30, 2012.

[40] Lara Janson, Rachel Durchslag, Heather Mann, Rachel Marro, and Allyson Matvey. "Our Great Hobby": An Analysis of Online Networks for Buyers of Sex in Illinois. Technical report, Chicago Alliance Against Sexual Exploitation, 2013.

[41] Michael J. Jones and James M. Rehg. Statistical Color Models with Application to Skin Detection. *International Journal of Computer Vision*, 46:81–96, 2002.

[42] Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn, and Joseph Le Roux. Foreebank: Syntactic Analysis of Customer Support Forums. In *Proceedings of EMNLP*, 2015.

[43] Ghassan O. Karame, Elli Androulaki, Marc Roeschlin, Arthur Gervais, and Srdjan Čapkun. Misbehavior in Bitcoin: A Study of Double-Spending and Accountability. *ACM Transactions on Information and System Security*, 18(1):2:1–2:32, 2015.

[44] Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and Linking Web Forum Posts. In *Proceedings of CoNLL*, 2010.

[45] April Kontostathis. ChatCoder: Toward the Tracking and Categorization of Internet Predators. In *Proceedings of the SIAM International Conference on Data Mining Text Mining Workshop*, 2009.

[46] Brian Krebs. Cards Stolen in Target Breach Flood Underground Markets. `http://krebsonsecurity.com/2013/12/cards-stolen-in-target-breach\-flood-underground-markets`, 2013.

[47] Brian Krebs. Who's Selling Credit Cards from Target? `http://krebsonsecurity.com/2013/12/whos-selling-credit-cards-from-target`, 2013.

[48] Nicholas Kristof. Every Parent's Nightmare. The New York Times, 2016. `http://www.nytimes.com/2016/03/10/opinion/every-parents-nightmare.html`.

[49] Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, and Dan Klein. An Empirical Analysis of Optimization for Max-Margin NLP. In *Proceedings of EMNLP*, 2015.

[50] Matthieu Latapy, Clemence Magnien, and Raphael Fournier. Quantifying Paedophile Activity in a Large P2P System. *Information Processing and Management*, 49:248–263, 2013.

[51] Mark Latonero. Human Trafficking Online: The Role of Social Networking Sites and Online Classifieds. Technical report, USC Annenberg School for Communication & Journalism, 2011.

[52] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[53] Marco Lui and Timothy Baldwin. Classifying User Forum Participants: Separating the Gurus from the Hacks, and Other Tales of the Internet. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA)*, 2010.

[54] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL: System Demonstrations*, 2014.

[55] India McGhee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra McBride, and Emma Jakubowski. Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce 15*, 3:103–122, 2011.

[56] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, and Stefan Savage. A Fistful of Bitcoins: Characterizing Payments Among Men with No Names. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 127–140. ACM, 2013.

[57] Tyler Moore, Richard Clayton, and Ross Anderson. The Economics of Online Crime. *The Journal of Economic Perspectives*, 23.3:3–20, 2009.

[58] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. An Analysis of Underground Forums. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pages 71–80. ACM, 2011.

[59] NIST. The ACE 2005 Evaluation Plan. In *NIST*, 2005.

[60] Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. Learning to Extract International Relations from Political Context. In *Proceedings of ACL*, 2013.

[61] Department of Justice. Major Computer Hacking Forum Dismantled. `https://www.justice.gov/opa/pr/major-computer-hacking-forum-dismantled`, 2015.

[62] Alexander Panchenkos, Richard Beaufort, and Cedrick Fairon. Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames. In *Proceedings of the LREC Workshop on Language Resources for Public Security Applications*, 2012.

[63] Ankur P. Parikh, Hoifung Poon, and Kristina Toutanova. Grounded Semantic Parsing for Complex Knowledge Extraction. In *Proceedings of NAACL*, 2015.

[64] Claudia Peersman, Christian Schulze, Awais Rashid, Margaret Brennan, and Carl Fischer. iCOP: Automatically Identifying New Child Abuse Media in P2P Networks. In *Proceedings of the IEEE Security and Privacy Workshops*, pages 124–131, 2014.

[65] Nick Pendar. Toward Spotting the Pedophile Telling Victim from Predator in Text Chats. In *Proceedings of the International Conference on Semantic Computing*, 2007.

[66] Polaris. Human trafficking. Polaris Project, 2017. `http://www.polarisproject.org/human-trafficking/`.

[67] Rebecca S. Portnoff, Sadia Afroz, Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Damon McCoy, Kirill Levchenko, and Vern Paxson. Tools for Automated Analysis of Cybercriminal Markets. In *Proceedings of the International World Wide Web Conference*, 2017.

[68] Rebecca S. Portnoff, Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. Backpage and Bitcoin: Uncovering Human Traffickers. In *Proceedings of the International World Wide Web Conference*, 2017.

[69] Nathan J. Ratliff, Andrew Bagnell, and Martin Zinkevich. (Online) Subgradient Methods for Structured Prediction. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.

[70] Dominique Roe-Sepowitz, James Gallagher, Kristine Hickle, and Jessica Smith. One-day sex trafficking snapshot of an internet service provider. Technical report, Arizona State University, 2012.

[71] Dorit Ron and Adi Shamir. Quantitative Analysis of the Full Bitcoin Transaction Graph. In *Proceedings of the 17th International Conference on Financial Cryptography and Data Security*, pages 6–24, 2013.

[72] Henry A. Rowley, Yushi Jing, and Shumeet Baluja. Large Scale Image-Based Adult-Content Filtering. In *Proceedings of the International Conference of Computer Vision Theory and Applications*, pages 290–296, 2006.

[73] Malika Saada Saar. Girl Slavery in America, 2010. `http://www.huffingtonpost.com/malika-saada-saar/girl-slavery-in-america_b_544978.html`.

[74] Asaf Shupo, Miguel Vargas Martin, Luis Rueda, Anasuya Bulkan, Yongming Chen, and Patrick CK Hung. Toward Efficient Detection of Child Pornography in the Network Infrastructure. *IADIS International Journal on Computer Science and Information Systems*, 2:15–31, 2006.

[75] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:1349–1380, 2000.

[76] Kyle Soska and Nicolas Christin. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. In *Proceedings of the 24th USENIX Security Symposium*, pages 33–48, 2015.

[77] Chad M.S. Steel. Child pornography in peer-to-peer networks. *Child Abuse and Neglect*, 33:560–568, 2009.

[78] Chad M.S. Steel. Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms. *Child Abuse and Neglect*, 44:150–158, 2015.

[79] Brett Stone-Gross, Thorsten Holz, Gianluca Stringhini, and Giovanni Vigna. The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns. In *Proceedings of the 4th USENIX Conference on Large-scale Exploits and Emergent Threats*, LEET'11, 2011.

[80] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting Spammers on Social Networks. In *Proceedings of the Annual Computer Security Applications Conference*, 2010.

[81] Mihai Surdeanu. Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling,. In *Proceedings of the TAC-KBP 2013 Workshop*, 2013.

[82] Athanasia Sykiōtou. Trafficking in Human Beings: Internet Recruitment: Misuse of the Internet for the Recruitment of Victims of Trafficking in Human Beings. Technical report, Directorate General of Human Rights and Legal Affairs, Council of Europe, 2007.

[83] Rob Thomas and Jerry Martin. The Underground Economy: Priceless. In *Proceedings of the USENIX Security Symposium*, volume 31, 2006.

[84] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL*, 2003.

[85] Adrian Ulges and Armin Stahl. Automatic Detection of Child Pornography Using Color Visual Words. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2011.

[86] U.S. Department of Justice. Dallas Felon Admits to Sex Trafficking a Minor and Possessing an Assault Rifle, 2011. `http://www.justice.gov/usao/txn/PressRel11/wilson_clint_ple_pr.html`.

[87] U.S. Immigration and Customs Enforcement. Maryland man pleads guilty in sex trafficking conspiracy involving 3 minor girls, 2009. `http://www.ice.gov/news/releases/0907/090716baltimore.htm`.

[88] Bimal Viswanath, M. Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Towards Detecting Anomalous User Behavior in Online Social Networks. In *Proceedings of the USENIX Security Symposium*, 2014.

[89] Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. Predicting Thread Discourse Structure over Technical Web Forums. In *Proceedings of EMNLP*, 2011.

[90] Janis Wolak, Marc Liberatore, and Brian Neil Levine. Measuring a year of child pornography trafficking by us computers on a peer-to-peer network. *Child Abuse and Neglect*, 38:347–356, 2014.

[91] Michael Yip, Nigel Shadbolt, and Craig Webber. Structural Analysis of Online Criminal Social Networks. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, pages 60–65, 2012.

[92] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American society for information science and technology*, 57(3):378–393, 2006.

Appendix A

This Appendix contains detailed tables for each Shared Author wallet, and for each Persistent Bitcoin Identity wallet. For each wallet, the authors grouped together using the relevant methodology are listed. Also included is information about the ads posted by each author: the locations the ads were posted in, the area codes for the phone numbers listed, the post subcategories, and the most frequently occurring demographic tokens. Finally, we also include the total number of ads posted, and dollars spent, for each author.

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805
36660	los angeles	714	f-escorts, m-escorts	girl, latina, thai, japanese, asian	649	$3020
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946
37140	los angeles	562, 714, 523	f-escorts, m-escorts	girl, latina, japanese, thai, young	1237	$5671
72789	NY	917	f-escorts	asian, girl, young	28	$2320
87912	northeast TX	817	f-escorts	hispanic	22	$105
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925

Table A.1. SA Wallet `1A3Bj` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
24770	IL	312	f-escorts	girl, korean, asian, japanese, young	48	$4805
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40

Table A.2. SA Wallet `1Abgk` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110828	sf bay	415	f-escorts	asian, girl, japanese, young, taiwanese	68	$7494.50
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40

Table A.3. SA Wallet `1ASPo` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110828	sf bay	415	f-escorts	asian, girl, japanese, young, taiwanese	68	$7494.50
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
85388	northeast TX	469, 972	f-escorts	girl, asian, japanese, korean, young	116	$4308.50
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
110827	sf bay	415	f-escorts	asian, japanese, girl, korean, female	48	$4111.50
110829	sf bay	415	f-escorts	girl, korean, asian, woman, taiwanese	48	$3612.25
140159	WA	224, 206	f-escorts	girl, young, asian, student, korean	165	$765
24772	IL	630	f-escorts	girl, asian, young	14	$1650
24482	IL	708, 224, 872, 773	f-escorts	girl, asian, young, japanese, chinese	116	$4714

Table A.4. SA Wallet `1BT6w` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110827	sf bay	415	f-escorts	asian, japanese, girl, korean, female	48	$4111.50
139514	WA	253	f-escorts	–	7	$980
139504	WA	425	f-escorts	girl, black, filipino	4	$865

Table A.5. SA Wallet `1D1di` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110827	sf bay	415	f-escorts	asian, japanese, girl, korean, female	48	$4111.50
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445
121231	south FL	754, 461, 305, 954, 247	f-escorts, dom&bdsm	girl, latina, female, woman, spanish	89	$545
105050	sfbay, sacramento	707	f-escorts	asian, girl	87	$415
64038	NV	702	f-escorts	gf	7	$1510
138908	WA	206	f-escorts	girl, asian, young, lady	28	$2325
37140	los angeles	562, 714, 523	f-escorts, m-escorts	girl, latina, japanese, thai, young	1237	$5671
24772	IL	630	f-escorts	girl, asian, young	14	$1650
24482	IL	708, 224, 872, 773	f-escorts	girl, asian, young, japanese, chinese	116	$4714

Table A.6. SA Wallet `1E4RK` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946
109120	san joaquin valley	669, 310, 626, 702	f-escorts	girl, youth	329	$1805
74000	NY	917	f-escorts	asian, girl, girlfriend, student	44	$3470

Table A.7. SA Wallet `1Gsuis` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
118112	southeast TX	832, 713	f-escorts	girl, asian, young, taiwanese, japanese	51	$4229

Table A.8. SA Wallet `1Hre7` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
41870	los angeles	323, 404, 213	f-escorts	young, african, american	3	$930
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185

Table A.9. SA Wallet `1Kh3x` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
121282	south FL	561, 954	f-escorts, m-escorts, adult jobs	girl, young	396	$1172
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946

Table A.10. SA Wallet `1KpyX` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
37140	los angeles	562, 714, 523	f-escorts, m-escorts	girl, latina, japanese, thai, young	1237	$5671
39230	los angeles	626, 424	f-escorts	french, american, girl	6	$1085

Table A.11. SA Wallet `1KtCW` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
110827	sf bay	415	f-escorts	asian, japanese, girl, korean, female	48	$4111.50
73502	NY	929	f-escorts, body rubs	girl, young, asian	39	$2670
110828	sf bay	415	f-escorts	asian, girl, japanese, young, taiwanese	68	$7494.50
64042	NV	702	f-escorts	young	11	$2210
10451	CO	303	f-escorts	latina, girl	9	$855
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946
74000	NY	917	f-escorts	asian, girl, girlfriend, student	44	$3470

Table A.12. SA Wallet `1Kyoc` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
110830	sf bay	408	f-escorts	girl, asian, taiwanese, japanese, young	72	$3483.40
110827	sf bay	415	f-escorts	asian, japanese, girl, korean, female	48	$4111.50
110828	sf bay	415	f-escorts	asian, girl, japanese, young, taiwanese	68	$7494.50
33620	KS, OR, los angeles	323, 585, 713, 503	f-escorts	asian, young, student	X	$999.65
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445
64532	NV	702	–	f-escorts	6	$940
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
24772	IL	630	f-escorts	girl, asian, young	14	$1650
24482	IL	708, 224, 872, 773	f-escorts	girl, asian, young, japanese, chinese	116	$4714

Table A.13. SA Wallet `1LetZ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110830	sf bay	408	f-escorts	girl, asian, taiwanese, japanese, young	72	$3483.40
36585	los angeles	909	f-escorts	girl, asian, chinese, young	193	$765
36660	los angeles	714	f-escorts, m-escorts	girl, latina, thai, japanese, asian	649	$3020
92917	OH	929, 512, 323	f-escorts	girl, asian	20	$940
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
140112	WA	337, 509	f-escorts	girl, asian, girlfriend, japanese	152	$623
69294	NY, MA, NJ	561, 917, 551, 203, 646	f-escorts	girl, brazilian, latinas, venezuelan	128	$1205
109652	san joaquin valley	818	f-escorts	–	3	$165
71536	NY	347	f-escorts, body rubs	young, asian, japanese, girl, korean	105	$1240
74649	NY, NJ	718, 917, 347, 516, 646	f-escorts, adult jobs	girl, lady, young, japanese, asian	40	$1950
37898	los angeles	310, 213	f-escorts	black, american	3	$770
74936	NY	917, 282, 247, 646, 435	body rubs	girl, young	11	$875
72690	NY, NC	929, 347, 353, 136, 919, 917	f-escorts	colombiana, latina, girlfriend, brazilian	187	$3394
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
72181	NY	631, 347, 516	body rubs	girl, woman, latina, dominican	127	$2230
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946
74000	NY	917	f-escorts	asian, girl, girlfriend, student	44	$3470
111826	sf bay	510, 909, 408	f-escorts	girl, asian, chinese, young	16	$1185

Table A.14. SA Wallet `1LYEQ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
73991	NY	646	f-escorts	girl, asian, girlfriend	5	$745
51088	sf bay, MA	209	f-escorts	girl, white	26	$120

Table A.15. SA Wallet `1MGDy` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
118112	southeast TX	832, 713	f-escorts	girl, asian, young, taiwanese, japanese	51	$4229
14950	DC	202	f-escorts	girl, japanese, asian, taiwanese	59	$2660.50
41721	sf bay, los angeles	323, 213	f-escorts	girl, asian	400	$1840
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805

Table A.16. SA Wallet `1MheR` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110830	sf bay	408	f-escorts	girl, asian, taiwanese, japanese, young	72	$3483.40
110828	sf bay	415	f-escorts	asian, girl, japanese, young, woman	68	$7494.50

Table A.17. SA Wallet `1Mufv` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
73502	NY	929	f-escorts, body rubs	girl, young, asian	39	$2670
85391	northeast TX	469	f-escorts	asian, girl, japanese	20	$2097.25
45994	NY, los angeles	929, 225, 646	f-escorts, body rubs	girl, japanese, korean, asian, young	14	$1331
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445
72789	NY	917	f-escorts	asian, girl, young	28	$2320
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946
24482	IL	708, 224, 872, 773	f-escorts	girl, asian, young, japanese, chinese	116	$4714

Table A.18. SA Wallet `1N7V4` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
118112	southeast TX	832, 713	f-escorts	girl, asian, young, taiwanese, japanese	51	$4229
66852	NJ	914, 845, 732, 757	f-escorts	girl, asian, student, young, korean	8	$1285
66789	NY, NJ	201, 845, 215, 415, 732, 303, 848, 929, 347, 914	f-escorts	girl, young, asian, japanese	47	X

Table A.19. SA Wallet `1P7n4` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
14950	DC	202	f-escorts	girl, japanese, asian, taiwanese	59	$2660.50
18491	GA, SC	803	f-escorts, body rubs	american, african, black	8	$286
118111	southeast TX	832, 713	f-escorts	girl, asian, japanese, young, taiwanese	79	$1672
85391	northeast TX	469	f-escorts	asian, girl, japanese	20	$2097.25
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445
4774	NV, san joaquin valley, AZ, los angeles	925	f-escorts	black, girl	12	$55
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805
24482	IL	708, 224, 872, 773	f-escorts	girl, asian, young, japanese, chinese	116	$4714

Table A.20. SA Wallet `1PesE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110828	sf bay	415	f-escorts	asian, girl, japanese, young, woman	68	$7494.50
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445

Table A.21. SA Wallet `1yVFE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
109746	sf bay, san joaquin valley	510	f-escorts	african	6	$20
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
72789	NY	917	f-escorts	asian, girl, young	28	$2320
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445

Table A.22. SA Wallet `12Xis7` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
118112	southeast TX	832, 713	f-escorts	girl, asian, young, taiwanese, japanese	51	$4229
97790	OR	503	f-escorts	asian, girl, young, taiwanese	22	$1390
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
73277	NY	917, 646	f-escorts	student, malaysian, girl, japanese, korean	9	$895
112186	sf bay	628, 669, 408	f-escorts	girl, asian, young, korean	40	$3070
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445
21270	GA, NY, south FL, IL, sfbay, MA, LA	708, 954, 315, 617, 731, 415	f-escorts, body rubs	girl, japanese, asian	728	$3368
72789	NY	917	f-escorts	asian, girl, young	28	$2320
138908	WA	206	f-escorts	girl, asian, young, lady	28	$2325
72181	NY	631, 347, 516	body rubs	girl, woman, latina, dominican	127	$2230
24770	IL	312, 708, 224, 872	f-escorts	girl, korean, asian, japanese, young	48	$4805
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946

Table A.23. SA Wallet `14FCt` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
10038	PA, south FL, KY, upstate NY, IN, NV, southeast TX, central TX	832, 317, 347, 215, 929	f-escorts	girl, asian, japanese, girlfriend	82	$1201
138908	WA	206	f-escorts	girl, asian, young, lady	28	$2325
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185

Table A.24. SA Wallet `14XUU` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
36660	los angeles	714	f-escorts, m-escorts	girl, latina, thai, japanese, asian	649	$3020
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
74649	NY, NJ	718, 917, 347, 516, 646	f-escorts, adult jobs	girl, lady, young, japanese, asian	40	$1950
74243	NY	347	f-escorts, body rubs	young, girl, asian	41	$1165
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
36741	NY, los angeles	714, 786	f-escorts, m-escorts	girl, asian, fella, guy, thai	1262	$5946
39230	los angeles	626, 424	f-escorts	french, american	6	$1085

Table A.25. SA Wallet `16iD4` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64042	NV	702	f-escorts	young	11	$2210
41656	san diego, los angeles	323, 626, 661	f-escorts	girl, ebony, black	28	$240
10317	CO	303, 720	f-escorts	girl, asian, japanese, korean, young	128	$5445

Table A.26. SA Wallet `17idT` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
41721	sf bay, los angeles	323, 213	f-escorts	girl, asian	400	$1840
21270	GA, NY, south FL, IL, sf-bay, MA, LA	708, 954, 315, 617, 731, 415	f-escorts, body rubs	girl, japanese, asian	728	$3368
72690	NY, NC	929, 347, 353, 136, 919, 917	f-escorts	colombiana, latina, girlfriend, brazilian	187	$3394
72789	NY	917	f-escorts	asian, girl, young	28	$2320
72181	NY	631, 347, 516	body rubs	girl, woman, latina, dominican	127	$2230
74000	NY	917	f-escorts	asian, girl, girlfriend, student	44	$3470

Table A.27. SA Wallet `18tTg` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
139378	WA	260, 206	f-escorts	girl, young, asian, lady, japanese	82	$5925
37140	los angeles	562, 714, 523	f-escorts, m-escorts	girl, latina, japanese, thai, young	1237	$5671
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185
85390	northeast TX	214, 469	f-escorts	girl, asian, japanese, young, lady	20	$2238.25
74000	NY	917	f-escorts	asian, girl, girlfriend, student	44	$3470

Table A.28. SA Wallet `194iD` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
110831	sf bay	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40
64042	NV	702	f-escorts	young	11	$2210
24482	IL	708, 224, 872, 773	f-escorts	girl, asian, young, japanese, chinese	116	$4714

Table A.29. SA Wallet `198xk` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
273	RI, DE, CO, WA, KS, HI, TN, IA, NV, ME, MS, NJ, OK, WY, MN, IL, AR, IN, MD, LA, TX, AZ, WI, NY, MI, NC, UT, VA, OR, DC, CT, MT, CA, MA, OH, AL, NH, GA, PA, SD, FL, AK, KY, NE, ID, MO, WV, NM, SC	994, 704, 415, 108, 347, 103, 917, 800, 188, 804	adult jobs, f-escorts, datelines, body rubs, strippers	woman, american, irish, english, italian	687	$1471.65
11540	FL, NV, CA, CO, WA	619	f-escorts	asian, girl, filipina	7	$57.75

Table A.30. PBI Wallet 1 EM `1M58i` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
338	AL, TN, AR	305	f-escorts, body rubs	ebony, african	4	$9.20
64042	NV	702	f-escorts	young	11	$2210

Table A.31. PBI Wallet 1 EM `1EyHa` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
355	RI, DE, CO, WA, KS, HI, TN, IA, NV, ME, MS, NJ, OK, WY, MN, IL, AR, IN, MD, LA, TX, AZ, WI, NY, MI, NC, UT, VA, OR, DC, CT, MT, CA, MA, OH, AL, NH, GA, PA, SD, FL, AK, KY, NE, ID, MO, WV, NM, SC, VE, ND	800, 210, 978	datelines, transsexuals	girl, shemale, lady, asian, baby	403	$87571.24
11642	NY, CA, CO	929, 858	f-escorts	baby, girl, black, lady, young	22	$1775

Table A.32. PBI Wallet 1 EM `1AFN6` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
865	AL, GA, MS, SC	704, 401	f-escorts	–	9	$44
131006	TN	615	body rubs	–	1	$102

Table A.33. PBI Wallet 1 EM `1MZG3` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
1265	MS, GA, NY, MI, NC, SC, OH, AL, LA, MD	404, 910	f-escorts	–	23	$97
26689	IL	414	f-escorts	ebony, young	10	$60

Table A.34. PBI Wallet 1 EM `14eoU` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
2092	AL, TN	901	f-escorts	–	23	$74
50	RI, CO, WA, KS, TN, IA, NV, ME, MS, NJ, OK, MN, IL, AR, IN, MD, LA, ID, AZ, WI, NY, MI, NC, UT, DC, OR, VA, CT, MT, CA, MA, OH, AL, NH, VE, GA, PA, SD, FL, KY, NE, ND, TX, MO, WV, NM SC	888	datelines	girl	54	$690

Table A.35. PBI Wallet 1 EM `16THW` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
4206	TX, NV, ND, DC, AZ	623, 150, 972, 200, 571, 701, 574, 410, 100, 236	f-escorts	china, asian, girl	14	$160
65030	NV	775	f-escorts	ebony, puerto rican	1	$5

Table A.36. PBI Wallet 1 EM `14nrA` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
5779	AZ	602	f-escorts	girlfriend, girl	3	$15
45806	LA, CA	415, 510	f-escorts	black, latina, woman	5	$25
112274	CA	415	f-escorts	woman	7	$56
110623	CA	925	f-escorts	lady	5	$265
86763	TX	214, 682	f-escorts	–	1	$75
128623	FL	727	f-escorts	white	4	$30
112186	CA	628, 669, 408	f-escorts	girl, asian, young, korean	40	$3070
75833	NY	916	f-escorts	–	1	$5
13830	NY, CT, NJ	347, 304	f-escorts	girl, italian	8	$27
36676	CA	714	f-escorts	girl	288	$1320
104392	CA	415, 209, 323, 016, 510, 916	f-escorts	girl, young, baby	85	$1100
41497	CA	347	f-escorts	–	5	$340
11680	CO, NY, MI, UT, IL, CA, NM	510, 562	f-escorts	girl, female, baby, black	22	$218
24770	IL	312, 708, 224, 872	NM, f-escorts	girl, korean, asian, japanese, young	48	$4805
118332	TX	346	f-escorts	young	4	$490
113571	NY, IN, CA	779	f-escorts	girl, asian, baby, malaysian	14	$301
73059	NY	347, 646	f-escorts	latin	1	$25
51967	MA	617	f-escorts	girl	2	$5

Table A.37. PBI Wallet 1 EM `1Nzho` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
6555	AR, MO	417, 702	f-escorts	girl	11	$43.60
99435	PA	646	f-escorts	chocolate	4	$20

Table A.38. PBI Wallet 1 EM `1Ks4n` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
6783	AR	501	f-escorts	black, baby	1	$18.75
4512	AZ, IA, CO, NC, FL, KY, IL, NE, OR, OH	321, 551, 754	f-escorts, body rubs	girl	17	$1168
45454	LA, TX	832	f-escorts	latina, baby, black	3	$637
24573	IL	288, 312	body rubs	african, girl, cougar, american	2	$175
38329	CA	626,	body rubs	asian, woman	2	$780

Table A.39. PBI Wallet 1 EM `1EKEg` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
7528	CA	818	f-escorts	–	8	$136
9929	CO, MI, NC, FL, VA, CT, MO, NV, TX	631, 203, 704, 337, 310, 954, 305, 130, 540, 404, 734, 100, 321, 424, 214	adult jobs, body rubs, f-escorts	black, espanola, african, girl	232	$497
67166	NJ	551	f-escorts, body rubs	woman	5	$275

Table A.40. PBI Wallet 1 EM `1JntX` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
8090	CA	619	f-escorts	young, girl	7	$12
112769	CA	925	f-escorts	asian, girl	2	$285.75
41173	AZ, CA	714, 949	body rubs, f-escorts	asian, filipino	25	$210

Table A.41. PBI Wallet 1 EM `1EChh` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
8353	TX	832, 281, 954	f-escorts	black, girl, young	8	$1760
42348	CA	951	f-escorts	girl	8	$45
41367	CA	909, 213, 773	body rubs, f-escorts	black, babe	8	$95
40618	NV, IL, CA, MA, NJ	201, 305, 661, 818, 727, 470, 702, 916	f-escorts	girl, woman, indian, young	33	$778
105260	CA	916	f-escorts	–	8	$25

Table A.42. PBI Wallet 1 EM `1Aah7m` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
11177	CO	303	body rubs	ladies, oriental, asian	8	$510
51222	MA	617	f-escorts	baby	1	$20
7739	CA	415	f-escorts	young, woman, girl, american	10	$260
48607	MD	443, 202, 410, 419	f-escorts	girl	2	$295
70287	NY, NJ	917	f-escorts	baby, asian, girl, chinese	12	$475
138908	WA	206	f-escorts	girl, asian, young, lady	28	$2325

Table A.43. PBI Wallet 1 EM `18SzY` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
11454	CO	303	f-escorts	babe	1	$152
110830	CA	408	f-escorts	girl, asian, taiwanese, japanese, young	72	$3483.40

Table A.44. PBI Wallet 1 EM `18SDy` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
12923	CT, MD, MA, DC	347, 804	f-escorts	black, girl, indian	14	$56
75941	NY	929	f-escorts	asian, young, girl, japanese	14	$775
139303	WA	425	f-escorts	girl, young, black	4	$500
6715	CA, AR	810, 310	f-escorts	–	3	$65
3408	AZ	200, 100, 520, 480	f-escorts	indian, black, chick, italian, spanish	37	$295

Table A.45. PBI Wallet 1 EM `1DLkr` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
14331	FL	786	f-escorts	transsexual, female	17	$191
104392	CA	415, 209, 323, 016, 510, 916	f-escorts	girl, young, baby	85	$1100

Table A.46. PBI Wallet 1 EM `1FhgN` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
14461	FL	386, 904	f-escorts	american, girl, puerto rican	4	$4
72542	NY	631, 347	f-escorts	girl	5	$295

Table A.47. PBI Wallet 1 EM `1AnJA` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
15416	NY, PA, FL, DC, NJ	305	transsexual	girl, transsexual, american	6	$65
67355	FL, NJ	908, 917	transsexual	girl, boy, transsexual, italian, european	6	$87

Table A.48. PBI Wallet 1 EM `1LGmB` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16758	PA, MD, DE, NJ	302	body rubs	woman	151	$215
46159	los angeles	225	f-escorts, body rubs	girl, asian, young, latina, japanese	19	$776

Table A.49. PBI Wallet 1 EM `14LMur` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16917	PA, DE, NJ	250, 267	adult jobs, f-escorts	lady, female	18	$93.80
75565	NY	917, 718	f-escorts	asian, girl	14	$1480

Table A.50. PBI Wallet 1 EM `1NBrV` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16929	NY, PA, MD, DE	301	f-escorts	–	7	$17
57196	MN	612	f-escorts	lady, young	1	$80
139412	WA	206	f-escorts	girl, asian, japanese, young, baby	12	$905

Table A.51. PBI Wallet 1 EM `1P6eT` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16929	NY, PA, MD, DE	301	f-escorts	–	7	$17
12730	NY, CT	203	body rubs, f-escorts	girl, young, korean, asian	37	$3081.20
93063	OH	216	transsexual	girl, transsexual, woman	5	$6

Table A.52. PBI Wallet 1 EM `1BZ91` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
20064	GA, FL	404	f-escorts	young, girlfriend, spanish	10	$277.70
36676	CA	714	f-escorts	girl	288	$1320

Table A.53. PBI Wallet 1 EM `1DFvc` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
21352	GA	216, 678, 904	adult jobs, f-escorts	baby, girl, lady, women	9	$7
37140	CA	562, 714, 523	m-escorts, f-escorts	girl, latina, japanese, thai, young	1237	$5761

Table A.54. PBI Wallet 1 EM `1MAoG` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
24586	IL	708	m-escorts, f-escorts	masculine, young, nigerian, guy	4	$1.30
43224	CA	530	f-escorts	girl, puerto rican	24	$375

Table A.55. PBI Wallet 1 EM `12r6t` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
25332	NY, IL	630, 847, 247	f-escorts	girl, white	5	$137
73060	NY	424	transsexual	transsexual	2	$60

Table A.56. PBI Wallet 1 EM `158DE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
25530	NY, OH, CA, PA, IL	562, 872, 323, 724, 513, 917	f-escorts, m-escorts, body rubs	latina, puerto rican, black	11	$300
42487	CA	702	f-escorts	ebony	2	$220

Table A.57. PBI Wallet 1 EM `18j9y` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
26115	IL	312	f-escorts	girlfriend	2	$72
41874	NV, CA	714, 909	f-escorts	baby	9	$40

Table A.58. PBI Wallet 1 EM `1K7rX` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
27897	IL	613	f-escorts	–	1	$10
40835	CA	626	f-escorts	girl, korean	375	$1930
66789	NY, NJ	201, 845, 215, 415, 732, 303, 848, 929, 347, 914	f-escorts	girl, young, asian, japanese	47	$4385
35759	CA	909	f-escorts	latina	2	$75
109120	CA	669, 310, 626, 702	f-escorts	girl, youth	329	$1805
62769	NV	702	body rubs, f-escorts	mature, woman	47	$215
68195	NJ	917	f-escorts	asian, japan, girlfriend, young	4	$825
75941	NY	929	f-escorts	asian, young, girl, japanese	14	$775
63603	NV	702	f-escorts	–	2	$115
40835	CA	626	f-escorts	girl, korean	375	$1930
74720	NY, NJ	929	f-escorts	latina, girl, lebanese, freshman, brazilian	20	$40
37788	CA	858, 602	f-escorts	girlfriend, girl	3	$295
75070	NY	929, 917	f-escorts	girl, asian, korean, japan	19	$1962
38536	LA, CA	323	f-escorts	asian	8	$269
50053	MA, NJ	917, 646	body rubs	–	6	$175
90435	FL, SC	352, 864	f-escorts	italian, black, girl, mature	3	$66
16949	PA, DE, VA, NJ	561	f-escorts	woman, italian	17	$38.05
108665	CA	559, 213	f-escorts	latina, girl, baby	7	$40
75941	NY	929	f-escorts	asian, young, girl, japanese	14	$775
95943	PA, OH, VA	614	f-escorts	girl, baby, asian	18	$435

Table A.59. PBI Wallet 1 EM `13Ges` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
31057	IN	765	f-escorts	–	1	$1.25
59679	MO	310	f-escorts	young, puerto rican, girl	7	$86
119296	TX	832, 014, 140	f-escorts	latina, spanish, girl	10	$50
81825	NC, VA	347, 973	f-escorts	baby	4	$102
87847	TX	832, 415, 247, 214, 682	f-escorts, body rubs	black, indian, ebony, baby, girl	22	$282

Table A.60. PBI Wallet 1 EM `1KCJ5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
34887	OH, MI, FL, KY	100, 808, 859	body rubs, f-escorts	black, girl, mature	21	$71
130776	TN	615	f-escorts	–	1	$67

Table A.61. PBI Wallet 1 EM `1GSWt` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
35498	CA	909	f-escorts	woman, black	3	$213
85439	TX	214	body rubs	woman	2	$25

Table A.62. PBI Wallet 1 EM `1DaRLu` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
37773	CA	949	f-escorts	–	3	$590
37773	CA	949	f-escorts	–	3	$590
113573	CA	732	f-escorts	girl	1	$295
36786	CA	949	f-escorts	–	1	$295
37773	CA	949	f-escorts	–	3	$590
36788	CA	949	f-escorts	–	2	$495
36786	CA	949	f-escorts	–	1	$295

Table A.63. PBI Wallet 1 EM `16ZvA` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
41149	CA	310	f-escorts	girl, black	2	$14
9913	OK, TX	316, 817	body rubs, f-escorts	girl, black	9	$167
13548	RI, NH, DE, FL, DC, CT, MA, MD	404, 617	m-escorts, f-escorts	girl, young	32	$171.75
138375	VA	757	f-escorts	–	2	$18

Table A.64. PBI Wallet 1 EM `1Nuf8` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
43274	CA	661	m-escorts	masculine	2	$6
96973	OK	832, 417, 918	f-escorts, body rubs	–	8	$14

Table A.65. PBI Wallet 1 EM `16Yyx` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
44282	NH, CA, MA	916	f-escorts	–	5	$14
40532	CA	410, 323, 310, 415, 520, 510	f-escorts	young, queen, black, baby	3	$10
139585	WA	509	f-escorts	asian, girl, taiwan, vietnam	3	$114
86101	OK, TX	281, 469	f-escorts		5	$183
85902	TX	469	body rubs	espana, girl	7	$170
97157	OK	918	strippers, f-escorts, body rubs	girl, baby	12	$9.50
11185	CO	510, 916	f-escorts	black, girl, young, ebony, baby	4	$130
69696	NY, PA, CT, NJ	203, 202, 339, 213, 323, 267, 412, 646, 484, 231, 480	transsexual	girl, transsexual	29	$261
140349	WA	425	f-escorts	girl	2	$8
118112	TX	832, 713	f-escorts	girl, asian, young, taiwanese, japanese	51	$4229

Table A.66. PBI Wallet 1 EM `1NbRn` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
45178	CA, OR	503	f-escorts, body rubs	girl, young, woman	10	$25
36599	CA	818	bdsm	mistress	2	$135.25

Table A.67. PBI Wallet 1 EM `1Lune` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
45464	LA, TX	504	adult jobs, f-escorts	–	7	$916.25
112000	CA	415, 707	f-escorts	girl, chinese, asian, woman, japanese	21	$1335

Table A.68. PBI Wallet 1 EM `1J1Cc` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
45994	NY, los angeles	929, 225, 646	f-escorts, body rubs	girl, japanese, korean, asian, young	14	$1331
45994	NY, los angeles	929, 225, 646	f-escorts, body rubs	girl, japanese, korean, asian, young	14	$1331

Table A.69. PBI Wallet 1 EM `1N7Af` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
48313	NY, MD	203, 469, 443, 571, 321, 410	f-escorts	girl, baby, feminine, transsexual	11	$14
110831	CA	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40

Table A.70. PBI Wallet 1 EM `13gqd` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
50060	PA, RI, MA	774, 857, 508, 617, 351, 917	m-escorts, strippers, f-escorts, datelines, body rubs	young, girl, egyptian, african, american	32	$265
66460	NJ	973	f-escorts	latina	4	$600
63757	NV	702	f-escorts	–	9	$1625
141487	TX	214, 806	f-escorts	–	2	$3

Table A.71. PBI Wallet 1 EM `19cYw` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
54662	NY, CT, MI	929, 248, 747	f-escorts	asian, girl, young, korean, chinese	13	$184
73991	NY	646	f-escorts	girl, asian, girlfriend	5	$745

Table A.72. PBI Wallet 1 EM `16qR6` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
56512	MN	651	f-escorts	costa rica	2	$20.20
56512	MN	651	f-escorts	costa rica	2	$20.20

Table A.73. PBI Wallet 1 EM `189Bu` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
57948	WI, MN	414, 262, 312, 608, 407, 920, 402, 803, 504, 419	f-escorts, body rubs	black, baby, girl, american, african	39	$345
140156	WA	929, 509, 360, 206	f-escorts	girl, asian, japanese, korean, lady	305	$1167
12694	CT	860, 247, 140, 475	f-escorts	–	21	$68
85390	TX	214, 469	f-escorts	girl, asian, japanese, young, lady	20	$2238.25
105050	CA	707	f-escorts	asian, girl	87	$415
79762	NC	980, 786	f-escorts	female	3	$10
76754	NY, FL	929, 305	f-escorts	young, girl, asian	14	$235

Table A.74. PBI Wallet 1 EM `1EKp8` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
59954	MO	417	body rubs	asian	1	$42
110827	CA	415	f-escorts	asian, japanese, girl, korean, female	48	$4111.50

Table A.75. PBI Wallet 1 EM `13qSM` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
59971	MO	314	body rubs	–	1	$12
38410	CA	323	f-escorts	asian, japanese, girl, young	12	$1050

Table A.76. PBI Wallet 1 EM `1NVVV` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
59987	MO	917, 063, 636, 314	body rubs	asian	16	$411
59987	MO	917, 063, 636, 314	body rubs	asian	16	$411

Table A.77. PBI Wallet 1 EM `1L8rv` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
63097	NV	702	f-escorts	–	2	$220
419	NY, PA, AL, GA, NJ	631, 845, 718, 717, 267, 408	m-escorts, f-escorts	girl, black, white, mexican, indian	27	$124
51384	MA	617	f-escorts	girlfriend, young, asian, girl	114	$550

Table A.78. PBI Wallet 1 EM `14cax` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
63324	NV	424	f-escorts	–	1	$280
104958	CA	916	body rubs, f-escorts	girl, american, baby, african	8	$57

Table A.79. PBI Wallet 1 EM `15ZAE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
63773	NV, CA	702, 408	f-escorts	latina, girl	8	$30.95
36741	NY, CA	786, 714	m-escorts, f-escorts	girl, asian, fella, guy, thai	1262	$5946

Table A.80. PBI Wallet 1 EM `1Fvev` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64174	NV	619	f-escorts	–	3	$685.90
36826	CA	714	body rubs	–	2	$757

Table A.81. PBI Wallet 1 EM `1EUAF` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64220	NV	702	f-escorts	freshman	5	$1720
73721	NY	631, 347, 516	m-escorts, f-escorts	boy	2	$180
72772	NY	347	f-escorts	girl, black	2	$230
124100	FL	916	f-escorts	–	2	$220
41028	CA	562	f-escorts	african, italian	3	$180
72175	NY	718	f-escorts	–	2	$285
63787	NV	725, 702	f-escorts	–	1	$1170
2991	AZ	480	f-escorts	lady, woman	1	$135

Table A.82. PBI Wallet 1 EM `1LDTv` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64869	NV	775	adult jobs, f-escorts	young	6	$118.20
14454	NC, FL, GA, SC	100, 786	f-escorts	cubanita	7	$22

Table A.83. PBI Wallet 1 EM `1Hu8a` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
68274	NY, NJ	917	f-escorts	girl, latina, puerto rican	10	$2105
24485	IL	312	f-escorts	girl, korean, asian	7	$600

Table A.84. PBI Wallet 1 EM `1EF3p` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
70534	NM	615	f-escorts	black	1	$107.60
15684	DC	718	f-escorts	–	1	$610

Table A.85. PBI Wallet 1 EM `1HEDb` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
70608	NM, TX	281, 702	body rubs, f-escorts	girl	14	$150
105836	CA	323	f-escorts	girl	4	$35
43322	CA	559	f-escorts	young, lady, female, latina	16	$30
41538	CA	323, 619	f-escorts	girlfriend, cuban	10	$325
87054	TX	469	f-escorts	–	1	$10

Table A.86. PBI Wallet 1 EM `1JWm4` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
70746	NM, WI, IL	609, 608, 518, 224, 409, 408	f-escorts	girl, singapore, asian, girlfriend	17	$654
53728	MI	734	body rubs	–	2	$3
1643	RI, CO, AK, TN, NV, ME, MS, NJ, OK, DE, AR, IN, LA, TX, KS, IO, NY, MI, NC, CT, MT, CA, MA, OH, NH, GA, PA, FL, HI, KT, NE, ND, AZ, MS, WV, AL	100	adult jobs	girl, female, guy, male	115	$755.58

Table A.87. PBI Wallet 1 EM `18FBc` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
71647	NY	917, 212	body rubs, f-escorts	young, asian, latina, girl, thai	18	$971
85268	TX	214	body rubs	asian, lady	14	$155
36660	CA	714	m-escorts, f-escorts	girl, latina, thai, japanese, asian	649	$3020

Table A.88. PBI Wallet 1 EM `15kcN` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
72448	NY, PA	631, 404, 517, 516	m-escorts, f-escorts	young, italian, spanish, girl, dominican	27	$588
38954	CA	209, 909	f-escorts	black, baby, latina, african	13	$1433

Table A.89. PBI Wallet 1 EM `1L1Pd` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
72618	NY	347	body rubs, f-escorts	girl	30	$2450
118607	TX	281	f-escorts, body rubs	asian, girl	9	$2770
110831	CA	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40

Table A.90. PBI Wallet 1 EM `1Jjm5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
72625	NY	347, 718	f-escorts	girl, hungarian, ukranian, slovakian, russian	2	$335
33993	KY	502	f-escorts, body rubs	girl	2	$28
41911	CA	714	body rubs	girl, european	3	$1480
3639	AZ	480	f-escorts	black	1	$45
63060	NV	702	body rubs, f-escorts	girl, asian	12	$200
18097	GA	470, 678	f-escorts	–	21	$55
87713	TX	520, 469	f-escorts	–	2	$560
121994	FL	786	f-escorts	–	1	$380
41028	CA	562	f-escorts	african, italian	3	$180
41028	CA	562	f-escorts	african, italian	3	$180
11642	NY, CA, CO	929, 858	f-escorts	baby, girl, black, lady, young	22	$1775
39971	CA	657	f-escorts	girl	5	$60
41028	CA	562	f-escorts	african, italian	3	$180
27286	CA, IL	323	f-escorts	girl	3	$140

Table A.91. PBI Wallet 1 EM `1M4aa` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
73285	NY	929, 917, 347	f-escorts	lady, chocolaty, older	5	$335
14025	CT, NJ	908, 732	f-escorts	milf, woman, matured	12	$106

Table A.92. PBI Wallet 1 EM `19mbJ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
73662	NY	646	body rubs, f-escorts	girl, asian, young	23	$1920
66713	NY, NJ	929	f-escorts	girl, asian, korean	12	$819

Table A.93. PBI Wallet 1 EM `1573u` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
73727	NY	929	f-escorts	girl, indian, panamanian	11	$165
111826	CA	510, 909, 408	f-escorts	girl, asian, chinese, young	16	$1185

Table A.94. PBI Wallet 1 EM `16CmR` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
73944	NY	718	body rubs	asian	1	$815
45848	los angeles	872, 609, 317, 701, 404, 720, 225, 504, 424, 804	f-escorts	girl, asian	3	$1515

Table A.95. PBI Wallet 1 EM `1BPCN` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
74554	NY	704	m-escorts, f-escorts	masculine, dude	3	$4
41027	CA	310	bdsm	–	1	$9

Table A.96. PBI Wallet 1 EM `1N4FE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
74582	NY	929, 917	body rubs	–	12	$618
53079	MI	586	transsexual, body rubs	girl, feminine, transsexual	8	$144

Table A.97. PBI Wallet 1 EM `15Ztx` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
75070	NY	929, 917	f-escorts	girl, asian, korean, japan	19	$1962
69696	NY, PA, CT, NJ	203, 202, 339, 213, 323, 267, 412, 646, 484, 231, 480	transsexual	girl, transsexual	29	$261
130414	TN	901	f-escorts	–	3	$39
21317	NC, GA, SC	504	f-escorts	–	3	$29

Table A.98. PBI Wallet 1 EM `1DqxW` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
79927	NC	919	f-escorts, body rubs	mature	4	$84
112606	CA	510, 100, 929, 669, 646	f-escorts	girl, young, asian, korean, japanese	41	$1280

Table A.99. PBI Wallet 1 EM `1Fv7D` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
84289	NC	910	f-escorts	–	2	$11.90
98830	OR	971	f-escorts	girl, young	2	$30

Table A.100. PBI Wallet 1 EM `1FpkQ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
85369	TX	469	f-escorts	latina	3	$90
111884	CA	678, 213, 475	f-escorts	girl, woman, latina	8	$155
24770	IL	312, 708, 224, 872	CA, f-escorts	girl, korean, asian, japanese, young	48	$4805
79749	NC	910	f-escorts, body rubs	young, girl	9	$25

Table A.101. PBI Wallet 1 EM `166vQ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
85388	TX	469, 972	f-escorts	girl, asian, japanese, korean, young	116	$4308.50
16309	PA, MN, DC, NY, MO, IN, IL	920	f-escorts	girl	10	$192.75
63776	NV, WA	253	body rubs, f-escorts	young, lady, hispanic, girl	34	$91.20

Table A.102. PBI Wallet 1 EM `1JY63` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
86101	OK, TX	281, 469	f-escorts	girl, latin, espanol	5	$183
68033	PA, NJ	732	m-escorts, f-escorts, bdsm	boy, women, men, female, asian	7	$226.95

Table A.103. PBI Wallet 1 EM `1Amh5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
86108	TX	214, 469	f-escorts	–	8	$705
9910	TX	424	f-escorts	baby	4	$112
41411	CA	909	adult jobs, f-escorts	ebony, girl	12	$300
36335	CA	626	f-escorts	–	1	$296.75
39762	CA	909	f-escorts	woman, young	3	$150
9222	NY, TX, NJ	512	f-escorts	–	5	$35
12145	CO, TX	803, 720, 816, 719, 817	m-escorts, adult jobs, f-escorts, datelines, body rubs	girl, lady	37	$189.15
6286	AR	832, 972, 870, 956, 732, 479	f-escorts	–	6	$77
63040	NV	657, 232, 702, 323	f-escorts, strippers	mature, milf, girlfriend	21	$6106.60
108825	CA	209	f-escorts	young, baby	9	$51.80
62641	NV	007, 702	body rubs, f-escorts	girl, asian	5	$410
63075	NV	702	body rubs, f-escorts	girl	4	$855
113947	CA	562	f-escorts	girl	3	$40
87608	TX	469, 432	f-escorts	woman, babe	7	$586
111077	CA	408	f-escorts	lady	15	$1515.65
73344	NY	917	f-escorts	young, girl	21	$1795
58136	MN	612, 218	f-escorts, body rubs	latina, girl	2	$5
71416	NY	929, 347	body rubs	russian, babe, girl	5	$1005
37824	CA	510, 714	f-escorts	mexican, girl	9	$210
2029	AL, GA, NC	901	f-escorts	girl	4	$68
63140	NV	702	body rubs	–	2	$115
63112	NV	702	f-escorts	girl	2	$355
107116	CA	858	f-escorts	–	2	$205
86528	TX	210	f-escorts	–	2	$70
73923	NY	917	f-escorts	–	1	$220
63112	NV	702	f-escorts	girl	2	$355
12914	RI, NH, CT	860, 702	f-escorts	asian, woman, girl, lady	5	$90
71404	NY	718	body rubs	girl, latina, asian, middle eastern, young	7	$1695
63159	NV	702	f-escorts	–	1	$30

Table A.104. PBI Wallet 1 EM `168GD` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
86486	TX	972	transsexual	transsexual, chica	1	$51
33620	KS, CA, OR	323, 585, 713, 503	f-escorts	asian, young, student	18	$999.65

Table A.105. PBI Wallet 1 EM `14BJR` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
86703	FL, TX	786	f-escorts	black, venezuela, latina, negro	4	$85
74575	NY	310	f-escorts	–	1	$530
38416	CA	562, 714	f-escorts, body rubs	young, thai, korean, white, latinas	8	$1820
110827	CA	415	f-escorts		48	$4111.50
123265	FL	786	body rubs	puerto rican	2	$185
121049	FL	954	body rubs, f-escorts	latina, girl	12	$1520
71417	NY	917, 347	body rubs	russian, female, babe	5	$965
73360	NY	646	body rubs	girl, young	7	$820

Table A.106. PBI Wallet 1 EM `1P6DB` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
91088	FL	850	f-escorts	–	1	$36
87029	TX	904	f-escorts	girlfriend	2	$90

Table A.107. PBI Wallet 1 EM `1KFPo` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
91750	OH	513, 937	body rubs, f-escorts	female	2	$782
98156	WA, OR	206	f-escorts	young, puerto rican, italian	3	$145
38411	CA	323	f-escorts	asian, girl, korean	12	$970
112689	CA	415	f-escorts	dutch, brazilian, african, russian	4	$105
15128	DC	872	f-escorts	–	5	$80
67597	PA, NJ	267	f-escorts	girlfriend	4	$390
19090	GA, FL	478	f-escorts	white	6	$175
72281	NY	516, 646	f-escorts	woman, babe	3	$600

Table A.108. PBI Wallet 1 EM `1EerL` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
93063	OH	216	transsexual	girl, transsexual, woman	5	$6
42549	CA	347	m-escorts	–	2	$4

Table A.109. PBI Wallet 1 EM `1JCre` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
93273	OH	323	m-escorts	black, masculine	2	$25.70
71389	NY	718	body rubs	–	1	$114
108183	CA	559	body rubs	girl	1	$57
206	AL	850	f-escorts	black, hispanic, baby	7	$62

Table A.110. PBI Wallet 1 EM `133b7` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
98367	OR	971	f-escorts	girl, woman	11	$250
110831	CA	408	f-escorts	girl, japanese, asian, young, taiwanese	85	$9348.40

Table A.111. PBI Wallet 1 EM `1CDaj` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
100788	PA	717, 631, 917	f-escorts	girl, asian, baby, young, korean	6	$132
4707	AZ	480	f-escorts	–	4	$24

Table A.112. PBI Wallet 1 EM `1Jsgm` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
103841	CA	916	f-escorts	chica	5	$18
715	TN, GA, FL	209	f-escorts	–	7	$6

Table A.113. PBI Wallet 1 EM `16Syp` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
103887	CA	916	f-escorts	girl, immature	16	$221
16447	NC, DC	980	f-escorts	italian, portuguese, girl	25	$120
6331	AR, MO	510	f-escorts	girl	3	$146
8227	TX	254	f-escorts	–	2	$5

Table A.114. PBI Wallet 1 EM `13dKY` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
104638	CA	916	f-escorts	young	19	$42
42837	CA	949	f-escorts	young, black	1	$5
104638	CA	916	f-escorts	young	19	$42
7053	GA, AR, TX	903, 857, 901, 305, 202, 502, 321	f-escorts	girl, baby, young, black	5	$195
104638	CA	916	f-escorts	young	19	$42
43145	CA	415	f-escorts	young, woman, puerto rican	7	$185

Table A.115. PBI Wallet 1 EM `1LSju` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
105276	CA	530	f-escorts	–	17	$130
10127	TX	832, 205, 210, 702, 254, 361	f-escorts	ebony, lady, girl, black, young	18	$500
63536	NV, UT	609, 702	f-escorts	girl, brazilian	2	$20

Table A.116. PBI Wallet 1 EM `17Xoc` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
106462	CA	858	body rubs	girlfriend, thai	1	$20
52717	MI	305	strippers, transsexual	italian, ladyboy, shemale	13	$145.75
10649	CO	702, 970	f-escorts	lady, girl	5	$23
74148	NY	646	body rubs	russian, siberia	2	$445
54671	MI, TN, TX	504	f-escorts	babe	3	$85

Table A.117. PBI Wallet 1 EM `1Je2z` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
106865	CA	408	f-escorts	latin, mature	5	$72
14541	FL	619	f-escorts	–	4	$5

Table A.118. PBI Wallet 1 EM `1JgQK` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
108150	CA	661	adult jobs, f-escorts, body rubs	female	11	$71.20
74304	NY	929, 247	f-escorts	girl, young, girlfriend, japanese, korean	4	$1270

Table A.119. PBI Wallet 1 EM `19TW2` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
108594	CA	559	adult jobs, strippers	lady, chica	13	$124.90
54669	MI, TN, TX	334	f-escorts	–	4	$110
118254	TN, TX	832	f-escorts	american, african	4	$390
38360	CA	626	f-escorts	girl, asian	4	$3140
74381	NY	347	f-escorts	espana	1	$5
121833	FL	786	f-escorts	black	2	$140

Table A.120. PBI Wallet 1 EM `1LxPF` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
111051	CA	405, 408	adult jobs, body rubs, f-escorts	latina, asian, young, korean, girl	4	$714.20
100911	PA	410	f-escorts	mediterranean, romanian, girl, italian, black	3	$80
67046	NY, NJ	908	f-escorts, body rubs	latina, russian, thai	35	$3816.90
73714	NY	917	body rubs	russia	1	$510
66390	NJ	732	body rubs	–	2	$790

Table A.121. PBI Wallet 1 EM `1AZ6T` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
112236	CA	925	body rubs	–	1	$720
26483	IL	224	f-escorts	baby	1	$35

Table A.122. PBI Wallet 1 EM `17dKq` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
113046	CA	415	f-escorts	latina, woman	9	$40
99439	PA	215, 267	f-escorts	ebony, chocolate, girl	15	$752
118112	TX	832, 713	f-escorts	girl, asian, young, taiwanese, japanese	51	$4229
64040	NV	702	f-escorts	–	10	$600
71085	NY	518	adult jobs, f-escorts	girl, latina, girlfriend	267	$3792
74243	NY	347	body rubs, f-escorts	young, girl, asian	41	$1165

Table A.123. PBI Wallet 1 EM `1NJA5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
117865	SD	915, 725	f-escorts	girl	1	$56
75376	NY	646	f-escorts	student, japan, girl, korea	8	$1200

Table A.124. PBI Wallet 1 EM `189T` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
118638	TX	281	f-escorts	–	3	$120
24932	IL	312, 708, 773	f-escorts	young	13	$430
73493	NY	917, 646	body rubs, f-escorts	young, girl, asian, japanese, chinese	28	$2070
24481	IL	773	f-escorts	girl, asian, taiwanese	13	$1593

Table A.125. PBI Wallet 1 EM `1Hjq` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
119472	TX	337	f-escorts	girl, young	1	$493.80
113573	CA	732	f-escorts	girl	1	$295

Table A.126. PBI Wallet 1 EM `19S7` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
121340	FL	954	body rubs	woman	1	$1380
11642	NY, CA, CO	929, 858	f-escorts	baby, girl, black, lady, young	22	$1775

Table A.127. PBI Wallet 1 EM `19HaT` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
122855	FL	786	f-escorts	cubanita, cubana	6	$80
140001	WA	509	f-escorts	young, girl, asian	11	$18

Table A.128. PBI Wallet 1 EM `1MXgv` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
124858	IA, FL, ME	929	f-escorts	girl, asian, japan, korean	10	$314
110830	CA	408	f-escorts	girl, asian, taiwanese, japanese, young	72	$3483.40
74163	NY	929	body rubs, f-escorts	japanese, young, korean	33	$1535

Table A.129. PBI Wallet 1 EM `1789c` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
127574	FL	941	f-escorts, body rubs	latina	8	$910
10961	NV, CO	720	f-escorts	–	7	$100

Table A.130. PBI Wallet 1 EM `1BTZ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
129981	TN	901	f-escorts	girl	1	$79.65
43629	CA, FL	469, 786, 973	transsexual, f-escorts	transsexual, men, mocha, girl, shemale	12	$190.40

Table A.131. PBI Wallet 1 EM `1Gb3` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
130221	TN	678, 757, 305, 662	f-escorts	black, girl	3	$28.75
23637	HI	808	body rubs, f-escorts	girl, young, asian	118	$309

Table A.132. PBI Wallet 1 EM `1DPE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
130511	TN	786, 347, 901	f-escorts	young, lady	6	$36
67964	NJ	609	f-escorts	woman, young	1	$15
16986	PA, DE	832, 702	f-escorts	girl, baby, korean, asian, japan	13	$119

Table A.133. PBI Wallet 1 EM `14y1o` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
131868	TN	832, 346	f-escorts	girl	3	$42
50528	NH, MA	401	f-escorts	girl, indian	5	$332

Table A.134. PBI Wallet 1 EM `135S` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
135481	UT	801	body rubs	lady, latina	1	$115
76331	NY	347	f-escorts	european	1	$325
121323	FL	702	f-escorts	girl	5	$235

Table A.135. PBI Wallet 1 EM `14ywq` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
135639	UT	435	f-escorts	mature	1	$90
24601	IL	630	f-escorts, body rubs	girl, russian, spanish, polish	2	$1260
62662	NV	702	body rubs, f-escorts	asian, girl, japanese	21	$3675

Table A.136. PBI Wallet 1 EM `1ExUN` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
135664	UT	786	f-escorts	–	2	$232
110829	CA	415	f-escorts	girl, korean, asian, woman, taiwanese	48	$3612.25
4318	AZ, TX	915, 480, 602	body rubs, f-escorts	puerto rican, girl, mexican, young, black	11	$131
54959	MI	269, 248	adult jobs, f-escorts, body rubs	young	31	$115

Table A.137. PBI Wallet 1 EM `1GEDX` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
135664	UT	786	f-escorts	–	2	$232
24770	IL	312, 708, 224, 872	UT, f-escorts	girl, korean, asian, japanese, young	48	$4805

Table A.138. PBI Wallet 1 EM `1DvN` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
135966	UT	801	f-escorts, body rubs	–	2	$175
138909	WA	425, 206	f-escorts	girl, asian, young, japanese, korean	82	$4185

Table A.139. PBI Wallet 1 EM `1HLqs` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
135987	UT	402	f-escorts	–	1	$172.80
66789	NY, NJ	201, 845, 215, 415, 732, 303, 848, 929, 347, 914	f-escorts	girl, young, asian, japanese	47	$4385
105811	CA	209	f-escorts	–	1	$5
109652	CA	818	f-escorts	–	3	$165

Table A.140. PBI Wallet 1 EM `1J3t` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
137443	VA	757	f-escorts	woman	2	$10
13718	RI, CT, MA	774, 310, 617	adult jobs, f-escorts, body rubs	female, woman, asian	158	$877.20

Table A.141. PBI Wallet 1 EM `1NmYX` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
139545	WA	425	adult jobs, f-escorts	baby	18	$36
117952	TX	832	f-escorts, body rubs	chinese, girl	13	$180

Table A.142. PBI Wallet 1 EM `1NT5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
10	CO	–	bdsm	–	1	$15
10	CO	–	bdsm	–	1	$15

Table A.143. PBI Wallet 1 EM `1H5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
62849	NV	702	body rubs, f-escorts	black, indian, milf, women	4	$1555
69662	NY, NJ	414, 347, 484, 608, 929	m-escorts, f-escorts	baby, dominicana, jamaica, puerto rico, girls	82	$1545
72575	NY	212	body rubs	girl, chinese, japanese, thai, korean	20	$1323
122696	FL	824	body rubs	–	3	$329.50

Table A.144. PBI Wallet MEM `14psc` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
239	CO, WA, KS, HW, TN, IO, NV, ME, MS, NJ, OK, DE, MN, IL, AK, IN, MD, LA, ID, AZ, WI, NY, MI, NC, UT, DC, OR, VA, CT, MT, CA, MA, OH, AL, NH, VT, GA, PA, SD, FL, AL, KY, NE, ND, TX, MO, WV, NM	678, 213, 194, 712, 940, 121, 167, 171	datelines	gay, adults, bisexual	276	$1052.41
8531	RI, CO, WA, TN, IO, NJ, OK, WY, MN, IL, IN, MD, TX, WI, NY, MI, NC, DC, VA, CT, MT, CA, MA, GA, PA, FL, KY, NE, ND, OH	866	datelines	young, sister, daughter, mommy	45	$72.37
51248	MA	508	f-escorts	–	15	$43.50

Table A.145. PBI Wallet MEM `1LzN` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
100009	PA	484	f-escorts	caucasian, woman	3	$288
137927	VA	305	f-escorts	girl	2	$14

Table A.146. PBI Wallet MEM `1H4h` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
1949	AL, VA	757	body rubs	adult, mature, ethnic	2	$17
46395	NV, LA, VA	773	bdsm	–	2	$6.15
60421	MO	314	f-escorts, body rubs	–	2	$294
60742	TN, MO	901	f-escorts	baby, hispanic	11	$40
66777	NY, NJ	201, 973, 862, 929, 347, 475, 917, 516	adult jobs, f-escorts	girl, dominican	11	$85

Table A.147. PBI Wallet MEM `1GWt` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
81715	NC	704, 980	body rubs	latina, asian, caucasian, girl	6	$4152.40
121258	FL	300, 200	f-escorts	asian, girl	1	$90

Table A.148. PBI Wallet MEM `1DCQ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
125216	FL	904	f-escorts	–	1	$3
137514	VA	757, 804	f-escorts	–	23	$95

Table A.149. PBI Wallet MEM `13DE` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
18886	GA, SC	404	f-escorts	–	2	$2
124214	FL	786	adult jobs	chicas	1	$24

Table A.150. PBI Wallet MEM `1CJy` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
7881	CA	323, 510	f-escorts	girl, woman	6	$118
32175	IA	641, 702	f-escorts	young, girl	6	$5
89799	FL	252, 352	f-escorts	ebony, jamaican, woman	19	$319

Table A.151. PBI Wallet MEM `1BBey` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
65580	NV	702	strippers	girlfriend	6	$80
70943	NM, TX	210	body rubs, t-escorts	girl, transsexual	14	$261.55

Table A.152. PBI Wallet MEM `15Gd` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
2994	AZ	415	f-escorts	–	4	$42
73820	NY	646	body rubs, f-escorts	asian, girls, young	28	$2045

Table A.153. PBI Wallet MEM `16pp` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16211	NY, DC	305	f-escorts	girl	2	$40
24399	IL	708	f-escorts	russian	13	$263
42755	CA	323	f-escorts	chinese, young	3	$175

Table A.154. PBI Wallet MEM `18zU` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
40978	CA	702	m-escorts	thai	3	$3
54236	MI	734	f-escorts, body rubs	chinese, girls, young	3	$180

Table A.155. PBI Wallet MEM `1Fox` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
4563	AZ, TX	201	f-escorts	colombian, girl	7	$53
12931	NY, CT, MA, NJ	914, 732, 973, 305, 508	f-escorts	girl, latina, brazilian	18	$394

Table A.156. PBI Wallet MEM `1BQ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16986	PA, DE	832, 702	f-escorts	girl, korean, lesbian, asian, japan	13	$119
112220	CA	510	f-escorts	mama, baby, lady	9	$48.30

Table A.157. PBI Wallet MEM `1KXo` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
11002	CO	719	body rubs	young	11	$541
43194	CA	626	body rubs	girl, asian, young	5	$360

Table A.158. PBI Wallet MEM `1Hyt` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
17338	PA, DE	610, 484, 267	f-escorts	girl	6	$193
121697	FL	561	f-escorts	girl	6	$134.40

Table A.159. PBI Wallet MEM `17uj` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
48317	MD, NJ	443	f-escorts, body rubs	–	5	$205
107111	CA	858	f-escorts	–	2	$405

Table A.160. PBI Wallet MEM `1FUk` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
16555	GA, NY, NC, DC, SC, MD	247, 580	m-escorts, f-escorts	young, cuban	11	$23
24460	IN, TN, MN, IL	734	f-escorts	–	6	$8

Table A.161. PBI Wallet MEM `1Nrs` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
12550	NV, CA, CO	707	body rubs, f-escorts	portuguese	11	$38
49912	MA	617, 866	adult jobs, strippers	girl, lesbian	7	$43.50

Table A.162. PBI Wallet MEM `16L5` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
196	CO, WA, TN, NV, ME, MS, NJ, OK, DE, MN, IL, AK, MD, LA, TX, AZ, NY, MI, KS, DC, OR, CT, MO, CA, MA, OH, AL, NH, VE, PA, SD, FL, AL, KY, ND, MO, WV, NM, SC	888, 866	datelines	mommy, milf, girlfriend, woman, daddy	94	$130.79
27575	NY, CA, MA, IL	202, 646	m-escorts	masculine, american, german, latin, french	9	$104

Table A.163. PBI Wallet MEM `1NYR` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64010	NV	702	f-escorts	–	7	$4055
66875	NJ	201	f-escorts	baby, girl	6	$505
69324	NY, NJ	442, 631	f-escorts	girl,american, young, daughter	43	$2846
72625	NY	347, 718	f-escorts	girl, hungarian, ukrainian, slovakian, russian	2	$335

Table A.164. PBI Wallet MEM `1AP1` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
4649	AZ	480	adult jobs, datelines, strippers	girl	67	$32
10474	KS, CO, MT	316	f-escorts	–	4	$15

Table A.165. PBI Wallet MEM `1K2J` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
17803	GA	917	t-escorts	transexual, girl, white	7	$273
47681	RI, NH, VT, NJ, NY, DE, FL, IL, PA, CT, MA, MD, ME	151, 216, 016, 877, 617, 518, 508, 802, 250, 375	adult jobs, strippers	girl, lesbian, babe, female, lady	322	$475.20

Table A.166. PBI Wallet MEM `1Dja` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
81465	NC	919	strippers	lady, girl	1	$4
100184	PA	607, 570	f-escorts	girl, lady	10	$6

Table A.167. PBI Wallet MEM `1B6G` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
36677	CA	323	f-escorts	young	1	$200
39708	NV, CA	323	f-escorts	girl, girlfriend	4	$335

Table A.168. PBI Wallet MEM `1BdD` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64403	NV	702	f-escorts	woman, lady	10	$20
67339	NY, PA, NJ	845, 718, 732, 929, 347, 646, 917, 484	f-escorts	girl, young, japanese, asian, korean	44	$4201
111772	FL, CA	650	f-escorts	girl	4	$835

Table A.169. PBI Wallet MEM `18Vf` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
123207	FL	305	f-escorts	–	1	$380
138783	WA	425	f-escorts	milf, girl, momma	22	$535

Table A.170. PBI Wallet MEM `1Ng` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
2518	AK	907, 754	f-escorts	girl	12	$6
7469	CT, VT, MA, ME	201	f-escorts	–	4	$13
13538	CT	203	f-escorts	feminine, latina	1	$2

Table A.171. PBI Wallet MEM `1Gvn` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
11555	CO	970	f-escorts	milf, young	7	$475
136120	UT	801	f-escorts	–	1	$188.40

Table A.172. PBI Wallet MEM `1LYx` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
3245	AZ, NJ, NY, WA, DE, FL, DC, IL, PA, CA, NE, TX	247, 888	datelines	latina, girl, baby, babe, girlfriend	23	$76.97
38750	CA	805	m-escorts	men	2	$16

Table A.173. PBI Wallet MEM `13jV` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
21950	GA	678, 912	f-escorts	girl	3	$20
81774	NC, SC	919	f-escorts	girl	10	$66

Table A.174. PBI Wallet MEM `1FUv` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
13891	CT	860	f-escorts	asian, girl, girlfriend	7	$14
114318	FL, CA	347	m-escorts	puerto rican, young	14	$48
137320	VA	707, 757	t-escorts	girl, tgirl	9	$13

Table A.175. PBI Wallet MEM `1Fs4` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
66702	CT, NJ	203, 862, 609, 305, 508, 914	f-escorts	lady, american, italian	5	$73
137094	VA	434	f-escorts	brazilian, woman	2	$2

Table A.176. PBI Wallet MEM `1K2H` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
49582	MD	240	f-escorts	–	2	$5
54132	MI	313	f-escorts	young, girl	3	$56
75933	NY, PA	917	m-escorts	boy	2	$41

Table A.177. PBI Wallet MEM `15db` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
85292	TX	469	body rubs	latina	3	$115.75
118990	TX	832	f-escorts, body rubs	female	2	$930

Table A.178. PBI Wallet MEM `1BrD2` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
74237	NY	917	f-escorts	girl	18	$2557
98449	OR	775, 503	f-escorts	–	5	$567.80
122703	FL	321, 267, 954	f-escorts	–	5	$1415

Table A.179. PBI Wallet MEM `197Z` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
39264	CA	747	adult jobs, t-escorts	girl, transsexual, boy	9	$18
49776	MA	617, 978	f-escorts, body rubs	italian	7	$235.25
74830	NY	646	m-escorts	guy, jamaican	6	$47

Table A.180. PBI Wallet MEM `1AA` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
62667	NV	702	body rubs	–	6	$125
63073	NV	702	f-escorts	girl, french	3	$660

Table A.181. PBI Wallet MEM `1A6M` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
64380	NV	702	f-escorts	–	1	$20
92351	OH	513, 859	f-escorts	girl, white	3	$693

Table A.182. PBI Wallet MEM `1P5Z` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
67094	NY, NJ	646	f-escorts	girl, asian, babe, korean	17	$878
76468	NY	718	body rubs	girl, young	19	$300

Table A.183. PBI Wallet MEM `1PeV` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
46067	los angeles	623, 561, 720	f-escorts	girl, brazilian, black	3	$57
75647	NY	917	f-escorts	girl	6	$1305
127520	FL	102, 239	body rubs	asian, lady	8	$564

Table A.184. PBI Wallet MEM `1KUT` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
15810	DC, VA	757	m-escorts, f-escorts	–	4	$51.30
34403	KY	859	f-escorts	–	1	$4

Table A.185. PBI Wallet MEM `1JNQ` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
12323	CO	719	body rubs	–	2	$28
51243	RI, NH, MA	917	f-escorts	woman, girl, white	10	$18

Table A.186. PBI Wallet MEM `1F3w` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
20222	GA, FL	323, 202	f-escorts	–	3	$8
111237	CA	831	body rubs	lady, white	3	$26

Table A.187. PBI Wallet MEM `14Qo` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
11429	GA, CO, FL, DC, CT, MA, TX	509	bdsm	lady	10	$28
11429	GA, CO, FL, DC, CT, MA, TX	509	bdsm	lady	10	$28
11429	GA, CO, FL, DC, CT, MA, TX	509	bdsm	lady	10	$28
11429	GA, CO, FL, DC, CT, MA, TX	509	bdsm	lady	10	$28

Table A.188. PBI Wallet MEM `1MD` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
63787	NV	725, 702	f-escorts	–	1	$1170
80678	NC, GA	443, 004	f-escorts, body rubs	woman	9	$24

Table A.189. PBI Wallet MEM `1DpT` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
10839	FL, CO	719	m-escorts, f-escorts	puerto rican	21	$230
128995	FL	150, 100, 772	f-escorts	woman, sicilian, girl	39	$30
136850	VA	757	f-escorts	woman	3	$135

Table A.190. PBI Wallet MEM `1L6E` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
6230	AZ	619	f-escorts	–	1	$3
38105	CA	916	f-escorts	asian	11	$260

Table A.191. PBI Wallet MEM `1MCS` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
7453	NH, CT, VT, MA, ME	617	body rubs	lady, girl	8	$64.75
72128	NY	605, 201, 786, 617, 646	f-escorts	girl, latina, brazilian	12	$1

Table A.192. PBI Wallet MEM `13vH` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
6888	AR	816, 949	f-escorts	girl	1	$18
30435	IN, VA	304, 574	f-escorts	girl	8	$102.90
99136	WA, OR	404, 206	f-escorts	lady, woman, ebony	9	$24

Table A.193. PBI Wallet MEM `1Lew` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
44629	CA	925	f-escorts	–	3	$323
46090	AR, AL, TN, los angeles	347, 901, 870	body rubs, f-escorts	girl	91	$231
80076	NC	919	f-escorts	young, baby, woman, girl,	4	$27

Table A.194. PBI Wallet MEM `1G2S` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
115594	CA	510	f-escorts	asian, girl	1	$30
115597	CA	707	f-escorts	young	1	$40

Table A.195. PBI Wallet MEM `12T9` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
27507	IL	217	body rubs	lady, youthful, asian	39	$40.60
46400	los angeles	225	body rubs	girl, asian, young, korean	6	$570

Table A.196. PBI Wallet MEM `13Yo` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
9940	TX	214, 469	f-escorts	–	2	$22
24265	IL	312	datelines, f-escorts	girl, lady, american, european	61	$2156.40

Table A.197. PBI Wallet MEM `13af` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
8	IL	–	bdsm	–	1	$8
9	IL	–	bdsm	–	1	$8

Table A.198. PBI Wallet MEM `16qB` Statistics

Author ID	Post Locations	Phone Area Codes	Post SubCategories	Top Demographic Tokens	No. Ads Posted	$ Spent
1	AR	–	f-escorts	–	1	$3
2	OR	–	f-escorts	–	1	$3
3	OK	–	f-escorts	–	1	$3
4	MD	–	f-escorts	–	1	$3
5	CO	–	f-escorts	–	1	$3
6	MS	–	f-escorts	–	1	$3
7	FL	–	f-escorts	–	2	$3

Table A.199. PBI Wallet MEM `1Ejb3` Statistics